UNIVERSITÀ DI TORINO

How artificial intelligence can further European multilingualism

Strategic recommendations for European decision-makers

edited by
Rachele Raus
Università di Bologna
Member of the Jean Monnet Centre of Excellence on
Artificial Intelligence for European Integration
Università di Torino

Artificial Intelligence for European Integration | Report - 2023

How artificial intelligence can further European multilingualism

Strategic recommendations for European decision-makers

edited by
Rachele Raus
Università di Bologna
Member of the Jean Monnet Centre of Excellence on
Artificial Intelligence for European Integration
Università di Torino

ai4ei
Artificial Intelligence for European Integration
Jean Monnet Centre of Excellence

Artificial Intelligence for European Integration | Report - 2023

UNIVERSITÀ DI TORINO

Artificial Intelligence for European Integration | Report - 2023

AI4EI
www.jmcoe.unito.it

Collane@unito.it
Università di Torino

ISBN ebook: 9788875902728
ISBN cartaceo: 9791256000142

ai4ei | Artificial Intelligence for European Integration
Jean Monnet Centre of Excellence

CPS CULTURE POLITICA SOCIETÀ

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

Ledizioni
The Innovative LEDIpublishing Company

Ledizioni LediPublishing
Via Antonio Boselli, 10
20136 Milan – Italy
www.ledizioni.it
info@ledizioni.it

INDICE

RECOMMENDATION 3

Europe must invest in developing multilingual corpora from authentic national material that reflects the range of diatopic variation

RECOMMENDATION 4

Europe must invest in developing language and computer technologies that are truly Made in EU

Rachele Raus
Università di Bologna
Member of the Jean Monnet Centre of Excellence on *Artificial Intelligence for European Integration*
Università di Torino

Introduction

This report is intended to provide the European Union's policy- and decision-makers with a solid basis for making informed choices about investing in artificial intelligence to promote European multilingualism, offering four specific recommendations.

Here we will use the term "artificial intelligence"[1] to mean the "theory and development of computer systems able to perform tasks normally requiring human intelligence, such as visual perception, speech recognition, decision-making, and translation between languages" (IATE: entry ID 3571274[2])[3]. Moreover, the term "language industry" will be taken to mean the set of "products, techniques, activities or services that entail natural language processing (from the French definition, IATE entry ID 921669).

With the spread of "large language models", or in other words computer models capable of automated unsupervised, self-supervised or semi-supervised deep learning based on enormous quantities of data generally taken from the Internet, as in the highly controversial case of ChatGPT, the time has come to turn objective, scientific attention to these models so that shared measures can be taken to safeguard one of the EU's key values: multilingualism.

To this end, the work group whose language activities I had the pleasure of coordinating with the network set up by the Jean Monnet Centre of Excellence on Artificial Intelligence for European Integration (AI4EI)[4] n Turin—a network which has since regrouped as the AI4EI Observatory[5] inaugurated at the Università di Torino on 12 December 2023—brought together experts in linguistics, language teaching, translation, discourse analysis, computer science and information engineering in a dialog to produce data and studies about AI's impact on the language industry and, ultimately, on multilingualism in the European Union. Right from the planning stage, it was decided to use the 'circular' research method, in a combination of the conventional bottom-up model with a top-down model that involved people with a wide range of functions—management, teaching, and even graduate students—of different ages, gender and levels of experience.

[1] The main terms used herein this report are defined in the *Glossary* provided at the end of this report

[2] https://iate.europa.eu

[3] All websites referenced in this report were checked on 31 July 2023.

[4] https://www.jmcoe.unito.it/home

[5] https://www.observatory.unito.it/

These studies have helped cast a sharp light on the current linguistic, social and cultural biases, or distortions of reality and errors that can lead to full-blown prejudices, promulgated by the deep learning-based language industry. In so doing, they have enabled us to draft four strategic recommendations for EU decision-makers to avoid the negative impact that AI could have on European multilingualism in the coming years.

This report is the outcome of an effort that began on 6 October 2020, when Laurent Romary, chairman of Technical Committee 37 of the International Organization for Standardization (ISO), was invited to participate in the Jean Monnet Centre of Excellence AI4EI project kick-off conference[6]. Romary discussed a question which has been raised in several quarters (Vetere 2023, Kim et al. 2019): the problematic dominance of English not only as the most common working and pivot language in drafting and translating European texts—which, moreover, are often used to train AI algorithms, but also because most language technology products are in English.

These initial explorations were followed up in an international conference on AI's impact on multilingualism held on 23 and 24 April 2021[7], some of whose presentations were then collected in the volume *Multilingualism and Language Varieties in Europe in the Age of Artificial Intelligence*[8]. Also presented on that occasion were the findings of a survey conducted during the 2020-2021 academic year, which was followed by similar polls in the two subsequent academic years. Over the three years, a total of 3328 questionnaires on the use of machine translation based on deep learning were administered to students at ten Italian universities and eight French universities before and after attending teaching modules designed to instil a critical understanding of how AI is employed in the language industry.

In this three-year period, moreover, the initial network of Italian and French universities was expanded via participation in a number of international events (e.g., the *Assises de la Francophonie scientifique*[9]) or research exchanges with other Jean Monnet Centres (including the Hawke EU Jean Monnet Centre at the University of South Australia[10]) and other universities abroad (the

[6] https://www.jmcoe.unito.it/content/kick-conference-ai4ei

[7] https://www.jmcoe.unito.it/content/linguistic-rights-and-language-varieties-europe-age-ai

[8] https://www.collane.unito.it/oa/items/show/132#?c=0&m=0&s=0&cv=0.

[9] https://www.auf.org

[10] https://www.unisa.edu.au/research/hawke-eu-centre-for-mobilities-migrations-and-cultural-transformations/

Universidade Estadual de Campinas[11] in São Paulo, Brazil, for instance). The network now consists of over thirty universities and research institutes based in Italy, Belgium, France, Spain, Australia, Canada and Brazil[12].

A second publication by the network dealing with its work on IA and natural languages in 2021-2023 will soon be posted on the AI4EI Centre of Excellence website.

These exchanges between experts from universities as well as non-academic settings (a number of representatives of private organisations also participated) made it possible to explore the problematic aspects of using artificial intelligence uncritically and, accordingly, foregrounded the need to invest in critical training about AI (Recommendation 1), not least because of the need to invest in the new professions and types of professional training demanded by the language industry job market (Recommendation 2).

Likewise, investing in the development of multilingual corpora is imperative (Recommendation 3). Here, it is essential to understand that even though AI learns from data, the data available on the Web or prepared for training purposes are made up of "corpora" (Rastier 2021), or in other words material that must be appropriately contextualised and selected for research and AI training. Data, in fact, give the impression of being 'neutral' information, whereas the texts used to train neural networks—and which call for human interpretation—do not. In this sense, the informed use of data considered as corpora, and above all the development of European-language corpora, would make it possible to preserve the diatopic variation of these languages—including the 'minority' languages—i.e., their specific features linked to a given culture and a given geographical area, and thus correct the disproportionate amount of IT resources available in English compared to other European languages.

This issue ties in with two further problems: first, there is a lack of AI actually 'made in Europe', given that most of the computer models and technologies employed in deep learning originate in non-EU countries; and second, the mushrooming number of unsupervised or self-supervised models—again, largely produced outside the EU—creates linguistic, social and cultural biases that are incompatible with the EU's goal of promoting social inclusion and multilingualism. Hence the fundamental importance of in-

[11] https://www.unicamp.br/unicamp/

[12] The home universities of the network members and the research centres that contributed to the Centre of Excellence's language research during the three years of Erasmus+ funding are listed in the annex.

vesting in AI that is truly *made in Europe*, (Recommendation 4), as the European Commission has urged since its 2018 Coordinated Plan on Artificial Intelligence (COM(2018)795: 1). For AI that is in fact made in Europe, "assembling" devices is not enough. Europe must develop research and computer technology informed by the need for human-supervised learning models and approaches that can avoid the biases entailed by large language models.

This report is thus aligned with other similar initiatives which call for greater attention to European-made AI from both the legislative and ethical standpoints. Examples include the LAION network's open letter[13], petitioning the European Union to 'establish large-scale supercomputing facilities of AI research, enabling the European research community to study open-source foundation models under controlled conditions with public oversight'[14], or the projects promoted by Humane AI Net, the European Network of Human-Centred Artificial Intelligence[15].
Our four recommendations can be summarised as follows:

1. Invest in new types of critical training about artificial intelligence that can promote an informed use of language industry tools employing deep learning algorithms.
2. Invest in new occupational profiles in the language industry.
3. Invest in developing multilingual corpora from authentic national material that reflects the range of diatopic variation.
4. Invest in developing language and computer technologies that are truly *Made in EU*.

In the following sections, the rationale for each of these recommendations is set out by some of the experts who took part in the AI4EI Centre of Excellence's work over the past three years, and have sought to support the recommendations with data and research findings.

In thanking all the people who contributed to the project for their constructive exchanges on these issues, we hope that this report will serve as a catalyst for incisive European policies on artificial intelligence at a time of sweeping technological and social changes that compel our attention, directing our thoughts to Europe's future and the cultural, language and computer models it will adopt.

[13] https://laion.ai/notes/letter-to-the-eu-parliament/

[14] https://www.unite.ai/laion-and-a-group-of-27/

[15] https://www.humane-ai.eu/research-roadmap

References

European Commission (2018). *Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions. Coordinated Plan on Artificial Intelligence,* COM(2018) 795 final. Bruxelles : European Commission.

Kim Yunsu, Petrov Petre, Petrushkov Pavel, Khadivi Shahram, Ney H. (2019). "Pivot-based Transfer Learning for Neural Machine Translation between Non-English Languages". In: *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, 2019*. Association for Computational Linguistics, 866-876. DOI: 10.18653/v1/D19-1080.

Rastier Francois (2021). "Data vs Corpora". In: Damon Mayaffre, Laurent Vanni (a cura di) *L'intelligence artificielle des textes : des algorithmes à l'interprétation*. Parigi: Champion, 203-245.

Vetere Guido (2023). "Elaborazione automatica dei linguaggi diversi dall'inglese: introduzione, stato dell'arte e prospettive". In: Rachele Raus (cur.) *et al. De Europa Special Issue. 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*. Torino, Milano: Università degli Studi di Torino, Ledizioni LediPublishing, 69-87. https://www.collane.unito.it/oa/items/show/13

RECOMMENDATION 1

Europe must invest in new types of critical training about artificial intelligence that can promote an informed use of language industry tools employing deep learning algorithms.

Artificial intelligence and European multilingualism

Dardo de Vecchi
KEDGE Business School

Technological shock though it may be, AI is now part and parcel of our daily lives inside and outside the EU's borders. Take, for example, machine translation: undeniably, it has made enormous progress, to the point where we normally trust the translations it gives us. On the other hand, determining whether an automatically translated text is reliable and accurate is not easy. Who can say whether a machine translation—a translation, moreover, that is both fast and free—is any good? Very few people are in a position to do so, while the number who rely on machine translation is incalculable. This mass of people who run their documents through machine translators is not made up only of students of the many disciplines that do not focus on language (de Vecchi 2022). It also includes a considerable number of firms for which languages are (only) a tool they need in order to operate, but not one of their raw materials. Business schools train corporate managers, but taking a critical look at AI does not appear to figure among their priorities. And yet, the current enthusiasm for these new technologies indicates that it should (de Vecchi 2022). In any case, it would be advisable for the public to understand both the extent of humans' capacity to carry out a certain number of tasks, and the advantages of using AI to perform them.

A strategic vision of artificial intelligence cannot limit itself to considering AI's functionalities, including its speed in processing data, but must also bear in mind the consequences of eliminating the human factor in a Europe that, paradoxically, wants to put people at the centre of its concerns (European Commission 2019).

Accordingly, our watchword should be *understanding*: understanding what we are dealing with from the technical and social standpoint. Any policy for AI should thus centre on preparing people for its use, an aspect that, consequently, should never be absent from educational programmes, especially in a multilingual Europe.

In the return to multilingualism, the human being must be the protagonist of applications based on deep learning about languages. AI does not have the 'capacity' to decide, for example, what reasons there may be for protecting, promoting, translating or teaching a given language. Even less does it have the ability to establish strategies for teaching or dealing with 'majority' or 'minority' languages[1]. What we must do, then, is reflect on our linguistic heritage. And this is an issue that AI cannot resolve with the tools that are now available.

Such reflection must involve all of Europe when thinking of interdisciplinary collaborations and long-term strategies that con-

[1] For a definition of 'minority' language and a discussion of these notions, see Agresti in this report.

sider English as an essential language that, seemingly, we cannot do without. Nevertheless, diglossia, or the use of two languages in the same community of speakers, tells us that different languages can perform specific functions within the same social group, and even within a society. Languages can be bridges rather than barriers. Can AI handle this linguistic dichotomy? If well managed, it very likely can.

Although negotiations and trade often rely on English, we normally prepare for both using our mother tongue, the only way to forge a bond of trust. The EU should thus invest not only in instilling an understanding of AI, but also in training people to reflect on its use. This, for instance, can prevent situations where people are misled in dealings that do not take place in their own language. The history of Internet reminds us of how a tool can get out of hand in ways its creators could never have foretold.

From a linguistic standpoint, it is thus essential to teach the differences between language as a human faculty, language as a specific tongue, and language as discourse. Above all, what must be borne in mind is the nature of language as a means of representing thought, of which it is an instrument. Specialists are well aware that these distinctions cannot be ignored, given that AI deals with—or rather, manipulates—different realities. And these distinctions must be taught and remembered, as it is crucial at the time AI is 'fed' (and the metaphor is an apt one) with data. The *No Language Left Behind* or NLLB-200 model claims to handle 200 languages; comforting though this is, multilingualism, and especially European multilingualism, must not be tempted by these siren songs. Rather, it must foster an awareness of the linguistic heritage, of its importance and impact in the society we live in. We may thus legitimately ask to know who will decide what languages are to be 'handled'.

Lastly, the key question revolves around the values at the basis of our actions, and for which we aspire to build a European multilingualism that can come up to our expectations. Do we want to do it ourselves, or do we want to leave it to the machines, eliminating all human intervention in the process? In his *The Last of the Vostyachs* (2012), Diego Marani writes that all the languages in the world are needed for humanity to survive. And with a well-thought-out, conscientious European multilingualism, the European citizen will survive.

References

De Vecchi Dardo (2022). « Le multilinguisme européen et l'IA. Enquête auprès des futurs décideurs ». In: Rachele Raus (cur.) *et al*. *De Europa Special Issue. 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*. Torino, Milano: Università degli Studi di Torino, Ledizioni LediPublishing, 215-245. https://www.collane.unito.it/oa/items/show/132.

European Commission (2019). *Building Trust in Human-Centric Artificial Intelligence* COM (2019)168. Bruxelles : European Commission.

Marani Diego (2009). *The Last of the Vostyachs*. Dublin : Dedalus.

Meta AI at the page: https://ai.facebook.com/research/no-language-left-behind/

R1

Artificial intelligence for professionalising multilingual competences in Europe

Maria Margherita Mattioda, Ilaria Cennamo
Università di Torino

Silvia Domenica Zollo
Università di Napoli "Parthenope"

The technological advances spurred by multilingual Artificial Intelligence, or AI (Yvon 2022), have a broad social and economic impact[1], that extends from the educational sphere[2] to the job market (ELIS 2022: 25; Zilner *et alii* 2021).

The surge in deep learning-based multilingual technologies invites us to reflect on which innovative teaching practices (Cennamo, De Faria Pires 2022) should be promoted in university foreign language (FL) programmes to meet the need to professionalise multilingual competences in the European digital market. Multilingual AI has gradually become an integral part of Europe's language and translation services (EMT 2022: 7), and is also sparking growing interest in the international business community[3]: studies addressing this sector (Lecomte *et alii* 2023) emphasise the strategic role of multilingual and translation competences in the internationalisation of European organisations, as well as the increased attention to new technologies, though integrating them in the workflow is still at the experimental stage (Wilmot 2022: 86).

Against this backdrop, the European Union sees multilingualism and AI as two areas of strategic importance, given the potential that both hold for the EU's economic and social growth. The European Commission has embarked on numerous initiatives to encourage the Member States to formulate national AI strategies for education and training (viz., *Artificial Intelligence for Europe* in 2018, the *Coordinated Plan on Artificial Intelligence*, also from 2018, the 2021 *Review of the Coordinated Plan on Artificial Intelligence*, and the 2020 *White Paper on Artificial Intelligence*). More specifically, the EU believes that AI and the associated language technologies can be used in developing or applying new didactic methods in such sectors as FL teaching and learning, and, consequently, can preserve multiculturalism, multilingualism and more generally, European and national cultural and linguistic diversity, in academia as elsewhere.

However, if using AI is to prove beneficial for multilingual university education, the EU must put more effort into promoting di-

[1] As discussed in the seminar "*Impacts sociétaux de l'intelligence artificielle*" held at the Université Bordeaux Montaigne on 10 November 2022. https://www.u-bordeaux-montaigne.fr/fr/actualites/recherche/impacts-societaux-de-l-intelligence-artificielle.html

[2] This impact was explored at the *Translating Europe Workshop* 'L'intelligenza artificiale per la traduzione: verso una nuova progettazione didattica?' organised by the European Union's Directorate-General for translation and the Università degli Studi di Torino on 3 December 2021. https://italy.representation.ec.europa.eu/system/files/2021-11/TEW_Torino%20-%20Programma.pdf

[3] As can be seen from the many generative applications and intelligent services developed for businesses. An example is: https://www.oneai.com/

gital and language education in order to support the development of the skills of the future through an ethical, responsible and cross-cutting approach to AI-based technologies.

From a pedagogical standpoint, we believe that this kind of interdisciplinary education should not be available only to students in STEM programmes (degree programmes in computer science and engineering, for example). Rather, it should encompass the largest possible number of disciplines, figuring prominently in the human, political and social sciences, economics, statistics and business studies, thus meeting the demands of a rapidly evolving job market that is increasingly digital and multilingual.

Accordingly, there is an urgent need to assess what targeted action can be taken to integrate digital technologies strategically in the new FL teaching and learning practices in all university settings, as well as how the EU can support the Member States in formulating their employment and education policies. An essential prerequisite for engaging with this profound change is the ability to train people for hybrid jobs calling for skills that are both transversal and instrumental (Zollo 2022) as well as investing in digital literacy and, more specifically, in machine translation literacy (Bowker, Ciro 2019; Bowker 2020; Bowker 2021; Loock, Léchauguette 2021) through lifelong learning and other means. Gaining these skills is also a question of adopting a critical teaching approach that, by analysing the potential and limitations of neural network-driven automatic services, can enhance human language, translation and communication skills and thus bring about a more informed interaction with the machine (Cennamo, Mattioda 2022). Such a teaching approach is an effective response to the need for modular professional profiles with multilingual skill sets (Miličević *et al*. 2021) for specialists working in many areas of international cooperation. Moreover, as multilingual technologies can be put to a multitude of uses, including machine translation (Monti 2019: 20)—which can serve as aids for translating, writing, understanding and interacting in foreign languages and as a resource in learning a new language—university education can draw on a plurality of tools and ways of integrating them that can be deployed in FL teaching and learning in the professional contexts discussed above, both as part of the core curriculum and as supplementary material. FL teaching should thus evolve towards interdisciplinary goals where AI-based language skills are gained with a view to applying them in such areas as business communication, sectoral languages, in-

[4] See https://education.ec.europa.eu/education-levels/higher-education/european-universities-initiative

stitutional settings, translation, verbal interaction and multilingual content creation. Interdisciplinarity should also be a means of building soft skills such as creativity, managerial ability and problem-solving in ever-more complex multilingual and technological settings. As the findings of the studies carried out under the aegis of the *Artificial Intelligence for European Integration* (AI4Ei) project (Raus *et al*. 2022) indicate, it would be particularly important to offer training programmes and work placements at European public agencies, research centres and firms where students are called upon to deal with real-world problems using AI language technologies and methods.

On the political front, the European Universities Initiative's[4] strategies outlined for 2024, which include establishing common criteria for awarding a joint European Degree, call for an explicit focus on multilingualism and digital literacy applied to multilingual technologies in the 'European' universities which, as such, plan to work together to promote the development of skills and knowledge that meet the needs of a multilingual European digital single market. Specifically, the strategy for European universities must clarify the role assigned to teaching/learning FLs other than English, especially in university programmes that promote international openness essentially through courses offered in English as a lingua franca. In this sense, it should be emphasised that European policies play a decisive part shaping future internationalisation strategies, as it is necessary to incentivise Europe's universities to develop educational programmes and international research projects that are representative of European innovation and diversity in the world.

In conclusion, we must stress the importance, first, of investing in innovative foreign language teaching, which cannot ignore the pervasive spread of AI and its innumerable multilingual applications. Second, from the standpoint of European internationalisation, it is no less important to provide multilingual university education centring on learning European languages other than English. By requiring students to attend courses in at least one other European language in addition to the courses held in English as a lingua franca, such programmes can help make an innovative, multilingual European university a reality.

References

Bowker Lynne (2021). "Promoting linguistic diversity and inclusion: Incorporating machine translation literacy into information literacy instruction for undergraduate students". *The International Journal of Information, Diversity and Inclusion*, Volume 5, Issue 3, 127-151.

Bowker Lynne (2020). "Machine translation literacy instruction for international business students and business English instructors". *Journal of Business and Finance Librarianship*, Volume 25, Issue 1-2, 25-43.

Bowker Lynne, Buitrago Ciro Jairo (dir.) (2019). *Machine Translation and Global Research: Towards Improved Machine Translation Literacy in the Scholarly Community*. Bingley: Emerald. DOI: 10.1108/9781787567214

Cennamo Ilaria, De Faria Pires Loic (2022). "Intelligence artificielle et traduction. Les défis pour la formation et la profession". In: Maggi Ludovica, Bordes Sarah (a cura di). *Intelligences pour la traduction. IA et interculturel: actions et interactions*, *Revue internationale d'interprétation et de traduction / International Journal of Interpretation and Translation FORUM [20:2] Special Issue*. Amsterdam: John Benjamins Publishing Company, 333-356. https://doi.org/10.1075/forum.20.2

Cennamo Ilaria, Mattioda Maria Margherita (2022). "La traduzione automatica neurale: uno strumento di sensibilizzazione per la formazione universitaria in lingua e traduzione francese". In: Rachele Raus (cur.) *et al.*. *De Europa Special Issue. 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*. Torino, Milano: Università degli Studi di Torino, Ledizioni LediPublishing, 307-331. https://www.collane.unito.it/oa/items/show/132

Curry Edward, Metzger Andreas, Zillner Sonja, Pazzaglia Jean-Christophe, García Robles Ana (2021). "Data economy 2.0: From big data value to AI value and a European data space". In: *The Elements of Big Data Value: Foundations of the Research and Innovation Ecosystem*. Cham: Springer International Publishing, 379-399.

ELIS (2022). *European Language Industry Survey 2022 Trends, expectations and concerns of the European language industry*. https://elis-survey.org/wp-content/uploads/2022/03/ELIS-2022-report.pdf?utm_source=elis-repository&utm_medium=website&utm_campaign=elis-report22&utm_id=elis-report-22

European Commission (2021). *Coordinated plan on artificial intelligence. 2021 Review COM(2021) 205 final,* Brussels : European Commission. https://digital-strategy.ec.europa.eu/en/library/coordinated-plan-artificial-intelligence-2021-review

European Commission (2020). *European Skills Agenda for sustainable competitiveness, social fairness and resilience COM(2020) 274 final*, Brussels, https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52020DC0274&from=EN

European Commission (2020). *White Paper on Artificial Intelligence—A European approach to excellence and trust COM(2020) 65 final* Brussels: European Commission. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52020DC0065

European Commission (2020). *The future of language education in Europe: case studies of innovative practices*. Luxembourg : Publications Office of the European Union. https://nesetweb.eu/wp-content/uploads/2020/05/NESET_AR_2020_Future-of-language-education_Full-report.pdf

European Commission (2018). *Coordinated Plan on Artificial Intelligence COM(2018) 795 final*. Brussels : European Commission. https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52018DC0795&from=DE

European Commission (2018). *Artificial Intelligence for Europe COM(2018) 237 final*. Brussels : European Commission. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM%3A2018%3A237%3AFIN

European Master's in Translation—EMT (2022). *Competence Framework 2022*. https://termcoord.eu/2022/11/updated-version-of-the-emt-competence-framework-now-available/

Lecomte Philippe, Vigier Mary, Gaibrois Claudine, Beeler Betty (2023). *Understanding the Dynamics of Language and Multilingualism in Professional Contexts*. Cheltenham: Edward Elgar Publishing.

Loock Rudy, Léchauguette Sophie (2021). "Machine translation literacy and undergraduate students in applied languages: report on an exploratory study". *Tradumàtica*, 19, 204-225.

Miličević Petrović Maja, Bernardini Silvia, Ferraresi Adriano, Aragrande Gaia, Barrón-Cedeño Alberto (2021). *Language data and project specialist: A new modular profile for graduates in language-related disciplines*. UPSKILLS Intellectual output 1.6. Zenodo. https://dx.doi.org/10.5281/zenodo.5030929

Monti Johanna (2019). *Dalla Zairja alla traduzione automatica. Riflessioni sulla traduzione nell'era digitale*. Naples: Paolo Loffredo Editore.

Raus Rachele, Silletti Alida, Zollo Silvia Domenica, Humbley John (2022). « Multilinguisme et variétés linguistiques en Europe à l'aune de l'intelligence artificielle ». In: Rachele Raus (cur.) *et al.*. *De Europa Special Issue. 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*. Torino, Milano: Università degli Studi di Torino, Ledizioni LediPublishing. https://www.collane.unito.it/oa/items/show/132.

Wilmot Natalie Victoria (2022). *Language Management: From Bricolage to Strategy in British Companies*. Bristol: Multilingual Matters.

Yvon François (2022). "Evaluer, diagnostiquer et analyser la traduction automatique neuronale". In: Maggi Ludovica, Bordes Sarah (a cura di). *Intelligences pour la traduction. IA et interculturel : actions et interactions*, *Revue internationale d'interprétation et de traduction /*

International Journal of Interpretation and Translation FORUM [20:2] *Special Issue*. Amsterdam/Philadelphia: John Benjamins Publishing Company, 315-332. DOI: https://doi.org/10.1075/forum.20.2

Zollo Silvia Domenica (2022). « De l'usage raisonné des ressources documentaires numériques pour le développement de la compétence instrumentale dans la didactique de la traduction spécialisée : retours d'expérience et perspectives en contexte LANSAD ». *Revue Traduction et Langues*, 21(1), 157-172.

Zollo Silvia Domenica, Calvi Silvia (2022). "Fraseologia, traduzione e *digital literacy* nel contesto universitario: riflessioni e proposte per un percorso didattico sperimentale". In: R. Raus *et al*. *De Europa Special Issue. 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*. Torino, Milano: Università degli Studi di Torino, Ledizioni LediPublishing, 263-284. https://www.collane.unito.it/oa/items/show/132

Sitography

European Commission. *European Universities Initiative*. https://education.ec.europa.eu/education-levels/higher-education/european-universities-initiative

European Commission. *Blueprint for sectorial cooperation on skills*. https://ec.europa.eu/social/main.jsp?langId=en&catId=1415

Artificial intelligence, machine translation and language learning

Alessandra Molino, **Lucia Cinato**
Università di Torino

Neural Machine Translation (MT) can be very useful in easing communication between speakers of different languages. Increasingly, it is an integral part of such commonly used tools as multilingual instant messaging, speech translation and online search engines. Neural MT also supports multilingualism for the European Union's institutions and citizens. Although this technology's contributions are undeniable, close watch must be kept on how it develops, investing in research and formulating policies that can channel its progress and use according to consensus principles. Neural MT must become a resource, not an encroachment on the rights and values of equality and inclusion. Europe must remain true to its motto "United in Diversity", and must continue to promote initiatives for linguistic diversity and learning at least two foreign languages, as called for by the 2002 Barcelona objective. It is also essential to raise the public's awareness of the limitations of MT and the benefits of learning a foreign language.

From this perspective, language learning policies cannot ignore the explosive role of neural MT and artificial intelligence (AI, and generative models in particular), which must be gradually integrated in educational programmes to benefit from their potential, as well as to draw attention to their limitations and prepare the rising generations for an informed, critical use of these tools. Research in what has been called 'MT literacy' is now getting under way. A noteworthy project, and the only one to date funded by the Erasmus+ programme (Key Action: Cooperation for innovation and the exchange of good practices), is MultiTraiNMT—Machine Translation training for multilingual citizens[1] coordinated by the Universitat Autonoma de Barcelona. The project's achievements include:

1) A syllabus for an MT course for multilingual citizens which also includes a foreign language learning module.
2) The MutNMT online platform that enables learners to gain insight into the internal workings of neural MT and train an ad hoc engine with user-uploaded corpora.
3) A book on MT for non-experts, with downloadable learning activities, supplementary teaching materials, and an online learning activity explorer.

What is needed now are further projects of this kind addressing different types of user and differentiating between the different scenarios where AI and MT are used in order to develop realistic, targeted policies.

[1] 2019-1-ES01-KA203-064245, 1 September 2019–31 August 2022, https://erasmus-plus.ec.europa.eu/projects/search/details/2019-1-ES01-KA203-064245?etrans=it

The almost complete lack of data and information about how MT can influence the processes and outcomes of language learning is a matter of concern for educators, language teachers and linguists. Existing studies (see, for example, Somers *et al*. 2006; Thue Vold 2018; Fredholm 2019; Carré 2022) suggest that the level of foreign language competence and understanding how MT works are important factors in recognising errors and shortcomings in machine translated texts. Such an awareness should limit the risk of using the language incorrectly and internalising certain standardised lexical and syntactic forms. Nevertheless, many questions remain. How are MT tools used by learners of different ages and levels of competence and education? Aside from sporadic personal use, how can MT and generative AI be integrated in language teaching? Does using online MT platforms have a positive effect on learners' written and spoken production? If it does, is the effect temporary or permanent? The *MTrill: Machine Translation Impact on Language Learning*[2] project coordinated by Dublin City University provided some early answers about MT's effects. The MTrill project conducted a syntactic priming experiment to determine whether participants spontaneously reuse the syntactic structures they had seen during a translation task using MT in their subsequent speech in English as a second language, finding that MT can in fact leave traces in the learning process: '[p]articipants trusted the GT [Google Translate] output enough to change their linguistic behaviour in order to mirror the system's choices' (Resende, Way 2012: 82).

The ethical dimension is a further aspect to be considered in teaching languages and shaping informed citizens. Teaching foreign languages is not just a question of learning their linguistic codes, but also of learning the culture linked to the languages. Learners must be made aware of cultural differences and sociolinguistic variation.

In conclusion, MT and AI, awareness, teaching, ethics, and digital tools in general—e-dictionaries and corpora, for instance—are all interconnected, indispensable parts of modern language education. As we have emphasized, however, their integration must be solidly based on more complete information, with more extensive testing together with international and interdisciplinary collaborations.

[2] Funded by EXCELLENT SCIENCE—Marie Skłodowska-Curie Actions, 25 April 2019–16 July 2021, https://cordis.europa.eu/project/id/8434550

References

Carré Alice, Kenny Dorothy, Rossi Caroline, Sánchez-Gijón Pilar and Torres-Hostench Olga (2022). "Machine translation for language learners". In: Dorothy Kenny (a cura di) *Machine translation for everyone: Empowering users in the age of artificial intelligence*. Berlino: Language Science Press, 187-207. DOI:10.5281/zenodo.6760024

Fredholm Kent (2019). "Effects of Google translate on lexical diversity: Vocabulary development among learners of Spanish as a foreign language". *Revista Nebrija de Lingüística Aplicada a la Enseñanza de Lenguas*, 13(26). DOI:10.26378/rnlael1326300

Resende Natalie, Way Andy (2021). "Can Google Translate rewire your L2 English processing?". *Digital*, 1, 66-85. DOI: https://doi.org/10.3390/digital1010006

Somers Harold, Gaspari Federico, Niño Ana (2006). "Detecting inappropriate use of free online machine translation by language students — a special case of plagiarism detection". In: *Proceedings of the 11th Annual Conference of the European Association for Machine Translation*. Oslo: European Association for Machine Translation, 41-48.

Thue Vold Eva (2018). "Using machine-translated texts to generate L3 learners' metalinguistic talk". In: Åsta Haukås, Camilla Bjørke, Magne Dypedahl (a cura di). *Metacognition in language learning and teaching*. New York: Routledge, 67-97. DOI: https://doi.org/10.4324/9781351049146

R1

Artificial intelligence and translation education: new skills for specialised translators and revisers

Maria Teresa Zanola
Osservatorio di Terminologie e Politiche Linguistiche (OTPL)
Università Cattolica del Sacro Cuore, Milano
REALITER — Rete panlatina di terminologia

Anna Serpente, Martina Alì
Università Cattolica del Sacro Cuore, Milano

Inevitably, artificial intelligence and the development of new technologies have revolutionised the role and tasks of professionals working in the language sector (*Commission de l'éthique en science et en technologie* 2019), including specialised translators and revisors who now must not only know their working languages, but must also master the use of technological tools for everything from managing the translation process to dealing with different types of text and terminology. This situation has brought new educational needs ranging from university programmes for aspiring translators and revisers to lifelong learning initiatives for professionals who have already embarked on their careers (Frérot, Karagouch 2016; Gambier 2009): programmes where theoretical knowledge is combined with practical skills to simulate the work of today's translators and revisors, preparing competent professionals who can compete on the job market.

In response to these needs, a group of experts from the European Master's in Translation network announced the updated EMT Competence Framework 2022 outlining the five main areas of competence for translation graduates: language and culture, translation (in the broad sense, including strategic and methodological competences and domain-specific knowledge in the professional's area of specialisation), technology, personal and interpersonal skills and, lastly, language service provision. As close scrutiny of this document shows, there is an urgent need for training programmes empowering translators and revisers to benefit from the opportunities held out by artificial intelligence and information technology, while making them aware of the limitations (Flöter-Durr 2022) that call for informed human intervention.

As part of the research project entitled 'Linguistic Rights and Language Varieties in Europe in the Age of AI', the *Osservatorio di Terminologie e Politiche Linguistiche* (OTPL) at the Università Cattolica del Sacro Cuore in Milano conducted trials with graduate students between April 2021 and May 2022 of new training programmes which raise learners' awareness of the need for an informed use of IT tools, with particular attention to the revision of machine translated texts (Calvi, Dankova 2022). These experiments addressed the use of machine translation for multilingual communication in specialised journals such as *Nature* and *National Geographic*, evaluating translation quality and the amount of human revision required (Guasco 2013). It was found that the performance of machine translation is poorer as regards lexical and terminological choice and the use of prepositions. An ad hoc investigation carried out in the domain of sustainable fashion, an

area involving particularly topical issues, found that machine translation has more difficulty in the terminological dimension (Zanola 2018). This finding was borne out by a further experiment centring on the relationship between machine translation and terminology in the domain of climate change and the environment.

This work has confirmed the need for innovative training paying close attention to technological or IT skills, without neglecting the linguistic and cultural competences which are still central to university programmes. In the field of specialised translation, moreover, analysis of machine translations has shown that good revision calls for a thorough theoretical understanding of terminological aspects. Studying the domain in different text types and sources enables the reviser to develop greater sensitivity, especially as regards diatopic, diachronic and diaphasic variation in terminology, which machine translation often does not take into consideration. Accordingly, a solid grounding in terminology theory, together with practical knowhow, is undoubtedly a strength in terms of professionalism and competitiveness. For over twelve years, OTPL has striven to provide such a grounding through its postgraduate programme in "Specialised terminologies and translation services".

References

Calvi Silvia, Dankova Klara (2022). « Industrie de la langue et formation des traducteurs spécialisés ». *Revue Traduction et Langues*, 21(1), 190-204. https://hdl.handle.net/10807/227695

Commission de l'éthique en science et en technologie (2019). *Les effets de l'intelligence artificielle sur le monde du travail. Document de réflexion*. Gouvernement du Québec.

European Master's in Translation (EMT) (2022). *Competence Framework 2022*. https://termcoord.eu/2022/11/updated-version-of-the-emt-competence-framework-now-available/

Flöter-Durr Margarete. (2022). "Epistemological limits of current digital techniques of artificial intelligence in translation". *Lebende Sprachen*, 67(1), 4-44. DOI: https://doi.org/10.1515/les-2022-0004

Frérot Cécile, Karagouch Lionel (2016). *Outils d'aide à la traduction et formation de traducteurs : vers une adéquation des contenus pédagogiques avec la réalité technologique des traducteurs*. *ILCEA* , 27 | 2016, mis en ligne le 08 novembre 2016, consulté le 10 juillet 2023. URL : http://journals.openedition.org/ilcea/3849 ; DOI : https://doi.org/10.4000/ilcea.3849

Groupe d'experts EMT (2009). *Compétences pour les traducteurs professionnels, experts en communication multilingue et multimedia*. Bruxelles: Commission européenne, Direction générale de la traduction.

Guasco Patrizia (2013). *La révision bilingue : principes et pratiques*. Milano: Educatt.

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

RECOMMENDATION 2

Europe must invest in new occupational profiles in the language industry.

Investing in new job profiles

Danio Maldussi
Università di Bergamo

Micaela Rossi
Università di Genova

In an ever-more interconnected society, where everyday communication relies increasingly on artificial intelligence systems (Vetere 2022), it is essential not to lose sight of the issues of cultural and linguistic diversity, and of respect for the parameters of inclusion and equal representation of the concerns of all genders or social, language and cultural groups. This essential need is reiterated in the Council of the European Union's Conclusions of 2022 (Council of the European Union 2022): 'Cultural and linguistic diversity is intrinsic to the European Union and its fundamental values. [...] An ambitious policy of cultural and linguistic diversity should fully integrate sustainability issues and draw on technological innovation, including in the digital field'. Now, however, the seemingly exponential growth of multilingualism in the applications that artificial intelligence can support—the European Parliament resolution of 3 May 2022 on artificial intelligence in a digital age (2020/2266(INI)) calls for the implementation and development of AI technology in minority languages, in the belief that this could boost their knowledge and use—does not necessarily mean a greater or more equitable attention to diversity (Larsonneur 2021). This disconnect between touted inclusiveness and the actual homogenisation resulting from models that are essentially trained on datasets from dominant speech communities can be more conducive to discrimination than to any real sharing of content, perpetuating dynamics of domination over certain social and cultural groups (Markl 2022) and denying these communities any role or empowerment in global communication.

Accordingly, specialists in communication must develop new skillsets: there is an urgent need to train experts who can navigate the complexity and diversity of linguistic expression while preserving its variety. Above all, it is crucial that people in such new jobs as prompt engineering, post-editing and programming show a mastery of diatopic, diamesic, diastratic and diaphasic variations in language systems and can train AI-based applications to recognise, define and identify them in large text corpora. Hence the related need to know how to build large corpora for machine learning designed specifically to safeguard linguistic variation from the kind of levelling-out that AI can cause. These skills will make it possible to defend plurality and diversity in effective multilingual communication.

Take, for example, the universe of small and medium enterprises where there is a growing demand for "tailor-made" machine translation systems that can cope with the linguistic and terminological needs of specific sectors and corporate environ-

ments. This need is not restricted to manufacturing, but is also felt in the service industry, as in the case of language service firms, for instance. At the level of individual firms, even more specific needs are posed by the in-house jargon and occupational variation (Bertaccini, Matteucci 2005) that can bring a sense of belonging and integration. Respect for plurality and diversity in communication is thus a strategic asset for all firms doing business in competitive areas.

However, managing complexity is not simply a question of identifying new professional roles. It also extends to designing output that can meet the need to optimise traditional language and terminological 'products' such as glossaries and user manuals, which professionals will be called on to replace with 'scalable' terminology databases in specific working languages and sectors. In addition, they will need to design adaptive machine translation systems.

References

Bertaccini Franco, Matteucci Alessandra. (2005). « L'approche variationniste à la pratique terminologique d'entreprise ». *Meta: Translators' Journal*, 50(4). https://doi.org/10.7202/019910ar

Council of the European Union (2022). *Council conclusions on reinforcing cultural exchanges through the mobility of artists and cultural and creative professionals, and through multilingualism in the digital era* (2022/C 160/07). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52022XG0413(02)

European Parliament (2022). *European Parliament resolution of 3 May 2022 on artificial intelligence in a digital age* (2020/2266(INI)). Bruxelles: European Parliament. https://www.europarl.europa.eu/doceo/document/TA-9-2022-0140_EN.html

Larsonneur Claire (2021). « Intelligence artificielle ET/OU diversité linguistique : les paradoxes du traitement automatique des langues ». *Hybrid* [en ligne], 7. DOI: https://doi.org/10.4000/hybrid.650

Markl Nina (2022). "Mind the data gap(s): Investigating power in speech and language datasets". In: Bharathi Raja Chakravarthi, B Bharathi, John P McCrae, Manel Zarrouk, Kalika Bali, Paul Buitelaar (Eds.). *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion.* Dublin: Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.ltedi-1, 11-12.

Vetere Guido (2023). “Elaborazione automatica dei linguaggi diversi dall’inglese: introduzione, stato dell’arte e prospettive”. In: Rachele Raus *et al*. *De Europa Special Issue. 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*. Torino, Milano: Università degli Studi di Torino, Ledizioni LediPublishing, 69-87. https://www.collane.unito.it/oa/items/show/132

R2

New professional profiles spanning languages and technologies. Findings of the needs analysis conducted in the Erasmus+ UPSKILLS project

Silvia Bernardini,
Adriano Ferraresi,
Maja Miličević Petrović
Università di Bologna

The changes in the job market arising from the unprecedented growth of technology, and especially of artificial intelligence, are far from being a novelty. In the language professions in particular, machine translation is but one factors that have not only affected work practices, but have also impacted educational needs, requiring university programmes to address the development of new technologies and machine translation models. As the growth of artificial intelligence continues to pick up speed, being familiar with systems such as ChatGPT is not enough: we must also have an idea of how they work, what they are based on, and what they can—and cannot—do. This leads to the need for new professional profiles spanning languages and engineering/programming, and stakeholders such as universities must do their part. To ensure a better match between the needs of the job market and those of higher education, the Erasmus+ project *UPgrading the SKIlls of Linguistics and Language Students—UPSKILLS*[1] has thus outlined a new, composite professional profile: the *language data and project specialist* (Miličević Petrović *et al*. 2021).

The specifications for this new professional profile in language- and linguistics-related fields stem from a detailed needs analysis, which found that there is a clear need for a new skillset, and above all for a new mind frame, to meet the professional challenges of the industry and its job market. The needs analysis started from a survey of the curricula of degree programmes in linguistics, modern languages and linguistic mediation at a sample of European universities. The survey found that additional learning content in line with job market requirements should be provided to empower students and educators to make the most of skills that are often already covered in the curricula, but only implicitly. The second component of the needs analysis, a review of the academic, institutional and professional literature, identified six important skill clusters, putting particular emphasis on transversal research skills. The third component was a corpus-driven empirical study of the most frequently recurring words and phrases in job advertisements. This study identified and classified the skills and competences typically required by the language industry, as well as the job titles given to these new, hybrid professions involving languages and technologies. This preliminary work served as the basis for a questionnaire administered to companies hiring linguists and language professionals. In turn, the skills and competences mentioned in the responses to the questionnaire were used to plan semi-structured interviews conducted with industry representatives. All components of the

[1] https://upskillsproject.eu

needs analysis confirmed that language programme curricula should place more emphasis on developing technological, managerial and transversal skills, especially those needed for research.

The profile of the *language data and project specialist* that emerged from the analysis is not linked to a single, specific industry position or job title, nor is it a proposal for a new degree programme. It is a deliberately generic profile, modular and adaptable to the needs of the job market and university programmes. To maintain this intrinsic flexibility, the profile's educational outcomes are structured around two dimensions: a vertical dimension focusing on the seven main domains identified in the UPSKILLS needs analysis—disciplinary, (inter)cultural, technical, data-oriented, research-oriented, organisational and transversal—and a horizontal dimension based on the standard elements of learning outcomes (knowledge, skills and competences). The central idea of the profile is to serve as a guide in selecting the skills and competences to be taught and learned, bearing in mind that most new jobs in the industry call for skills in all of the seven main clusters, with each cluster's contribution depending on the specific job. Figure 1.1 summarises the profile's typical tasks and responsibilities, and the required knowledge, skills and competences (Figure 1.2).

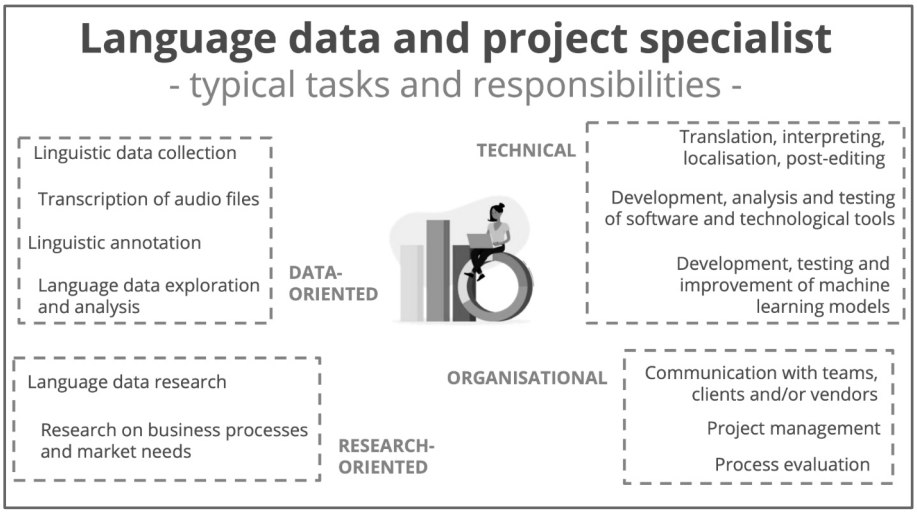

Figure 1.1 - Language data and project specialist tasks and skills
(Source: https://upskillsproject.eu/wp-content/uploads/2021/09/sess1.pres7_.Profile.pdf)

In addition, four more specific sub-profiles were defined, two which focus more on research (*language data analyst and language data scientist*) and two which focus more on management (*language data manager and language project manager*). A distinction is also made based on the level of responsibility and seniority: greater for the language data scientist and language

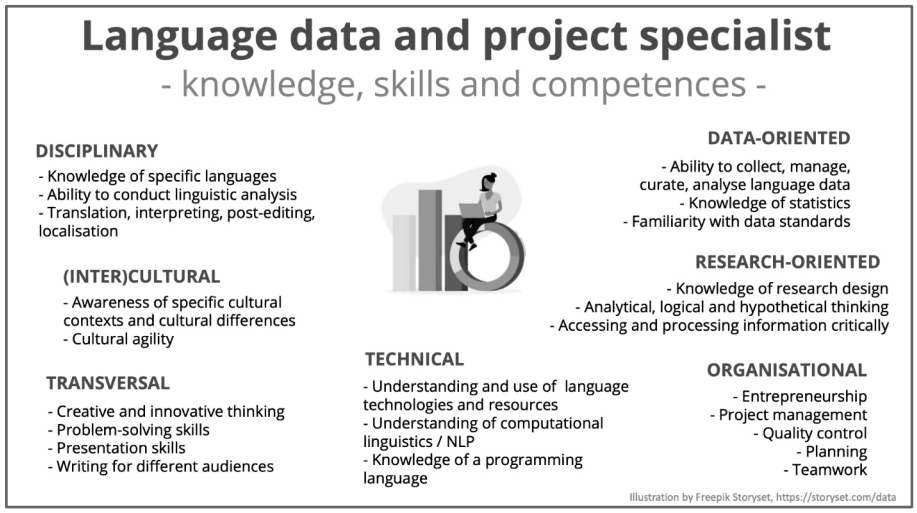

Figure 1.2 - Language data and project specialist tasks and skills
(Source: https://upskillsproject.eu/wp-content/uploads/2021/09/sess1.pres7_.Profile.pdf)

project manager, smaller for the language data analyst and language data manager. By way of example, Figure 2 shows some of the responsibilities and tasks associated with the sub-profiles. The skills and competences they involve are subsets of those for the general profile. For more information on the profiles and sub-profiles, as well as the methodology and findings of the needs analysis, see Miličević Petrović *et al.* (2021). We believe that the

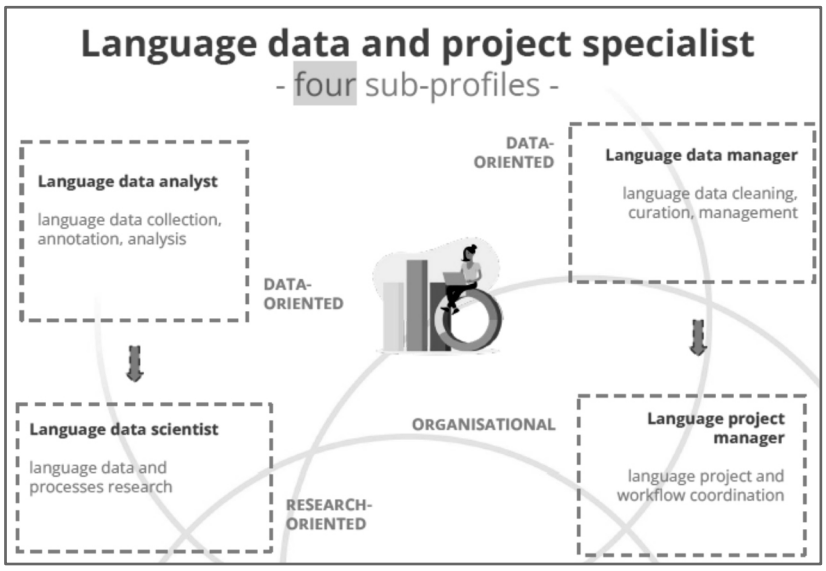

Figure 2 - Responsibilities and tasks for the four sub-profiles
(Source: https://upskillsproject.eu/wp-content/uploads/2021/09/sess1.pres7_.Profile.pdf)

profile and sub-profiles, in addition to summarising the wealth of academic and industry data that emerged from the UPSKILLS needs analysis, can help to bring order to a complex, multidisciplinary and highly varied field, guiding educators in designing new

programmes and teaching materials, and empowering students to put their interests and the knowledge they have gained to good use in pursuing the burgeoning career prospects in the language industry.

References

Miličević Petrović Maja, Bernardini Silvia, Ferraresi Adriano, Aragrande Gaia, Barrón-Cedeño Alberto (2021). *Language data and project specialist: A new modular profile for graduates in language-related disciplines*. UPSKILLS Intellectual output 1.6. Zenodo. https://dx.doi.org/10.5281/zenodo.5030929

A professional profile for more reliable data from Artificial Intelligence: the annotator

Michela Tonti
Università di Bergamo

The digitalisation now sweeping through global society has amplified the growing need for reliable linguistic data to train language technologies for specific application domains. Neural algorithms need human neurons much more than past research would have us believe. Being able to produce and rely on trustworthy data has long been a prime concern for computational linguistics and engineering when building and modelling language technologies (see, for example, Carletta 1996; Artstein, Poesio 2008; Bayerl, Paul 2011) where English is the language of choice. According to EUROSTAT (Vetere 2022: 79), digital media are in general more widely used by firms and private individuals in areas where English is the mother tongue or commonly spoken. Consequently, the better the technologies, the more attractive the languages in which they are available will be, and the more widely they will be used. In turn, the more a language is able to attract users, the more technological resources will be channelled into it.

To preserve linguistic diversity, multilingualism and plurilingualism,[1], training neural networks in other languages is crucial. The proposals for a regulation of the European Parliament and the Council on artificial intelligence (IA) and on digital services of 21 April 2021 and 15 December 2020 demonstrate European institutions' growing awareness of the issues that technological progress has brought to the geopolitical chessboard. Natural language technologies are at the centre of one of the most topical of these issues, as well as being among AI's major sectors.

AI is a computer's ability to mimic cognitive functions of human beings such as learning and problem solving. A computer uses mathematical and logical tools to simulate the ways human beings reason in order to learn new information and make decisions. Machine Learning (ML) is an application of AI where mathematical models are used to help a computer learn independently from experience. To a large extent, this experience is possible thanks to data annotation, a process whereby data are labelled by adding metadata. This is done to show the ideal result to the machine, a result that the model should then be able to replicate on data that have not been analysed beforehand. Annotation is combined almost exclusively with corpora, and serves to facilitate the extraction of linguistic or discursive information. Consequently, an annotated corpus contains the elements that ML must learn to recognise so that it can also be used in future

[1] 'Multilingualism' refers to the presence of more than one languages in a geographical area or, in our case, in an organisation and its translation work, whereas 'plurilingualism' refers to individuals' knowledge of more than one language (see Gaboriaux, Raus, Robert, Vicari 2022: 9).

projects. This type of annotation work takes place before corpora are made public. This means that it is chiefly done by specialists in computational linguistics, for which we have listed several bibliographic references. These specialists provide provisional lists of significant linguistic categories, such as the morphosyntactic components and the contextual conditions for disambiguating them. The rest of the job, which entails complex annotation, falls to the professional role we will discuss here: the annotator.

As annotators' work has until now been restricted to preparing monolingual or parallel corpora, their tasks are specified case by case, according to the categories to be annotated: phonetic, morphological, syntactic, stylistic or discursive. As indicated in the selection of publications focusing on annotation by computational linguists listed in the references, the reliability of the end result depends to a significant extent on the correctness of the annotation and pre-editing processes, both of which should be recognised as increasingly important. While our goal here is to shed light on the role of the annotator, mention must also be made of the pre-editor, given that we are dealing with work preceding the output that the end user obtains. The pre-editor is an integral part of producing/translating texts, and the text that has been edited or translated will be published and thus read by humans. The pre-editor's linguistic skills differ from those of the annotator, whose *modus operandi* is entirely new and unlike the traditional role of the discourse analyst who analyses actual discourse after it has been produced. By contrast, the annotator attempts to predict the effects that could ensue if a given word is used in different discourses, as semantic ambiguities are particularly insidious and costly in terms of post-editing requirements. Inconsistent or incoherent output and other morphosyntactic shortcomings that call for post-editing thus result in time-consuming work and additional expense that can be avoided thanks to the supervised learning that annotation provides. Another plus is that several parts of the annotation process can be automated: for example, *Sketch Engine* software features a standard component that can be used whenever a new corpus is loaded, automatically tagging all morphological categories and a certain number of syntactic categories with a small percentage of errors. Manual annotation is required if it is necessary to tag units that are significant from a syntactic, stylistic and pragmatic standpoint.

In fact, most of the situations now calling for manual annotation involve pragmatic-stylistic elements, or thematic and informational units. This means that a wide variety of non-traditional cat-

egories must be defined and identified, and consequently the skills that annotators are called upon to use change radically from one project to the next. This is an important point, as it underscores the high level of adaptability, appetite for research, and willingness to continue learning that an annotator must show.

An example of the annotation and development of a model is given in Figure 1. The annotator's tasks are represented in the blue boxes, while those in the green boxes fall to the annotation lead who coordinates what the annotators have done: as a large number of texts are needed to train the system, an equally large number of annotators will be required. It is thus essential to harmonize the annotations to ensure that output data are reliable.

We thus believe that supervised machine learning and the role of the annotator are essential for the AI market.

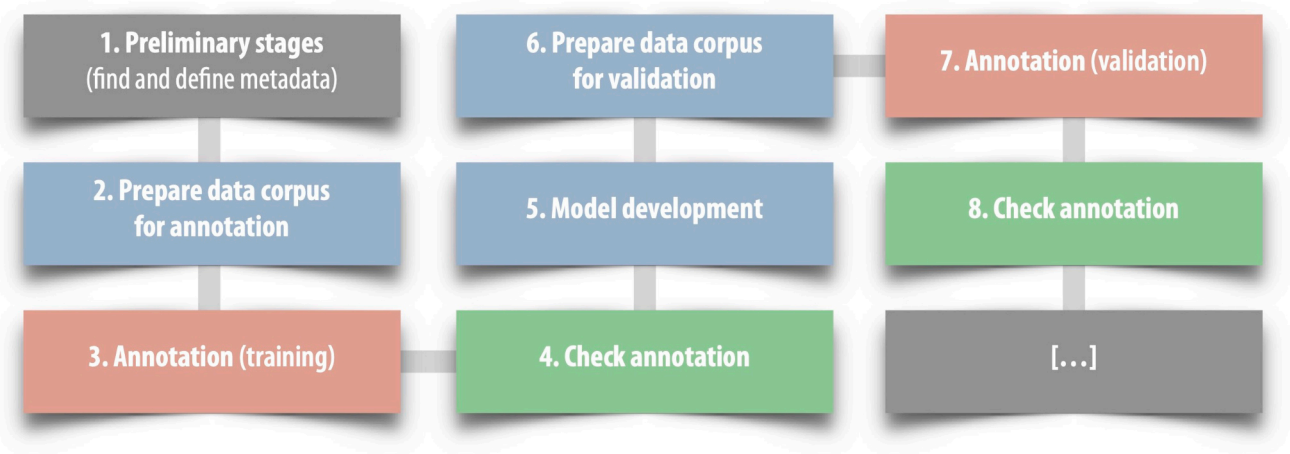

Figure 1. Stages of the annotation and development process: an example

Examples of work conducted with the aid of manual annotation

ELMo: Deep contextualized word representations by Matthew E. Peters *et al*. This paper presents ELMo, a pioneering LLM (Large Language Model) that generates contextualised word representations using deep bidirectional language models.

Probing Neural Network Comprehension of Natural Language Arguments by Timothy Niven and Hung-Yu Kao. This study investigated LLMs' ability and limitations in understanding natural language arguments.

CoQA: A Conversational Question Answering Challenge by Siva Reddy *et al*. This project presented the CoQA dataset centring on answering questions appearing in conversations. In this project, LLMs were required to understand questions and generate valid answers.

SemEval-2020 Task 11: Detection of Propaganda Techniques in News Articles by Giovanni Da San Martino *et al*. This paper presents the challenging task of detecting the propaganda techniques used in news articles, encouraging the use of LLMs and manual annotation to improve detection accuracy.

References

Artstein Ron, Poesio Massimo (2008). "Inter-coder agreement for computational linguistics". *Computational Linguistics*, 34(4), 555–596.

Artstein Ron, Poesio Massimo (2005). "Bias decreases in proportion to the number of annotators". *Proceedings of FG-MoL 2005*, 141–150.

Bayerl Petra Saskia, Karsten Paul Ingmar (2011). "What Determines Inter-Coder Agreement in Manual Annotations? A Meta-Analytic Investigation". *Computational Linguistics,* 37 (4), 699–725, DOI: https://doi.org/10.1162/COLI_a_00074

Carletta Jean (1996). "Assessing agreement on classification tasks: The kappa statistic". *Computational Linguistics*, 22(2), 249–254.

Da San Martino Giovanni, Barrón-Cedeño Alberto, Wachsmuth Henning, Petrov Rotislav, Nakov Preslaw. (2020). "SemEval-2020 Task 11: Detection of Propaganda Techniques in News Articles". In: Herbelot A. *et al*. *Proceedings of the Fourteenth Workshop on Semantic Evaluation*. Barcellona: SemEval, 1377–1414.

European Commission (2021). *Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts*. Bruxelles : European Commission. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206

European Commission (2020). *Proposal for a Regulation of the European Parliament and of the Council on a Single Market For Digital Services (Digital Services Act) and amending Directive 2000/31/EC*. Bruxelles : European Commission. https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52020PC0825

Gaboriaux Chloé, Raus Rachele, Robert Cécile, Vicari Stefano (a cura di) (2022). « Le multilinguisme dans les organisations internationales ». *Mots. Les langages du politique*, n°128. DOI: https://doi.org/10.4000/mots.29135

Niven Timothy, Kao Hung-Yu (2019). "Probing Neural Network Comprehension of Natural Language Arguments". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Firenze: Association for Computational Linguistics (ACL), 4658–4664. DOI: 10.18653/v1/P19-1459

Peters Matthew E., Neumann Mark, Iyyer Mohit, Gardner Matt, Clark Christopher, Lee Kenton, Zettlemoyer Luke (2018). "Deep Contextualized Word Representations". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. New Orleans: Association for Computational Linguistics, 2227–2237. DOI: 10.18653/v1/N18-1202

Reddy Siva, Chen Danqi, Manning Christopher D. (2019). "CoQA: A Conversational Question Answering Challenge". In: *Transactions of the Association for Computational Linguistics*, 7. Cambridge: MIT Press, 249–266.

Seretan Violeta, Roturier Johann., Silva David, Bouillon Pierre (2014). "The ACCEPT Portal: An Online Framework for the Pre-editing and Post-editing of User-Generated Content". In: *Proceedings of the EACL 2014 Workshop on Humans and Computer-assisted Translation*. Gothenburg: Association for Computational Linguistics, 66–71.

Vetere Guido (2023). "Elaborazione automatica dei linguaggi diversi dall'inglese: introduzione, stato dell'arte e prospettive". In: Raus R. *et al*. *De Europa Special Issue. 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*. Torino, Milano: Università degli Studi di Torino, Ledizioni LediPublishing, 69-87. https://www.collane.unito.it/oa/items/show/132.

Sitography

Sketch Engine, http://www.sketchengine.eu

R2

Gender bias and artificial intelligence: a lack of skills?

Mara Floris
Università Vita-Salute San Raffaele

In November 2017, the American writer Alex Shams caused something of a stir on social media when he tweeted 'Turkish is a gender neutral language. There is no "he" or "she"—everything is just "o". But look what happens when Google translates to English'. The text was followed by the screen grab shown in Figure 1, where a few short phrases with the neutral Turkish pronoun 'o' are translated into English (Figure 1):

Personal pronouns are translated as masculine or feminine according to the gender stereotype for the profession in question: 'engineers' and 'doctors' are men, 'nurses' and 'secretaries' are women. The same thing occurs when feelings and attitudes are associated with personal pronouns: the translations reinforce the stereotype of the fragile and emotional woman (Prates *et al*. 2020).

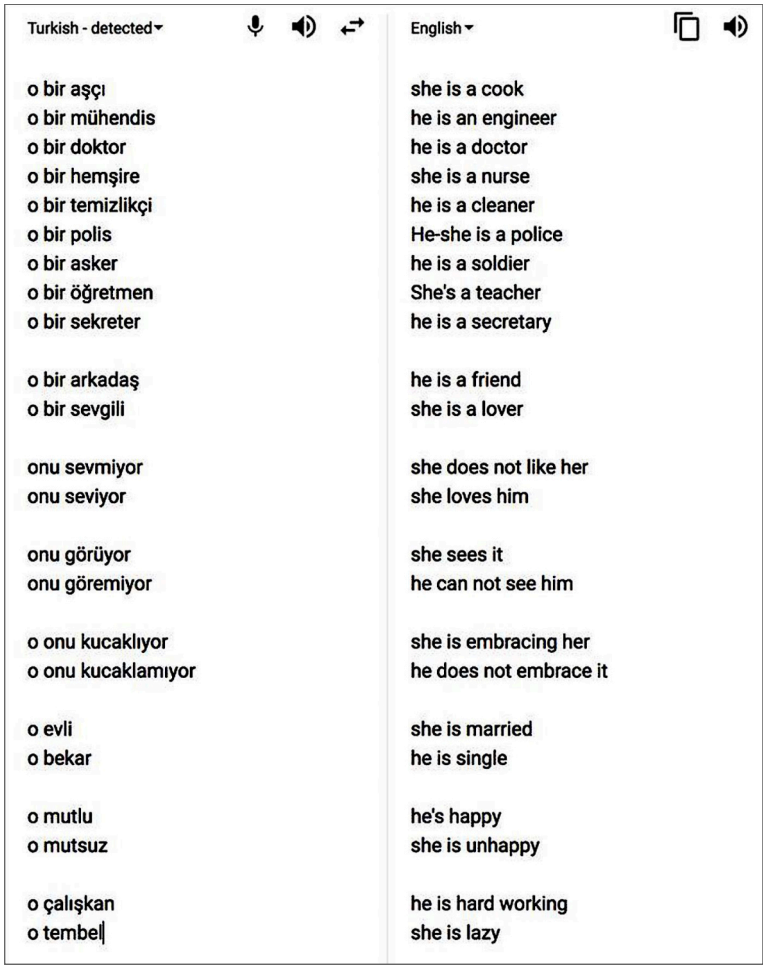

Figure 1. Screengrab from Alex Shams' tweet of 28 November 2017 (source: https://twitter.com/alexshams_/status/935291317252493312)

The literature provides ample evidence of gender bias in machine translations and other Natural Language Processing (NLP) systems (Chen *et al*. 2021; Costa-jussà 2019; Sun *et al*. 2019). The problem is two-fold: there is a lack of specific training to address gender bias, and there is also a lack of specific linguistic skills in languages other than English.

After all the comments on social media, Google made a few tweaks. Now, if you have Google translate *'o bir doktor'* from Turkish into Italian, for instance, you will get 'lei è un dottore', i.e., 'she is a doctor'. Much better, although the correct term in Italian is 'dottoressa', an error that could have easily been avoided by someone with the appropriate language skills who is aware of the gender bias problem.

The problem is especially widespread in natural language processing (NLP) and translation (see Luccioli *et al*. 2020). Gender bias can be defined as favouritism or systematic discrimination towards a particular gender, resulting in unequal treatment and

opportunities. Although producers' declarations of intent state that artificial intelligence (IA) aims to be objective and impartial, it often reflects biases present in the data used to train AI models. This data can contain preconceptions inherited from the society in which it was designed and which lead to biased output.

Like machine translation, sentiment analysis systems are also affected by gender bias. When an NLP system analyses the sentiments in a text, it may mismatch words or expressions typically used by women or men with positive or negative sentiments, creating a distorted interpretation and reinforcing gender stereotypes.

Résumé screening systems are another example of gender bias in NLP. If an automated CV screening system trained on historical data shows a preference for certain term or types of experience typically associated with a specific gender, this preference could discourage female or male candidates, depending on the direction of the bias.

Several major steps are needed to deal fully and completely with all the possible consequences of this discrimination in the world of NLP technologies. New professions must be developed, linguists must gain specialised skills, and a joint effort must be made to put a stop to the processes whereby gender biases are reflected in the technologies we produce. Here, it is crucial to create new professional profiles with a solid linguistic grounding—particularly in languages with grammatical gender—and an understanding of gender biases and their linguistic manifestations. Traditionally, the development of NLP has relied chiefly on the skills of computer scientists and engineers, who may not be fully aware of the nuances and the lack of linguistic inclusion. By involving language specialists with a background in languages other than English, we can ensure a broader and culturally diversified perspective on the development process. With their experience in language analysis, and a sensitivity to gender issues, these specialists can help identify potential discrimination and propose strategies for mitigating them.

In addition, linguists trained in sociolinguistics, discourse analysis and gender studies could drive a better understanding of how gender biases are manifested in language. Drawing on their experience, they can contribute actively to the development of algorithms and models that are more sensitive to different gender identities and expressions.

Blocking the process whereby gender biases crop up in NLP technologies is a fundamental goal. Reaching it will mean identifying and correcting these biases throughout the development

process. By establishing rigorous assessment methods, we can systematically check NLP models for instances of gender discrimination and perfect them iteratively to ensure fairness and inclusion. It is also essential to create diversified and representative training datasets. This can be accomplished by incorporating the views of marginalised communities and consulting people with different gender identities during the data collection and annotation processes. By working actively to reduce gender prejudices in NLP systems, we can promote fairer and more equitable technologies that contribute to a more inclusive society.

References

Chen Yan, Mahoney Christopher, Grasso Isabella, Wali Esma, Matthews Abigail, Middleton Thomas, Nije Mariana, Matthews Jeanna (2021). "Gender bias and under-representation in natural language processing across human languages". In: *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES '21)*. New York: Association for Computing Machinery, 24–34. DOI: https://doi.org/10.1145/3461702.3462530

Costa-jussà Marta R. (2019). "An analysis of gender bias studies in natural language processing". *Nature Machine Intelligence*, 1(11), 495-496.

Luccioli Alessandra, Dolei Ester, Xausa Chiara (2020). "Investigating Gender Bias in Machine Translation. A Case Study between English and Italian". In: Adriano Ferraresi, Roberta Pederzoli, Sofia Cavalcanti, Randy Scansani (a cura di), *MediAzioni* 29: B29-B49. http://www.mediazioni.sitlec.unibo.it

Prates Marcelo O., Avelar Pedro H., Lamb Luís C. (2020). "Assessing gender bias in machine translation: a case study with google translate". *Neural Computing and Applications*, 32, 6363-6381.

Sun Tony, Gaut Andrew, Tang Shirlyn,Yuxin Huang, ElSherief Mai, Zhao Jieyu, Mirza Diba, Belding Elizabeth, Chang Kai-Wei, Yang Wang William (2019). "Mitigating gender bias in natural language processing: Literature review". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Firenze: Association for Computational Linguistics (ACL), 1630-1640. https://aclanthology.org/P19-1159/

RECOMMENDATION 3

Europe must invest in developing multilingual corpora from authentic national material that reflects the range of diatopic variation.

Linguistic diversity, a threatened resource

Linguistic diversity, one component of biodiversity (Le Coadic 2010: 53-56), is currently under threat from the economic-technological paradigm, as well as from a 'neo-babelic' ideology[1]. According to UNESCO, the world's approximately 6700 spoken languages are very unevenly distributed:

1) In terms of absolute number, most languages are on average spoken by a relatively small group of people, while a few languages—Mandarin Chinese, English, Spanish and so forth—are spoken by the majority of the world's population.
2) In terms of status, several widespread languages, generally co-official or second languages, such as Standard Arabic or Kiswahili, are the mother tongues of a modest number of people.
3) In geographic terms, areas with a very high concentration of linguistic diversity (the Indian subcontinent and the Himalayas, Southeast Asia, Central and South America, and sub-Saharan Africa, not to mention the major metropolitan cities) alternate with relatively more homogeneous areas such as Western Europe, North America, North Africa, Northeast Asia, etc.

It is by no means easy to define what a 'minority' language is, and drawing up a list is even more complicated, since being 'minority' depends very much on the context[2]. Defining what makes a language 'threatened' is more straightforward: the *Atlas of the World's Languages in Danger*[3] includes some 2500 languages—around 40 per cent of the languages spoken in the world today—that are at risk of extinction in the coming years.

Clearly, this 'threat' of extinction is not limited to the strictly linguistic level. It also involves serious repercussions in social, economic and environmental terms—as the Pope, for example,

[1] The enormous number of languages has traditionally been, and to a large extent still is, seen as a barrier to economic and practical communication, standing in the way of communication more generally and, consequently, of the peaceful coexistence between peoples. However, the last two wars that have bloodied Europe—the Balkan conflict of the 1990s and the current war in Ukraine— bear tragic witness to the fact that speaking the same language is no guarantee of peace, stability and dialogue between different countries. Pinning hopes for building European democracy on a single shared language (De Mauro 2014) now smacks more of biblical prophecy or a tool of globalising financial hegemony than of providing any reasonable prospect of fairness (Gazzola 2016) and of sustainable development in the environmental and social sense).

[2] Italian is a minority language in the European Parliament but a majority language in Italy; Catalan is a majority language in several cities in Catalonia, a minority language in Spain, and an ultra-minority language in Sardinia; Apulia-Calabrian Greek is always ultra-minority... and so on.

[3] https://unesdoc.unesco.org/ark:/48223/pf0000187026

Dealing with linguistic diversity and artificial intelligence: risks and opportunities

Giovanni Agresti
Centre National de la Recherche Scientifique — CNRS
Université Bordeaux Montaigne

stressed in his encyclical *Laudato si'* (Francis 2015: 111-114). Indeed, the three dimensions we mentioned of language distribution—number, status and geography—make it possible to sense and, at times, foretell potential conflicts, as every language is linked to highly sensitive factors such as individual and collective identity, historical memory, political and economic power, and the rights and duties that obtain in these spheres (Poggeschi 2015). Dealing with linguistic diversity, in the sense of a multidimensional complexity that is never only linguistic or cultural and touches many swathes of society, some narrow, some broad, is ultimately a highly delicate assignment, involving 'top-down' public policies, 'bottom-up' civil society activism and efforts midway between the two (Djordjevic 2018). Consequently, it calls for the utmost attention.

Dealing with linguistic diversity and artificial intelligence: risks and opportunities

Dealing with linguistic diversity has always meant governing networks of human and social relationships of extraordinary complexity and depth—and thus wielding power: the myth of the Tower of Babel, however interpreted, offers universally known proof. Today, digital technologies amplify these networks exponentially, multiplying the levels of interaction: no longer is interaction only in person (verbal and synchronous) and remote (written and asynchronous), but is also in person *and* remote (synchronous remote communication), to say nothing of the question of online access to texts through national libraries, search engines, databases, corpora and all the rest. These technical innovations rewrite the paradigm of human communication, and most notably the notions of 'distance' between interactants, of 'space' and 'time' in communication, and of 'speech community'. By increasing the functions and range of everyday communication to a dizzying extent, the digital accentuates, or rather, exacerbates, the economic value of natural languages, thus encouraging—*de jure* and *de facto*—what is often called linguistic centralisation. To operate efficiently, this process calls for simplification on at least four fronts, viz.:

1) Reducing the number of languages spoken in any given context (region, nation, continent, world) and/or adopting a common language to reduce the 'barriers' to mutual understanding.
2) Simplifying each language's vocabulary.
3) Standardising or 'dumbing down' discourse by using predict-

able, recurrent patterns that can be rapidly produced and are thus potentially stereotypes.

4) Putting the power to manage linguistic diversity in the hands of technology and its masters.

With such culturally alarming prospects, we must take a watchful and keenly critical stance, asking ourselves two questions:

1) How should minority language cultures and speech communities react to this creeping centralisation, especially those that are numerically and culturally most fragile, and geographically and culturally most peripheral, where communication usually takes place at close hand?
2) What function can be performed by artificial intelligence (AI), now at the leading edge of digital technology and the subject of much debate, ethical and otherwise[4]? Will it continue to accentuate linguistic-cultural-economic (and hence political) centralisation, or bring it into better balance[5]? Once again, technologies which are in themselves neutral show that they can be used for good or evil[6].

In the realm of risks, we see a glaring lack of symmetry. While huge linguistic corpora are available in the major international languages, databases in the minority languages are few and far between: AI works well with the 'big languages' and, inescapably, less so with the 'little languages'. From this standpoint, we could say that 'it never rains but it pours' and AI, without a concerted effort at language planning, can only fuel more linguistic-cultural centralism on a global scale. What is needed is more and better documenting of a growing number of less widely used languages (corpus planning), an effort that could provide the users of these languages with a larger and more effective store of digital resources, starting with machine translation from and for these languages. Machine translation is perhaps the most promising opportunity for dealing successfully with linguistic diversity, as it

[4] As regards AI, suffice it to say that the ideological positions could not be more varied: some claim that 'Artificial intelligence does not exist' (Julia 2020); others that 'There is no scientific discipline that has changed the world as much as artificial intelligence' (Ganascia 2021:150). For its part, UNESCO speaks of artificial intelligence as an 'ensemble of advanced ICTS that enable "machines capable of imitating certain functionalities of human intelligence, including such features as perception, learning, reasoning, problem solving, *language interaction*, and even producing creative work"' (UNESCO 2019:10; emphasis ours).

[5] For an extensive discussion of the relationship between minority languages and AI, see Agresti 2023.

[6] On the social impact and use of AI, see OECD 2019, Kiyindou 2019. As regards ongoing research on the issue, mention should be made of the work by the FrancophoNéA Néo-Aquitain research network (https://httpfrancophonea.fr), 'Numérique' ('digital') group, coordinated by Alain Kiyindou.

would ease the competition between languages on the market and in terms of building human capital. But reaching this goal calls for investing robustly in research projects: in collecting spoken and written forms of language X in the field, in orthographic standardisation (long a deeply felt and highly sensitive problem for minority languages), preferably 'polynomic' (Marcellesi *et al*. 2003), or in other words accounting for diatopic variation[7]; in digitalisation and building freely accessible linguistic corpora, preferably in multimedia form[8]; and in developing minority language spell checkers and writing aids for modern communication devices such as smartphones and tablets, as the latter can contribute to 'dusting off' a minority language's image and thus improve its status at the level of social representation and, consequently, at the level of use and transmission (Strubell 1999).

References

Agresti Giovanni (2023). « Intelligence artificielle et langues minoritaires : du bon ménage? Quelques pistes de réflexion ». In: Rachele Raus (cur.) *et al.*. *De Europa Special Issue. 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*. Torino, Milano: Università degli Studi di Torino, Ledizioni LediPublishing, 47-68. https://www.collane.unito.it/oa/items/show/132.

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

Djordjević-Léonard Ksenija (2018). « Linguistes, activistes et locuteurs : trois terrains croisés (vepse, tabarquin, croate molisain) ». *Études finno-ougriennes*, 49-50. DOI: https://doi.org/10.4000/efo.9951

Francis (2015). *Encyclical Letter Laudato si' of the Holy Father Francis on Care for Our Common Home*, Rome: Libreria Editrice Vaticana.

Ganascia Jean-Gabrielle (2021). « Intelligence Artificielle : Des Big-Data au Cerveau». In *Les signatures neurobiologiques de la conscience. Neurobiologie fonctionnelle, phénomènes de conscience, cognition, automates « intelligents », éthique*. Les Ulis: EDP Sciences, 147-162.

[7] The polynomic approach is often fundamental. In minority (and especially ultra-minority) language areas, there is in general no capital or academy that can set a single standard, as a state can for its national language. Establishing a standard that can seem unfamiliar, unauthentic, and thus artificial to the speakers of that minority language is too high a risk, and should be avoided.

[8] An example of a multimedia archive of the oral memory of European mountain communities is the Tramontana Network project (https://www.retramontana.org/), which hopes to make a significant contribution to creating linguistic and ethnographic corpora for 'peripheral' and highly conservative areas.

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 (a cura di). *Représentations sociales des langues et politiques linguistiques. Déterminismes, implications, regards croisés. Actes du Premier Congrès mondial des droits linguistiques, Vol. I*. Roma: Aracne, 269-286.

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

Kiyindou Alain (2019). *Intelligence artificielle. Pratique et enjeux pour le développement*. Parigi: L'Harmattan.

Le Coadic Ronan (2010). « Diversité, liberté, vitalité ». In: Giovanni Agresti, Mariapia D'Angelo (a cura di). *Rovesciare Babele. Economia ed ecologia delle lingue regionali o minoritarie*. Roma: Aracne, 51-72.

Marcellesi Jean-Baptiste, Bulot Thierry, Blanchet Philippe (a cura di) (2003). *Sociolinguistique. Epistémologie, Langues régionales, Polynomie*. Parigi: L'Harmattan.

OCDE (2019). *Considérations de politique publique. L'intelligence artificielle dans la société*. Parigi: Éditions OCDE. https://doi.org/10.1787/b7f8cd16-fr.

Poggeschi Giovanni (2015). «La mediazione linguistica e culturale come strumento esemplare per la vigenza dei diritti linguistici di prima specie». *Lingue e Linguaggi*, 16, 435-443. DOI: 10.1285/i22390359v16p435

Strubell Miquel (1999). «From Language Planning to Language Policies and Language Politics». In: Pater J. Weber (a cura di). *Contact + Confli(c)t: language planning and minorities*. Bonn: Dummler, 237-248.

UNESCO (2019). *Steering AI and Advanced ICTs for Knowledge Societies*. Parigi: Unesco.

R3

Corpora as resources for digital equality between official EU languages

Federico Gaspari
Università di Napoli "Federico II"

The following pages will discuss the role of corpora as resources for achieving equality between the official languages (and, consequently, between all citizens) of the European Union. From the beginning, multilingualism has been one of the constitutive features of the EU's cultural, social and political identity, and its continuing importance has been forcefully reasserted in the resolution on language equality in the digital age approved with a large majority by the European Parliament in September 2018 (European Parliament 2018), partly in response to the growing concerns about the weakening of certain European languages and the risk of their death and extinction (see for example Moseley 2010; Rehm, Uszkoreit 2012; Kornai 2013; Ceberio Berger *et al.* 2018).

Against this backdrop, where justified alarm vies with avid hopes for protecting and promoting multilingualism in Europe, we will focus on why electronic corpora are essential resources for ensuring that European languages enjoy equal digital vitality, not on the basis of an abstract egalitarian principle of interest only to scholars and linguists, but to guarantee the wellbeing and relevance of all communities of speakers in all spheres: educational, cultural, social, political and economic. Today's artificial intelligence (AI) based language technologies require mono- and/or multi-lingual digital data: from spell checkers to machine translation tools, up to speech synthesis and recognition systems, AI's advances and advantages can be exploited only if large quantities of high quality electronic language data are available for the areas and domains where the technologies are used (Vetere 2023).

Two international projects funded by the European Commission, the *European Language Grid*[1] (ELG) (Rehm 2023) and *European Language Equality*[2] (ELE) (Rehm *et al.* 2022; Rehm, Way 2023) (ELE; Rehm *et al.* 2022; Rehm, Way 2023) have amassed data and language tools and consulted extensively with experts and representatives of all of Europe's speech communities, dealing not only with scholars and linguists, but also with industry and market stakeholders, users, consumers, activists, politicians and decision-makers from the EU institutions, Member States, regional administrations and local authorities to lay the groundwork for an ambitious programme for achieving digital equality between all European languages. This massive effort demonstrated that the availability of electronic corpora is a decisive factor for levelling up all European languages' prospects for vitality and sustainability in the digital age.

[1] https://live.european-language-grid.eu

[2] https://european-language-equality.eu

A point that emerged from the analysis of currently available language resources is that the situation is one of clear inequality and stark imbalance in favour of a very few dominant languages, while most others are at a disadvantage. ELG and ELE have developed an interactive online dashboard[3] where the user can browse through dynamic graphs based on regularly updated data to view the current level of digital support for all European languages and compare their digital readiness, computed for the various types of language tools, resources and applications in the ELG Catalogue and groups thereof. The availability of electronic corpora for the EU's 24 official languages is indicated in Figure 1[4], showing the total number for all types of corpora (panel 1), as well as the numbers of monolingual corpora (panel 2), bilingual corpora (panel 3) and multilingual corpora (panel 4).

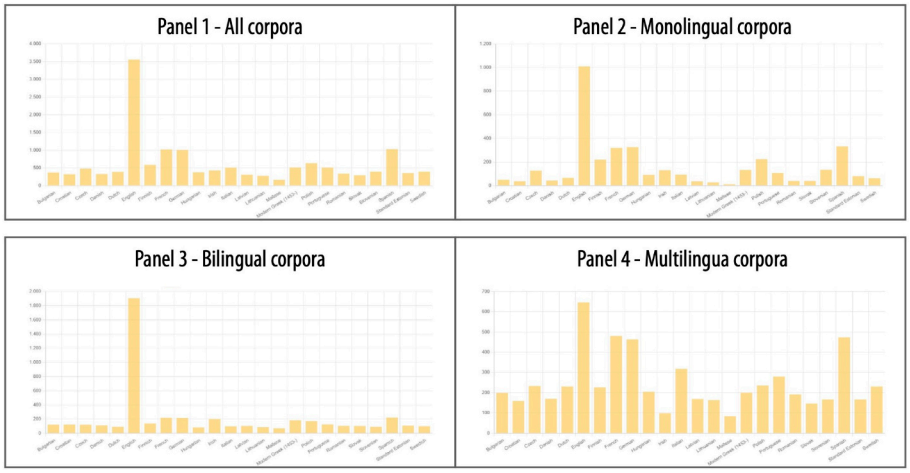

Figure 1. Availability of electronic corpora for the official EU languages in the ELG Catalogue (Source: https://live.european-language-grid.eu/catalogue/dashboard)

While the privileged position of certain languages is to be expected in view of the size of their community of speakers and their widespread international use in Europe and beyond (with English, Spanish, French and German standing out for the quantity of corpora), there are also a number of situations that are as surprising as they are worrying: first, the enormous gap in terms of the availability of corpora of all kinds between English and the other official EU languages, including those that are relatively well-resourced. Second, official languages used in some of the largest European countries (Italy and Poland, for instance) have

[3] The dashboard is available at https://live.european-language-grid.eu/catalogue/dashboard

[4] The graphs shown in Figure 1 are taken from the ELG/ELE dashboard and are based on data as of 31 May 2023. They refer only to the official EU languages, but the dashboard can also be used to visualise and compare up-to-date data for some seventy-odd other recognised and protected regional, minority or co-official European languages (Council of Europe 1992), some of which are spoken by communities of only a few thousand people.

few corpora compared to the size of their populations. Moreover, corpora with data in two or more languages chiefly embrace the dominant languages, starting with English, and are much less likely to include less-resourced languages. Lastly—and limiting ourselves for the sake of brevity to the major inequalities, digital corpora of all the types considered in the ELG are conspicuous for their absence in most of the official EU languages. As can be seen from the online dashboard, many of the official languages are in fact not much better off in this respect than the 70 or so regional, minority or co-official European languages covered by the ELG Catalogue.

It should be borne in mind that this brief overview is based entirely on the number of corpora for the official EU languages in the ELG Catalogue, without considering their size, actual quality or diversity in terms of data type (i.e., text only and/or also oral and/or video); likewise, it does not take the domains and text categories they cover into account (for a further discussion of the importance of more varied and innovative types of corpora, see Gaspari 2022: 50*ff*). Unsurprisingly, closer scrutiny of these factors shows that corpora availability is even more skewed in favour of English, followed—at a considerable distance—by the trio of languages mentioned earlier, while all the other official EU languages lag far behind (in this connection, see also Vetere 2023).

As we have seen from empirical data validated by the communities of experts consulted by the ELG and ELE projects, there are stark inequalities between the official EU languages as regards the availability of digital corpora, which are essential if we are to benefit from the latest advances that AI has brought in the development of language technologies. Far from being an abstract question divorced from real life and of interest only to scholars of linguistics and technology, these inequalities touch directly on Europe's citizens as members of their respective communities of speakers. Indeed, the gaps we have discussed reflect, and at the same time aggravate, disparities in the current support for the speech communities of Europe in the digital age and their future prospects for prosperity, not least as regards their educational, cultural, social, political and economic progress. These asymmetries have an impact on the internal relationships among European citizens and, from a broader perspective, on relationships between countries in the EU and beyond its borders in an increasingly globalised and interconnected world where there is mounting pressure to knuckle under to dominant linguistic-cultural models—pressure that is also exerted through technological supremacy.

In conclusion, there is an urgent need for all European citizens to be more aware of the issues presented by digital language equality and its potential for driving growth. There is an equally pressing need for courage and farsightedness in promoting well-funded R&D projects on the part of politicians and decision-makers, starting from those at Europe's institutions and extending to the Member States, the regional administrations and local authorities who have their community's linguistic rights and sociocultural identities at heart.

References

Ceberio Berger Klara, Gurrutxaga Hernaiz Antton,Baroni Paola, Hicks Davyth, Kruse Eleonore, Quochi Valeria, Russo Irene, Salonen Tuomo, Sarhimaa Anneli, Soria Claudia (2018). *Digital Language Survival Kit: The DLDP Recommendations to Improve Digital Vitality*. Bruxelles: European Commission; Erasmus+ Programme. www.dldp.eu/sites/default/files/documents/DLDP_Digital-Language-Survival-Kit.pdf

Council of Europe (1992). *European Charter for Regional or Minority Languages*. Strasbourg: Council of Europe. https://rm.coe.int/1680695175

European Parliament (2018). *European Parliament resolution of 11 September 2018 on language equality in the digital age (2018/2028(INI))*. Bruxelles : European Parliament. http://www.europarl.europa.eu/doceo/document/TA-8-2018-0332_EN.html

Gaspari Federico (2022). "Expanding the Reach of Corpus-Based Translation Studies: The Opportunities that Lie Ahead". In: Sylviane Granger e Marie-Aude Lefer (a cura di). *Extending the Scope of Corpus-Based Translation Studies*. London: Bloomsbury, 42-63.

Kornai András. (2013). "Digital Language Death". *PLoS ONE* 8(10): e77056. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0077056

Moseley Christopher, Nicolas Alexandre (a cura di) (2010). *Atlas of the World's Languages in Danger*. 3a edizione. Parigi: UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000187026 (accessed 23/07/2023)

Rehm Georg (a cura di) (2023). *European Language Grid: A Language Technology Platform for Multilingual Europe*. Berlino: Springer.

Rehm Georg, Gaspari Federico, Rigau German, Giagkou Maria, Piperidis Stelios, Grützner-Zahn Annika, Resende Natalia, Hajič Jan, Way Andy (2022). "The European Language Equality Project: Enabling digital language equality for all European languages by 2030". In: Željko Jozić e Sabine Kirchmeier (a cura di). *The Role of National Language Institutions in the Digital Age*. Budapest: Hungarian Research Centre for Linguistics, 17-47.

Rehm Georg, Uszkoreit Hans (a cura di) (2012). "About META-NET". In: Georg Rehm, Hans Uszkoreit (eds). *The Danish Language in the Digital Age*. *White Paper Series*. Berlin: Springer. DOI: https://doi.org/10.1007/978-3-642-30627-3_10

Rehm Goerg, Way Andy (a cura di) (2023). *European Language Equality: A Strategic Agenda for Digital Language Equality*. Berlino: Springer.

Vetere Guido (2023). "Elaborazione automatica dei linguaggi diversi dall'inglese: introduzione, stato dell'arte e prospettive". In: Rachele Raus *et al*. *De Europa Special Issue. 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*. Torino, Milano: Università degli Studi di Torino, Ledizioni LediPublishing, 69-87. https://www.collane.unito.it/oa/items/show/132

R3

The Italian language: cushioning language varieties from the impact of artificial intelligence

Chiara Russo
Università degli Studi di Catania

By virtue of its dynamic, complex nature, human language is heavily influenced by social, cultural and geographic factors and variables arising in a particular spatiotemporal setting. Analysing the varieties in a speech community reveals a staggering array of different forms and manifestations that confirm the universal proposition that every language presents a certain internal variability (Berruto, Cerruti 2019).

Every speaker, in producing linguistic forms and constructs along the four dimensions of variability[1], departs from the standard[2] by modifying and adapting their language system to meet everyday practical needs, thus gradually expanding the concept of what constitutes the norm[3].

The rapid growth of artificial intelligence (AI) poses a threat to this diversified language system. Language varieties, which global industry regards as a nuisance, risk being cancelled in pursuit of that 'ideal world' where everyone speaks the same language[4]. Even if it were possible to develop AI capable of communicating in all the world's languages, there would still be no way of dealing with all the many internal varieties. Take, for example, the case of personal digital assistants: programmed to use a standardised language, they limit creativity and expressiveness in interactions.

On the Italian language scene, the spread of AI has raised questions concerning the preservation of linguistic diversity and uniqueness. Consequently, it is essential to take a global, integrated approach that engages local communities and diverse stakeholders in creating and validating models that can understand and use the distinctive linguistic and cultural features found throughout the country[5].

Though the linguistic features of standard Italian are described and represented in a variety of language corpora, this cannot be said of all of the other varieties, widely used among communities of speakers but still poorly represented on the technological front. The databases and corpora available for

[1] Diatopic, diastratic, diaphasic and diamesic variability.

[2] In linguistics, the term Standard Italian denotes a specific variety which speakers take as a model, and which is thus free from social or regional connotations (Marzullo M. 2005).

[3] In the evolution of the Italian language, both the spoken and the standard variety have had a fundamental role: standard Italian has adapted to everyday needs, while spoken Italian has exerted pressure on the structures of the written language, leading to a series of structural changes in the language system.

[4] In the world of industry, there is a strong temptation to use English as a 'pivot language', relying entirely on machine translation.

[5] This is true for all European languages, which are equally rich in internal variety.

written and spoken varieties of Italian show problems of representativeness, flexibility and scalability[6].

Protecting the many language varieties from the effects of a massive use of artificial intelligence is a daunting challenge. But a number of measures and strategies can be adopted to overcome its complexities.

One fundamental measure consists of collecting high quality representative data so that AI can be trained on a diversified dataset. This would help prevent language models based on one or a few varieties, which could jeopardise the ability to reflect the real complexity of the language and the local cultures that use it. In view of the difficulty of this task, it is to be hoped that national or regional specifications can be introduced to define models ensuring high accuracy.

It is also of fundamental importance that linguistic experts be involved in continually monitoring and updating the collected data. Summarising, Italy's language diversity is a precious cultural heritage, and AI, if correctly used, can be an excellent means of enhancing, preserving and protecting varieties, not least through the creation of digital language resources such as teaching and learning apps[7].

References

Berruto Gaetano (2012). *Sociolinguistica dell'italiano contemporaneo*. Roma: Carocci.

Berruto Gaetano, Cerruti Massimo (2019). *Manuale di Sociolinguistica*. Novara: De Agostini Scuola SpA.

De Mauro Tullio (2021). *Storia linguistica dell'Italia repubblicana dal 1946 ai nostri giorni*. Bari-Roma: Laterza.

De Mauro Tullio. (2020). *Storia linguistica dell'Italia unita*. Bari-Roma: Laterza.

[6] The list of currently available databases, corpora and text archives is available at: https://accademiadellacrusca.it/it/contenuti/banche-dati-corpora-e-archivi-testuali/6228

[7] An interesting project in this connection is the *Lahjoita puhetta* (Donate Speech) large-scale corpus of spoken Finnish sponsored by the Finnish Broadcasting Company Yle and the Finnish State Development Company Vake Oy with the contribution of experts from the University of Helsinki. Winner of the Prix Europa Best European Digital Audio Prize in 2021 and open to the entire population, the project collected short samples of spontaneous colloquial speech from volunteers to develop AI that can understand and model the characteristics of Finnish language varieties. The goal was to develop voice-controlled apps and services that can be used without problems by all speakers. See https://www.yle.fi/lahjoitapuhetta

Marzullo Mara (2005). *Etimologia e origine della parola standard*. Firenze: Accademia della Crusca. https://accademiadellacrusca.it/it/consulenza/etimologia-e-origine-della-parola-standard/154

Sabatini Francesco (1985). *"L'italiano dell'uso medio: una realtà tra le varie linguistiche italiane"*. In: Gunther Holtus, Edgar Radtke (Hrsg.). *Gesprochenes Italienish in Geschichte und Gegenwart*. Tübingen: Narr.

R3

Multilingual corpora that reflect the range of diatopic variation: the case of Quebec

Valeria Zotti
Università di Bologna

Over the last sixty years, numerous studies have addressed the need to recognize the linguistic legitimacy of the French spoken in the countries of francophonie and, in particular, the French spoken in francophone Canada[1]. This battle for the recognition of the diatopic variation of the French language is now being continued by other linguists, including Nadine Vincent and Wim Remysen (2016), authors of the first dictionary based entirely on text corpora and conceived in Quebec for all francophones and francophiles interested in an 'open' description of French[2].

While we can now say that the linguistic legitimacy of Quebec French, long looked down upon as a 'deviant' version of the French of France, has been acknowledged on the lexicographic front (Zotti 2012) both transnationally and nationally (take, for example, the inclusion of numerous 'francophonisms' in the latest editions of the Le Robert and Larousse French dictionaries, see Cormier *et al.* 2013), this is still far from being true in the world of machine translation. A sample survey (Zotti 2021) of Québécois literary texts[3] found that there are almost no Québécois diatopic variations of French in the corpora employed by the major free online machine translation tools, Google Translate and DeepL. The Canadian Parliament's bilingual texts are an important French/English database, but not for other language pairs such as French/Italian, and in any case coverage is limited to the legal and administrative domain. The proportion of untranslated or incorrectly translated Québécois French words is thus quite high, both for lexematic variants (forms used only in Québécois French) and for semantic variants (forms existing in standard French with a different meaning). Because of these lacunae or errors, there is a risk that incorrect translations will spread in many areas of knowledge, which would be especially dangerous in the sphere of education.

A few words are thus in order about the data used to train machine translation tools. Though the literature on gender bias (Temmerman 2021) and the so-called 'algorithm biases' (Council of Europe 2019) is now fairly large, far fewer scholars have conducted studies and tests of MT performance in translating texts in 'poorly endowed' languages (Le 2019), or in other words, 'regional variants' and languages for which there are few of the re-

[1] Take, for example, the studies by Claude Poirier, for many years director of the *Trésor de la Langue Française au Québec* research laboratory at the Université Laval in Québec, and by his students, of whom mention should be made of Louis Mercier and Hélène Cajolet-Laganière.

[2] USITO, see https://usito.usherbrooke.ca/

[3] Poetry: Gaston Miron; prose: short stories and novels of the "terroir", with strong sociocultural connotations and an abundance of amerindianisms and realia.

sources such as parallel corpora that are essential for the development of high performance natural language processing systems. This links up with another concern about the rampant growth of plurilingual translation resources such as BabelNet based on web crawling, i.e., the automated extraction of unverified data from the web, which inevitably generate imprecise translations in specialised areas, together with cultural and linguistic stereotypes.

Statistical machine translation systems' inability to produce diversified output and their tendency to reproduce the more common 'patterns' and ignore the less frequent is all too obvious. The problem of diversified output has also been noted in the neural models for activities involving language generation (e.g., ChatGPT). For both learning models, in fact, producing accurate translations was the main goal, but maintaining lexical richness and creating diversified output were not seen as priorities. For a simple but particularly telling illustration concerning diatopic variability in French, if we ask an MT tool to give us the French translation of an English syntagma such as 'the mayor of [city]', the machine translation system must produce both 'le *maire* de [Lyon/Paris/Bordeaux]' and 'le *bourgmestre* de [Bruxelles/Liège/Anvers]', given that the term 'bourgmestre' is used in Belgium. Similarly, '*la mairesse* de ...', the feminine form of 'maire' used in Quebec, as opposed to the form '*une maire*" used in France for the female gender, should be translated with the corresponding English 'the mayoress of...', rather than with the more widespread masculine form 'the mayor' (see *Google Translate* e *DeepL*, last accessed 14 April 2023).

We believe that the tendency to generalise shown by today's machine translation systems can be a serious problem, leading to a monolithic vision of the linguistic complexity of human societies. Overgeneralising from input and giving even greater prominence to dominant forms could not only bring a loss of lexical choice, but could also be an underlying cause of worsening social prejudices against linguistic minorities. As the Council of Europe (2019) has pointed out, 'The lack of diversity and inclusion in the design of AI systems is [...] a key concern: instead of making our decisions more objective, they could reinforce discrimination and prejudices by giving them an appearance of objectivity'. AI is a strategic technology that offers many benefits for citizens, companies and society as a whole, provided it is ethical, sustainable, human-centric and respects fundamental rights and values. And provided that it allows for multilingualism and plurilingualism (Temmerman 2021).

The work of specialists in corpora linguistics to improve neural machine translation models by using 'quality' monolingual data (Sennrich *et al*. 2016) that respect internal and external multilingualism is essential, and preserving diversity, though not hitherto considered a priority in this area, is also important in machine translation.

References

Bentivogli Luisa, Bisazza Arianna, Cettolo Mauro, Marcello Federico (2016). "Neural versus Phrase-Based Machine Translation Quality: a Case Study". In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin: EMNLP, 257–267.

Conseil de l'Europe (2019). *Combler l'écart. Comment garantir les droits de l'homme pour tous. Compilation du Carnet des droits de l'homme*. Articles publiés en 2018 et 2019 par Dunja Mijatović, Commissaire aux droits de l'homme du Conseil de l'Europe. Bruxelles: Conseil de l'Europe.

Cormier Monique C., Francœur Aline, Boulanger Jean-Claude (a cura di) (2003). *Les dictionnaires Le Robert : Genèse et évolution*. Montréal: Presses de l'Université de Montréal. http://books.openedition.org/pum/13849

Le Ngoc Tan (2019). *Traduction automatique pour une paire de langues peu dotée*. Thèse de Doctorat en informatique cognitive. Montréal: Université du Québec.

Mercier Louis, Cajolet-Laganière Hélène (a cura di) (2004). « *Français du Canada - Français de France VI* » *Actes du Sixième colloque international d'Orford, du 26 au 29 septembre 2000* (*Canadiana Romanica*, 18). Tübingen: Niemeyer, 365.

Sennrich Rico, Haddow Barry, Birch Alexandra (2016). "Improving Neural Machine Translation Models with Monolingual Data". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (ACL). Berlin: Association for Computational Linguistics, 86–96. Doi: 10.18653/v1/P16-1009

Temmerman Rita (2021). « La créativité plurilingue, un obstacle pour l'IA? », intervento al convegno *Linguistic Rights and Language Varieties in Europe in the Age of AI*, Università di Torino, 21 aprile 2021. https://www.jmcoe.unito.it/content/linguistic-rights-and-language-varieties-europe-age-ai

Wim Remysen, Nadine Vincent (a cura di) (2016). *La langue française au Québec et ailleurs: Patrimoine linguistique, socioculture et modèles de référence*. Frankfurt: Peter Lang.

Zotti Valeria (2021). « Intelligence Artificielle et diversité linguistique : quel regard sur la langue française au Québec ? ». Intervento al Convegno internazionale *'Regards croisés sur le Québec et la France'*. Centro Interuniversitario di Studi Quebecchesi (Trento, 20-22 maggio 2021). USITO: https://usito.usherbrooke.ca

Zotti Valeria (2019). « Ressources numériques pour la traduction des mots désignant des *realia* ». *Etudes de Linguistique Appliquée*, 194, 227-246.

Zotti Valeria (2012). « La légitimité linguistique du français québécois passe-t-elle par la reconnaissance lexicographique d'une langue et d'une culture autres? La contribution d'un dictionnaire différentiel bilingue français québécois / italien ». In: *Lexiques Identités Cultures*. Verona : QuiEdit, 471–489.

Multilingual corpora and special languages: preserving diatopic variation

Marta Muscariello
Università IULM di Milano

In current IT resources and artificial intelligence technologies for language and text processing, the balance tilts heavily towards English. Most open-source software and data, often the only kinds small and medium enterprises can afford, are in English, leaving all the other languages without coverage (Vetere 2022). Consequently, there is a barrier to accessing the most advanced services technology can offer, a barrier resulting simply from the fact of belonging to a non-anglophone speech community. This situation limits the specificity and, above all, the quality and development potential of the markets tied to languages other than English. They are thus very much in the economic and political minority, with the risk that the entire European market will find itself unable to create innovation in many key growth areas.

One of the linguistic areas where automated data processing now has the greatest impact is that of languages for specific purposes, indispensable tools for effective communication in economic sectors and political life (Morresi 1998) which are not without repercussions on linguistic contact and even the evolution of the styles and expressive means of common languages (Cortelazzo 1994). Invariably, technological innovation has brought new coinages of specific terms. Indeed, we can say that the status of a country's special languages on the international scene reflects its capacity for growth and investment in a given sector of the economy. Take, for example, the special language of ICT, where English holds sway, or, conversely, the special language of wine connoisseurship, which bespeaks the leadership that Italy and France enjoy on the global market.

Comparing multilingual corpora thus enables us to monitor the situation of special languages on the international level, identifying the translation equivalents and, at the same time, the differences and specificities of each local market. Machine learning in this field is particularly challenging because ideally it should use the so-called onomasiological approach (or in other words, learning the translation equivalents from language to language starting from concepts) rather than relying entirely on a semasiological approach (translating with meanings as the starting point). From this perspective, the precision that special languages call for means that machine translation must be followed by careful post-editing.

A particularly sensitive field in this respect is that of law, where there is a continual tug of war between the demand for international harmonisation and the need to ensure the specificity of each national system. Creating specific corpora for each country (and thus having separate corpora for countries using the same lan-

guage), makes it possible to compare the linguistic and conceptual situation of jurisprudence and law both in Europe and vis-à-vis the broader global scene. Each country's legal language can thus be delineated, both in its own specific features and in its dynamic relationship with the supranational European legal system now taking shape (Rossini Favretti 1999; Felici, Mori 2019).

References

Cortelazzo Michele (1994). *Le lingue speciali: la dimensione verticale*. Padova: Unipress.

Felici Annarita, Mori Laura (2019). "Corpora di italiano legislativo a confronto: dall'Unione Europea alla Cancelleria svizzera". In: Bruno Moretti, Aline Kunz, Silvia Natale e Etna Krakenberger (a cura di). *Le tendenze dell'italiano contemporaneo rivisitate. Atti del LII Congresso Internazionale di Studi della Società di Linguistica Italiana (Berna, 6-8 settembre 2018)*. Milano: Officinaventuno.

Morresi Ruggero (a cura di) (1998). *Le lingue speciali. Atti del convegno di studi, Università di Macerata, 17-19 ottobre 1994*. Roma: Il Calamo.

Rossini Favretti Rema (1999). "Equivalenze traduttive in corpora giuridici multilingue". *Quaderni di Libri e Riviste d'Italia*. Roma: Istituto Poligrafico e Zecca dello Stato, 47-66.

Vetere Guido (2023). "Elaborazione automatica dei linguaggi diversi dall'inglese: introduzione, stato dell'arte e prospettive". In: Rachele Raus *et al*. *De Europa Special Issue. 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*. Torino, Milano: Università degli Studi di Torino, Ledizioni LediPublishing, 69-87. https://www.collane.unito.it/oa/items/show/132

RECOMMENDATION 4

Europe must invest in developing language and computer technologies that are truly Made in EU

R4

Technological independence and cultural diversity in European artificial intelligence

Moreno La Quatra
Università degli studi di Enna "Kore"

The burgeoning growth of artificial intelligence has led to an ever-greater dependence on AI resources developed by tech giants hailing mostly from outside Europe. This calls Europe's technological independence into question, and also raises concerns about the security of European users' data. In such a situation, passively using AI natural language processing models carries a risk of linguistic and cultural homogenization. Developing autonomous and independent AI resources in Europe could help preserve the continent's diverse languages and cultures, protect its citizens' data, and promote technological independence.

The artificial intelligence industry is a major strategic resource for many countries. In Europe, much of the AI technology used for natural language processing is supplied by large producers headquartered outside Europe's borders. As a result, Europe is becoming increasingly dependent on outsourced tech. This entails a number of risks, as imported technologies might not meet Europe's distinctive needs or comply with its personal data protection legislation. In addition, foreign technology producers may be unwilling to share resources and knowhow with the European authorities, making it difficult for the EU to pursue its technological independence.

It should be pointed out that many of the artificial intelligence models available today were developed chiefly in English or adapted from English for other languages. Far fewer models are available for European languages than for English[1]. Developing new AI models tailored to EU languages is thus a challenge, and failure to meet it would reduce growth potential and hamper Europe's ability to compete globally in AI.

Accordingly, investing in homegrown artificial intelligence resources for European languages is crucial. A major effort in this area is the European Language Grid (ELG)[2] project, a centralized platform for collecting language data from all over Europe. A consensus strategy for putting this data to concrete use would enable Europe to develop customised language models meeting the needs of the European population and complying with EU personal data protection legislation.

In addition to solving the problems stemming from technological dependence, developing autonomous AI resources would preserve Europe's cultural and linguistic diversity. Using AI models uncritically and training them on unlabelled corpora could lead to homogenisation in this respect, threatening language

[1] As of April 2023, the huggingface.co platform hosted 14,819 models for English, 618 (4%) for Italian, and 1,303 (9%) for French.

[2] https://live.european-language-grid.eu/

variety in the EU. Access to independent resources could help safeguard cultural diversity and Europe's minority languages, which technologies from developers outside Europe are all too prone to ignore.

The importance of linguistic diversity in European artificial intelligence

Europe's linguistic and cultural diversity presents a unique challenge for developing artificial intelligence solutions. At least 24 official languages are recognized by the European Union, in addition to many other minority and regional languages. This means that the AI tools used in Europe must be able to understand an extremely wide range of languages if they are to guarantee that linguistic variety is equitably represented.

It is important to emphasise that many of the languages spoken in the European Union have features that are not present or significant in other languages. The gendered nature of French and Italian is an example. To meet this challenge, the artificial intelligence resources developed for European languages must take such language-specific features into account (La Quatra, Cagliero, 2022; Sarti, Nissim 2022; Martin *et al*. 2020). Here, it should be borne in mind that inclusivity in language is not simply a question of dealing with grammatical gender, but also concerns the representation of specific groups and minorities such as the disabled, LGBTQ+ individuals and people of colour. Using tools based on inclusive models adapted to the specificities of European languages can contribute to promoting diversity and inclusion in all settings, from corporate communication to education and politics (Attanasio *et al*. 2021; Raus *et al*. 2022).

Developing European artificial intelligence models is an answer to this multilingual challenge, making it possible to create highly customised tools meeting the specific linguistic needs of the European population. Such language models would be able to analyse, recognise and generate text preserving each region's and each country's distinctive linguistic and cultural nuances.

Unlabelled corpora: risks and challenges in language models

To be effective, language models must be trained on large datasets (Raffel *et al*. 2020). Once trained, natural language models

show a remarkable capacity for generalisation and can handle a wide range of tasks effectively. In this scenario, the data used for training purposes are often obtained from unlabelled web resources such as social media, blogs and websites.

The use of unlabelled corpora entails a number of risks. First, unlabelled data may be affected by implicit biases, which can then propagate in the models. These biases can involve questions of gender, race, ethnicity and so forth, and can lead to discriminative and unpredictable output (Dodge *et al.* 2021). Unlabelled corpora can contain hate speech or other inappropriate content that can be incorporated in the language models. Second, when unlabelled corpora are used it is not possible to determine the variety or source region of the text used during training. This can result in language models that do not reflect the linguistic and cultural variety needed in a European setting.

Preventing these risks will require the use of labelled culture- and language-specific training resources. Though this will entail a major effort to collect and tag masses of data, it can lead to more inclusive artificial intelligence models that are better suited to the European setting. In addition, using labelled corpora can help in mitigating implicit bias and halting the spread of inappropriate content (Meade *et al.* 2022).

European artificial intelligence: challenges and opportunities for the future

Investing in the creation of European artificial intelligence resources is a crucial step towards maintaining Europe's technological independence and guaranteeing continuing advances in this fast-evolving field. However, achieving this goal will require a long-term commitment to developing technological skills and formulating European policies that guarantee the utmost protection for citizens' privacy.

From a European perspective, the quality of the data used to train artificial intelligence models is another crucial aspect. To preserve language variety and prevent the risk of discrimination, it is important that this data be representative and inclusive. Collaboration between universities and government is essential for the development of European artificial intelligence. Universities play a vital part in building advanced technological skills, while governments can promote policies encouraging R&D in artificial intelligence, as well as its adoption across industry.

References

Attanasio Giuseppe, Greco Salvatore, La Quatra Moreno, Cagliero Luca, Tonti Michela, Cerquitelli Tania, Raus Rachele (2022). "L'analyse du discours et l'intelligence artificielle pour réaliser une écriture inclusive: le projet E-MIMIC. *SHS Web of Conferences*, Vol. 138, 01007, 1-15. DOI: https://doi.org/10.1051/shsconf/202213801007

Attanasio Giuseppe, Greco Salvatore, La Quatra Moreno, Cagliero Luca, Tonti Michela, Cerquitelli Tania, Raus Rachele (2021). "E-MIMIC: Empowering Multilingual Inclusive Communication". In: *2021 IEEE International Conference on Big Data (Big Data)*. 4227-4234. DOI: 10.1109/BigData52589.2021.9671868

Raffel Colin, Shazeer Noam, Roberts Adam, Lee Katherine, Narang Sharan, Matena Micheal, Zhou Yanqi, Li Wei, Liu Peter J. (2020), "Exploring the limits of transfer learning with a unified text-to-text transformer". *Journal of Machine Learning Research*, Vol. 21,1-67.

Dodge Jesse, Maarten Sap, Marasović Ana, Agnew William, Ilharco Gabriel, Groeneveld Dirk, Mitchell Margaret, Gardner Matt (2021), "Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus". In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana: Association for Computational Linguistics, 1286–1305.DOI:10.18653/v1/2021.emnlp-main.98

La Quatra Moreno, Cagliero Luca (2023). "BART-IT: An Efficient Sequence-to-Sequence Model for Italian Text Summarization". *Future Internet*, 15, 15. DOI: https://dx.doi.org/10.3390/fi15010015

Louis Martin, Muller Benjamin, Ortiz Suárez Pedro Javier, Dupont Yoann, Romary Laurent, de la Clergerie Éric, Seddah Djamé, Sagot Benoit (2020). "CamemBERT: a Tasty French Language Model". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, 7203–7219. DOI: 10.18653/v1/2020.acl-main.645

Meade Nicholas, Poole-Dayan Elinor, Reddy Siva (2022). "An Empirical Survey of the Effectiveness of Debiasing Techniques for Pre-trained Language Models". In: *Vol.1: Long Papers* di *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*. Dublin: Association for Computational Linguistics, 1878–1898.

Sarti Gabriele, Nissim Malvina (2022). "IT5: Large-scale Text-to-text Pretraining for Italian Language Understanding and Generation". arXiv preprint arXiv:2203.03759.

Alida Maria Silletti
Università di Bari

Title of proposed action
Machine transcription to promote and preserve diatopic variation in the EU

Proposed action
Training secondary school and university educators and learners to be able to review (in the case of educators) and revise (learners) machine-transcribed audiovisual documents in an EU 'minority' or 'ultra-minority' language dealing with political/institutional communication at mother-tongue or foreign/second language level, enabling them to build the greater understanding of these languages needed to develop and implement oral and transcribed corpora.

Aims
1. Preserving Europe's diatopic variation and multilingualism by creating 'minority' or 'ultra-minority' language corpora—or implementing any such corpora that already exist—to preserve these languages and for use in linguistic investigation, viz.
2. in developing machine transcription software, and
3. as a starting point for learners to reflect on the specific features of these languages emerging from machine transcription of political/institutional audiovisuals, and to solve morphosyntactic and semantic problems during post-editing.

Details of the proposed action
As shown by Basque as an example of a 'minority' EU language (Sarasola *et al*. 2023)—here we will adopt the terms 'minority language' and 'ultra-minority language' proposed by Agresti (2023)—AI language technologies can be drivers of revitalisation for languages whose repertoire is shrinking with the drop in the number of active and passive speakers using them in settings that are to varying degrees institutional and formal. Among the measures that can counteract this phenomenon, training educators who can teach 'minority' or 'ultra-minority' languages at mother tongue or second/foreign language level in schools and universities is an essential resource, as France's programmes for Occitan have shown (Verny 2009). Another measure consists of developing generalist and specialised corpora covering the use of the language in certain contexts, such as the media, administration or political/institutional communication. Any effort of this kind calls for significant investments and, as Vetere (2023) has pointed out, a major trend affecting language investments in the EU is the 'anglicisation of European linguistic life', resulting in less

cultural and language diversity (Fischer, Pulaczewska 2009). Whereas Vetere (2023) sees more investment in machine translation as an effective means of preserving multilingualism, we propose to defend the EU's diatopic variation and multilingualism—in particular for the 'minority' and 'ultra-minority' languages—and put their use in political/institutional settings on a more systematic basis by leveraging the potential of machine transcription. This application of AI involves a three-step process: acoustic analysis, mapping sound frequencies to words, and analysing the word thus obtained with a language model, a pronunciation model and a phonetic model. In addition to calling for significant economic resources, this also requires an enormous amount of date in order to identify the most likely sequence of words in a given acoustic signal[1]. If the system is to yield the utterance which is most likely on the basis of the input data, it is clear that the more data is available, the lower the statistical probability of incorrect recognition will be, and consequently, the higher the performance that will be achieved. This, at least, is the situation that now holds for English, which boasts the most powerful software (Vetere 2023). It is thus equally clear that investing in this AI tool for 'minority' and 'ultra-minority' languages offers tremendous potential. Indeed, a classroom experiment conducted in Italy with intermediate-level university students of French as a foreign language (Silletti 2022) found that revising a machine transcription of an audiovisual document dealing with political/institutional communication—i.e., oral material that is more structured than spontaneous conversation—entails skills in prosody and transcription that also call for a significant effort at the grammatical, morphosyntactic and semantic level. It is thus to be hoped that machine transcription software for the EU's 'minority' and 'ultra-minority' languages can be developed, and that any such tools that already exist can be improved, thus enriching the repertoire available for these languages starting from authentic speech data. This will require major investments by META, the Multilingual European Technology Alliance (Vetere 2023) as well as by regional and local investors in order to build a European-made network of technologies that can preserve the EU's diatopic variation.

[1] https://www.journaldunet.fr/web-tech/guide-de-l-intelligence-artificielle/1501849-reconnaissance-vocale/

References

Agresti Giovanni (2023). « Intelligence artificielle et langues minoritaires : du bon ménage? Quelques pistes de réflexion ». In: Rachele Raus (cur.) *et al*. *De Europa Special Issue. 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*. Torino, Milano: Università degli Studi di Torino, Ledizioni LediPublishing, 47-68. https://www.collane.unito.it/oa/items/show/132.

Crochet-Damais Antoine (2022). "Reconnaissance vocale : définition, algorithmes et fonctionnement". *Journal du Net, 31/05/2022*. https://www.journaldunet.fr/web-tech/guide-de-l-intelligence-artificielle/1501849-reconnaissance-vocale/

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

Sarasola Kepa, Aldabe Itziar, Aranberri Nora (2023). "Enabling additional official languages in the EU for 2025 with language-centred AI". In: Rachele Raus (cur.) *et al*. *De Europa Special Issue. 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*. Torino, Milano: Università degli Studi di Torino, Ledizioni LediPublishing,91-105. https://www.collane.unito.it/oa/items show/132.

Silletti Alida Maria (2022). « La macrosyntaxe à l'épreuve de la transcription générée automatiquement : le cas des 'parenthèses'» intervento al Convegno internazionale *Franc'parler. Français parlé: données. Répresentatins, questionnements théoriques*, Università di Torino: 16-17 giugno 2022.

Verny Marie-Jeanne (2009). "Enseigner l'occitan au XXIe siècle. Défis et enjeux". *Tréma*, 31, 69-83. DOI: https://doi.org/10.4000/trema.962.

Vetere Guido (2023). "Elaborazione automatica dei linguaggi diversi dall'inglese: introduzione, stato dell'arte e prospettive". In: Rachele Raus *et al*. *De Europa Special Issue. 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*. Torino, Milano: Università degli Studi di Torino, Ledizioni LediPublishing, 69-87. https://www.collane.unito.it/oa/items/show/132

R4

Towards transparent European artificial intelligence

Giuseppe Attanasio
Università Bocconi

The race to Artificial Intelligence: a far from crowded field

As recent studies of the current forms of generative artificial intelligence[1] have shown, there can be no doubt that the quality of the output, be it a text or an image, often hinges on the size of the upstream investments in the AI tool (Bowman 2023: 1). The need for large, targeted investments and, at the same time, for skilled R&D centres, has meant that modern AI development has been entirely in the hands of a few tech giants, none headquartered in the European Union[2].

The privately-driven 'race to Artificial Intelligence'—now explicitly oriented towards building a consumer product—is both a risk and a major incentive for the development of new AI tools by entities accredited in the EU. The following pages will address two broad areas that constitute a risk for the EU, viz., the increasing lack of transparency and the social biases that AI threatens to perpetuate, and will conclude with a list of the challenges and possibilities to come.

The risks of poor transparency

Motivated by—admittedly legitimate—concerns about maintaining a possible economic edge, *all* producers of the most common generative AI tools refuse to disclose such key information as the data used for training, while the models themselves are accessible only via paid interfaces. For a number of reasons, this practice poses a series of risks for the EU and its Member States.

The European Commission's *Ethics Guidelines for Trustworthy AI*[3] emphasise the importance of transparency[4] and require that it be possible to measure parameters such as the presence of stereotyped concepts or environmental sustainability. The almost complete absence of details about the development and use of today's AI makes it impossible to abide by these guidelines. It is not possible, for example, to employ techniques stemming from Explainable AI, now a consolidated stream of research, calling for

[1] The term 'generative' AI means, generically, a technology capable of generating new and almost always original content starting from a form of input request. Examples include GPT-3 for language and DALL-E for images, where users can ask for such things as 'a sonnet about a thrush in the style of Dante', or a 'a painting of the Eiffel Tower in the style of Starry Night' by Picasso, respectively. Both GPT-3 and DALL-E were developed by the US company OpenAI.

[2] The recent ChatGPT, DALL-E, Bing, Bard, LLaMA and Claude were all developed by private companies in the United States.

[3] https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines/1.html

[4] For example, it must be possible to explain an algorithm's decision-making process to a human being.

access to the model assigning a given output to a given input. Likewise, estimating environmental impact is difficult, since neither the size of the model—which is positively correlated with CO_2 emissions from training (Strubell *et al*. 2019: 2)—or the quantity and type of training data are known. Lastly, limited access stands in the way of the new proposals for "watermarking" content to detect whether it was produced by IA[5], which require partial or complete access to the model and its functions and are crucial for verifiability.

More generally, "closed" models cannot be inspected and the proprietary owner—who does not always comply with current European legislation[6] —keeps the entire verification chain under wraps, infringing the EU guidelines' principle of independent verifiability. Verification and validation are extremely important, and all the more so when these models show themselves to be fragile and dangerous: there is already evidence that chatbots can be aggressive towards users[7], provide false information or 'hallucinations'[8], or be easily manipulated into changing their behaviour and ignoring their safeguards by means of carefully worded prompts[9].

The added complexities of these models makes it impossible to ensure that AI is *safe*: to date, there are no reliable methods for 'steering' AI to abide by principles or guidelines (Bowman 2023: 1). It follows that developing European AI is essential if this trend is to be reversed. New investments will drive new research on AI validation and safety, encouraging the evaluation of existing models as well as the construction of new 'EU-made" models where all details—from the data up to the model's components—are freely accessible to research institutions and industrial partners.

Ideologies and stereotypes perpetuated by AI

An equally pressing issue concerns the data used for training. Here, the risk is twofold.

First, the kind of data and where they come from can be problematic. It is common knowledge that the primary source of training data is the Internet. The web, however, does not reflect all the

[5] https://aiguide.substack.com/p/on-detecting-whether-text-was-generated

[6] https://www.independent.co.uk/tech/chatgpt-ban-italy-gdpr-data-protection-b2311738.html

[7] https://time.com/6256529/bing-openai-chatgpt-danger-alignment/

[8] https://cybernews.com/tech/chatgpts-bard-ai-answers-hallucination/

[9] https://www.washingtonpost.com/technology/2023/02/14/chatgpt-dan-jailbreak/

ways in which people view the world, but only the views held by those who can access it more or less all the time (Bender *et al.* 2021: 3). As a result, the outlooks, the ideologies—and the stereotypes—of a certain slice of the population can be learnt and perpetuated by a tool—AI—that reaches many more people. For example, recent studies have found that AI models that generate images will respond to the prompt 'a CEO' with an image of a man, while 'a terrorist' generates faces with Middle Eastern features (Bianchi *et al.* 2023: 4).

Were this not enough, *almost all* state of the art language corpora—or in other words, those that are large enough to train AI for language—are in English.

Constructing Anglocentric AI tools, however, intrinsically limits their potential for use with other languages in many cases**.** Nozza *et al.* (2022: 5), for example, show that models trained on a combination of English and Italian datasets detect hate speech more effectively than their monolingual counterparts. English, moreover, does not share the features of many European languages. Gender inflection in the grammar of Romance languages is an example: here, issues include the relationship between work roles and gender in machine translation (Stanovsky *et al.* 2019: 6), as well as gender inclusivity in language and translation (Attanasio *et al.*: 7, Piergentili *et al.* 2023: 8).

Constructing corpora under the supervision of entities accredited in the EU serves two purposes. The first goal is to provide resources of verifiable quality whose encoded values can be measured so that stereotypes can be mitigated before the corpora are used for training. The second and more general goal is to increase the presence of the Member States' languages to counterbalance the primacy of English and capture linguistic nuances that would otherwise be forgotten.

The coming challenges

The European Union must channel new resources into artificial intelligence to reverse the trends that have resulted in models that are neither transparent nor safe, and assign priority to English alone.

The are multiple incentives for doing so, ranging from the conclusive findings of recent cost-benefit studies, the support expressed by independent groups, and, above all, the need for safe AI complying with the guidelines laid down by the European Union.

References

Attanasio Giuseppe, Greco Salvatore, La Quatra Moreno, Cagliero Luca, Tonti Michela, Cerquitelli Tania, Raus Rachele (2021). "E-MIMIC: Empowering Multilingual Inclusive Communication". In: *2021 IEEE International Conference on Big Data (Big Data)*. 4227-4234. DOI: 10.1109/BigData52589.2021.9671868

Bender Emily M., Gebru Timmit, McMillan-Major Aangelina, Shmitchell Shmargaret (2021). "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?". In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*. New York: Association for Computing Machinery, 610–23. DOI: https://doi.org/10.1145/3442188.3445922

Bianchi Federico, Kalluri Pratyusha, Durmus Esin, Ladhak Faisal, Cheng Myra, Nozza Debora, Hashimoto Tatsunori, Jurafsky Dan, Zou James, Caliskan Aylin (2022). "Easily accessible text-to-image generation amplifies demographic stereotypes at large scale" . In: *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23).* New York: Association for Computing Machinery, 1493-1504. DOI: https://doi.org/10.1145/3593013.3594095

Bowman Samuel R. (2023), "Eight Things to Know about Large Language Models". *arXiv preprint arXiv:2304.00612*.

Nozza Debora, Bianchi Federico, Attanasio Giuseppe (2022). "HATE-ITA: hate speech detection in Italian social media text". In: *Proceedings of the Sixth Workshop on Online Abuse and Harms* (WOAH). Seattle: Association for Computational Linguistics, 252-260. DOI: 10.18653/v1/2022.woah-1.24

Piergentili Andrea, Fucci Dennis, Savoldi Beatrice, Bentivogli Luisa, Negri Matteo (2023). "From Inclusive Language to Gender-Neutral Machine Translation". https://arxiv.org/abs/2301.10075

Stanovsky Gabriel, Smith Noah A., Zettlemoyer Luke (2021). "Evaluating gender bias in machine translation". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*,. Florence: Association for Computational Linguistics, 1679–1684. DOI: 10.18653/v1/P19-1164

Strubell Emma, Ganesh Ananya, McCallum Andrew (2019). "Energy and Policy Considerations for Deep Learning in NLP". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence: Association for Computational Linguistics, 3645–3650. DOI: https://doi.org/10.18653/v1/P19-1355

A proposed European Union workgroup for developing multilingual and multimode corpora in response to multi-crisis situations

Federico Garcea
Università di Bologna

In responding to national or international crises and emergencies, having technological tools for communicating quickly with the affected population is a fundamental part of strategies for mitigating damage, providing aid, or simply forewarning the public to prevent worse damage.

In an internationally interconnected society, overcoming cultural and language barriers is often an imperative in times of crisis. Machine translation tools can be enormously helpful in such situations, but prior preparation is needed in terms of collecting and organising information, creating one or more multilingual corpora, and forging the scientific and technical skills involved in developing machine translation systems.

In the case of the 2010 earthquake in Haiti, the local emergency services were overwhelmed, and international responders from governments and NGOs had the problem of being able to communicate in the local language (Haitian Kreyòl). Through voluntary collaboration between academic groups and private companies, a human translation system was quickly set up whereby text messages asking for help could be distributed and triaged. Within another week, this was followed by a free machine translation system based on a Kreyòl-English corpus created by researchers from the collaborating organisations and the same translators who had been involved in the text messaging system.

The operation's success inspired a set of recommendations for creating similar resources in crisis situations (Lewis, Munro, Vogel 2011: 501).

More and more often, we find ourselves faced with international crises or multi-crises, where there is no single, localized event—like Haiti's earthquake or another natural disaster, or a terrorist attack—but a cascading series of events that start from a specific country or region, spreading across a spiralling number of countries and potentially reaching every corner of the globe. This is the case of the 2020 COVID-19 pandemic, or the humanitarian and energy multi-crisis sparked by the Russian invasion of Ukraine in 2022.

In such situations, the effects of the crisis do not make themselves felt immediately on the international scene, and the chief objective is to disseminate reliable, verified information in order to coordinate research, trade, logistics and communication in general.

During the COVID-19 crisis, the problem was to make the information held by the World Health Organization available in 'minor' languages, or in other words the languages that are not included among those into which each new release by the WHO

is rapidly translated but are needed in order to reach hundreds of millions of people in Africa and Asia. This effort was undertaken by the TICO-19 (Translation Initiative for COvid-19) virtual team, which created a corpus in 35 different languages with the information then available about the virus, including mitigation and containment strategies, known treatments and their observed effects, etc. (Anastasopoulos *et al.*, 2020).

This corpus was a fundamental resource both for human translators through a collection of translation memories, and for machine translation systems and researchers, as it made it possible to readily adapt existing tools with appropriate terminology and correct, timely information.

It is also a useful resource for fighting the spread of disinformation and instrumentalization for political or criminal ends, as it is possible to check whether information is trustworthy using computer assisted or machine translation.

It should be emphasised that this type of resource is even more important today, with the advent of LLMs (Large Language Models) and GPT (Generative Pre-trained Transformers) that can generate information in smooth, coherent—but not necessarily factual—language: resources can be *consumed* by GPT systems as they update themselves, or by government and non-governmental bodies to create GPT-like apps to inform the public directly and answer their queries automatically or semi-automatically.

In the United States, the University of Washington (UW) has banded together with other academic institutions to found the Language Technologies for Crisis Preparedness and Response (LT4CPR) group to continue the work done in this field and ensure better coordinated efforts in future crises.

We believe that a European Union workgroup should be set up to deal with these issues by creating and maintaining multilingual corpora on the topics of greatest concern for Europe's population in the event of humanitarian or economic crises.

Resources must be provided in all official EU languages, the languages of the linguistic minorities recognised by the Union, and the languages which are most widespread among the international communities residing in the EU (e.g., Turkish, Arabic, Ukrainian, etc.).

It is also important to create multimodal corpora containing both text and audiovisual information. Text information alone is not sufficient to cope with the volume and types of communication that are often necessary to reach all segments of the population.

There are situations in which healthcare, social and aid workers must communicate face to face (in person or online), and being

unable to use speech recognition and synthesis systems and multimodal machine translation is often a limitation.

European-wide multilingual and multimodal resources reflecting the continent's diatopic variation would be an effective means of improving the quality and timeliness of the health, legal, social and economic information exchanged among all citizens and residents of the European Union, and could literally save more lives in crisis situations in addition to mitigating their impact on the population and its most vulnerable segments.

References

Anastasopoulos Antonios, Cattelan Alessandro, Dou Zi_Yi, Federico Marcello, Federmann Christian, Genzel Dmitriy, Guzmán Francisco, Hu Junjie, Macduff Hughues, Koehn Philippe, Lazar Rosie, Lewis Will, Neubig Graham, Niu Mengmeng, Öktem Alp, Paquin Eric, Tang Grace, Tur Sylwia (2020). "TICO-19: the Translation Initiative for COvid-19". In: *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*. Association for Computational Linguistics. https://aclanthology.org/2020.nlpcovid19-2.5/

Lewis William, Munro Robert, Vogel Stephan (2011). "Crisis MT: Developing A Cookbook for MT in Crisis Situations". In: *Proceedings of the Sixth Workshop on Statistical Machine Translation*. Edinburgo: Association for Computational Linguistics, 501–511. https://aclanthology.org/W11-2164/

R4

Plurilingual terminology resources complying with the FAIR guiding principles for the Semantic Web

Silvia Calvi
Klara Dankova
Lucrezia Marzo
Maria Teresa Zanola
Università Cattolica del Sacro Cuore, Milano

Being able to rely on authentic, current and reliable plurilingual terminology resources is an invaluable asset in applications ranging from translation and interpreting to terminological analysis to facilitate the transfer of knowledge. Accordingly, the REALITER Pan-Latin Terminology Network[1] works with specialised terminology in the Romance languages (Gilardoni 2011, Zanola 2014). In over thirty years of activity, the network has addressed a broad array of topics with plurilingual glossaries for fields ranging from computer science to commerce, from biotechnology to medicine, and from fashion to sport. In addition to specialised translators, the glossaries are intended for professionals such as experts in the discipline covered, language consultants, journalists, revisors, editors and many others, as well as for anyone wanting to be better informed about how to use certain terms correctly in plurilingual communication (Calvi 2020). The work underpinning REALITER glossaries is based on precise methodological principles, aiming to 'promote the harmonic development of the Neo-Latin languages, taking their common origin into account'[2]: all of the Network's languages, including their variants, are equally important in the group's projects. From the theoretical/methodological standpoint, the variationist and systemic approaches are used; the Network encourages collaboration in terminological work, not only between terminologists for different languages and countries, but also between experts on the sectors addressed by the Network. Lastly, to ensure quality products, the Network applies the principles of accessibility, currency and reliability (*Principi metodologici del lavoro terminologico* 2000, Zanola 2012).

Since 2021, collaboration with CLARIN-IT (the Italian Common Language Resources and Technology Infrastructure) has been fundamental in making REALITER's plurilingual glossaries more accessible. To this end, REALITER glossaries have been converted into formats complying with the FAIR principles for the Semantic Web, which call for data to be Findable, Accessible, Interoperable and Reusable (Cimiano 2020). The glossaries and associated metadata have been included in the 'REALITER – OTPL'[3] collection and uploaded on the Social Sciences and Humanities Open Cloud—SSHOC[4] terminology platform, a tool for sharing data and, consequently, knowledge.

Investing in terminological work of this kind means amassing a scientific and cultural heritage that can undoubtedly improve the

[1] https://www.realiter.net/

[2] https://www.realiter.net/presentazione/regolamento

[3] https://dspace-clarin-it.ilc.cnr.it/repository/xmlui/handle/000-c0-111/565

[4] https://sshopencloud.eu/

effectiveness and efficiency of artificial intelligence applications. These resources can be integrated in machine and computer assisted translation systems to ensure quality output.

References

Calvi Silvia (2020). "Il lessico dell'arrampicata sportiva: metodologia per la progettazione ed elaborazione di un lessico plurilingue". In: Manuel Célio Conceição, Maria Teresa Zanola (a cura di). *Terminologia e mediação linguística: métodos, práticas e atividades*. Universidade do Algarve, 287-301. http://hdl.handle.net/10400.1/15043

Cimiano Philipp *et al*. (2020). *Linguistic Linked Data. Representation, Generation and Applications*. Berlino: Springer. DOI: https://doi.org/10.1007/978-3-030-30225-2

Gilardoni Silvia (2011). "I lessici della Rete Panlatina di Terminologia". In: Maria Teresa Zanola, Francesca Bonadonna (a cura di). *Terminologie specialistiche e prodotti terminologici*. Milano: EDUCatt, 101-112.

Monachini Monica, Frontini Francesca (2016). "CLARIN, l'infrastruttura europea delle risorse linguistiche per le scienze umane e sociali e il suo network italiano CLARIN-IT". *Italian Journal of Computational Linguistics*, 2, 11-30. DOI: https://doi.org/10.4000/ijcol.387

REALITER. *Principi metodologici del lavoro terminologico*. http://www.realiter.net/wp-content/uploads/2013/06/Principi-metodologici-del-lavoro-terminologico.pdf

Zanola Maria Teresa (2014). "Le réseau Realiter, un acteur du plurilinguisme". *Plaisance*, 11, 149-165.

Zanola Maria Teresa (2012). *Costruire un glossario: la terminologia dei sistemi fotovoltaici*. Milano: Vita e Pensiero.

Zanola Maria Teresa, Villa Maria Luisa, Dankova Klara (2023). "Langages et savoirs : intelligence artificielle et traduction automatique dans la communication scientifique". In: Rachele Raus (cur.) *et al*. *De Europa Special Issue. 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*. Torino, Milano: Università degli Studi di Torino, Ledizioni LediPublishing, 107-127. https://www.collane.unito.it/oa/items/show/132.

For a common European framework for evaluating AI-based translation technologies

Philippe Langlais
RALI, DIRO, Università di Montréal

François Yvon
Sorbonne Université e CNRS

Deep neural networks, and especially the Transformer architecture (Vaswani *et al.*, 2017), have brought tremendous progress in machine translation (Sutskever *et al.*, 2014; Bahdanau *et al.*, 2016). Many services based on this technology can produce good quality translations, though they are still often literal (Bhardwaj *et al.* 2020), contain contradictions or omissions, and are less pertinent in certain specific areas such as the financial and automotive industries.

To improve machine translation for professional purposes, it is essential to develop a better European framework based on practice-oriented metrics and pertinent, absolutely multisectorial data. This will entail involving the many actors in the world of translation: scholars in NLP and translation studies, companies providing translation devices and software, as well as translators and translation services. Doing so at the European level is essential in order to pursue a successful strategy for reducing the technological inequalities between European languages[1] and, above all, enabling Europe to take the lead in integrating technologies for professional use.

A few words are thus in order concerning current efforts to evaluate machine translation, such as the WMT campaigns[2], which concentrate on evaluating translation technologies and output quality, or their IWSLT[3] counterparts for spoken language translation (interpreting). In the WMT work, data are annotated according to the MQM taxonomy by translation professionals whose tasks include evaluating post-editing effort and predicting whether a translation contains so-called 'catastrophic' errors.

Despite this work, it is still difficult to gauge the results of quality evaluation efforts. Though some system have gone beyond simply detecting errors, few or none analyse them. It thus comes as no surprise that a review of the principal research papers dealing with machine translation published between 2010 and 2020 (Marie *et al.* 2021) found that BLEU scores (Papineni *et al.* 2002) continue to be used to measure how close a machine translation is to a reference human translation by counting the words and phrases they share.

Document-level translation quality evaluation is still uncommon (Specia *et al.* 2020; Zerva *et al.* 2022), though it is extremely useful from a professional standpoint.

[1] See the European Language Equality project at https://european-language-equality.eu/wp-content/uploads/2022/11/ELE__Deliverable_D3_4__SRIIA_and_Roadmap__final_version_-1.pdf

[2] http://www2.statmt.org/wmt23/

[3] https://iwslt.org

Lastly, there are too few studies addressing the management of Translation Memories (TMs) and their use in the translation process, though they are essential professional tools.

As for how deep learning systems are trained and tested, data used for this purpose have been collected from the published proceedings of the European Parliament (Koehn 2005), United Nations documents (Ziemski *et al*. 2016) and from parallel corpora harvested from the Internet (Esplà *et al*. 2019). In addition, specific data have been evaluated in the past for particular sectors, for the media, and so forth.

Despite the abundance of this data (at least for some language pairs), however, no attention has been devoted to splitting training and test corpora in any functionally targeted way. Test corpora are often packed with stereotype-laden and extremely repetitive phrases which, moreover, are already present in the training corpora. This risks 'contaminating' the tests, with repercussions that include overoptimistic evaluations. Few studies have addressed the evaluation of the examples used to train models, where quantity trumps quality. And yet, selecting data according to specific criteria would make it possible to train more robust models and build more consistent datasets (in this connection, see the exemplary case described by Varshney *et al*. 2022).

It should also be borne in mind that the document is a secondary element in organising data and that corpora are usually segmented in equivalent sentences for language pairs (the so-called aligned corpora). This makes it difficult to produce a cohesive translated text, given that the basic unit is the sentence.

More generally, developing a single system capable of dealing with multiple domains, though increasingly fundamental, is still an underinvestigated—and hence unsolved—problem (Pham *et al*. 2021). Most of the studies in this area have addressed a small number of highly diverse sectors (biomedicine, finance, technology) and thus do not encompass the broad array of domains that translation services must consider. For example, it has been found (Frenette 2021) that a generic neural translation system had difficulty in translating texts in several of the sectors handled by the Canadian Government's Translation Bureau, and that technical attempts to provide the system with further information in these sectors proved useless.

Summarising, we can say that despite the undeniable advances in machine translation, current frameworks for evaluating MT are inadequate, and that a thorough rethinking is required in order to develop more useful technologies meeting professional needs. This calls for more work in data preparation and annotation, for-

mulating representative metrics and developing new technologies (interactive translation and/or TM pre-translation, devices for managing translation flows, and so forth.

Undoubtedly, developing a common European evaluation framework is an ambitious project requiring synergistic efforts on the part of all the actors in the world of translation. But is also a challenge that Europe, with its multilingual strengths, is certain to overcome.

References

Bahdanau Dzmitry, Cho KyungHyun, Bengi Yoshua (2014). "Neural Machine Translation by Jointly Learning to Align and Translate". *ICLR*. https://arxiv.org/pdf/1409.0473.pdf

Bhardwaj Shivendra, Hermelo David Alfonso, Langlais Phillippe, Bernier-Colborne Gabriel, Goutte Cyril, Simard Michel (2021). "Human or Neural Translation?". In: *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona: International Committee on Computational Linguistics, 6553–6564. DOI: 10.18653/v1/2020.coling-main.576

Devlin Jacob, Chang Ming-Wei, Lee Kenton, Toutanova Kristina (2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Volume 1 (Long and Short Papers)* di *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Minneapolis: Association for Computational Linguistics, 4171–4186. DOI: 10.18653/v1/N19-1423

Espla Miquel *et al*. (2019). "ParaCrawl : Web-scale parallel corpora for the languages of the EU". In: *Proceedings of Machine Translation Summit XVII: Translator, Project and User Tracks*. Dublino: European Association for Machine Translation, 118–119. https://aclanthology.org/W19-6721.pdf

Frenette Xavier (2021). *Utilisation du plongement du domaine pour l'adaptation non supervisée en traduction automatique*. Tesi di laurea magistrale, Università di Montréal. DOI: https://doi.org/1866/26528

Koehn Philipp (2005). "Europarl: A parallel corpus for statistical machine translation". In: *Proceedings of Machine Translation Summit X: Papers*. https://aclanthology.org/2005.mtsummit-papers.11.pdf

Marie Benjamin, Fujita Atsushi, Rubino Raphael (2021). "Scientific Credibility of Machine Translation Research: A Meta-Evaluation of 769 Papers". In: *Vol. 1: Long Papers. 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*. Association for Computational Linguistics, 7297–7306. https://aclanthology.org/2021.acl-long.566.pdf

Papineni Kishore, Roukos Salim, Ward Todd, Zhu Wei-Jing (2002). "Bleu: a Method for Automatic Evaluation of Machine Translation". In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 311-318. https://doi.org/10.3115/1073083.1073135

Pham MinhQuang *et al.* (2021). "Revisiting Multi-Domain Machine Translation". *Transactions of the Association for Computational Linguistics*, 9:17–35. DOI**:** 10.1162/tacl_a_00351

Specia Lucia *et al.* (2020). "Findings of the WMT 2020 Shared Task on Quality Estimation". In: *Proceedings of the Fifth Conference on Machine Translation*. Association for Computational Linguistics,743-764.https://aclanthology.org/2020.wmt-1.79.pdf

Sutskever Ilya, Vinyals Orion, Le Quoc V. (2014). "Sequence to sequence learning with neural networks". In: *Advances in neural information processing systems*. https://proceedings.neurips.cc/paper/2014/file/a14ac55a4f27472c5d894ec1c3c743d2-Paper.pdf

Varshney Neeraj *et al.* (2022). "ILDAE : Instance-Level Difficulty Analysis of Evaluation Data". In: *Volume 1: Long Papers : Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*. Dublin: Association for Computational Linguistics, 3412-3425. DOI: 10.18653/v1/2022.acl-long.240

Vaswani Ashish, Shazeer Noam, Parmar Niki, Uszkoreit Jakob, Jones Llion, Gomez Aidan N., Kaiser Lukasz, Polosukhin Illia (2017). "Attention is All you Need". In: *Advances in Neural Information Processing Systems*. DOI: https://doi.org/10.48550/arXiv.1706.03762

Zerva Chrysoula *et al.* (2022). "Findings of the WMT 2022 Shared Task on Quality Estimation". In: *Proceedings of the Seventh Conference on Machine Translation (WMT)*. Abu Dhabi: Association for Computational Linguistics, 69-99. https://aclanthology.org/2022.wmt-1.3.pdf

Ziemski Michal *et al.* (2016). "The United Nations Parallel Corpus v1.0". In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. Portoroz: European Language Resources Association (ELRA), 3530-3534. https://aclanthology.org/L16-1561.pdf

Annex

Universities and research institutions whose personnel collaborated in the studies conducted by the Jean Monnet Centre of Excellence on Artificial Intelligence for European Integration panel on linguistic rights and AI

Centre National de la Recherche scientifique — CNRS
REALITER — Rete panlatina di terminologia
Kedge Business School
Sorbonne Université
Universidad del País Vasco
Universidade Estadual de Campinas
Università Bocconi
Università Cattolica del Sacro Cuore — Milano
Università degli Studi di Enna “Kore”
Università degli Studi di Napoli “Federico II”
Università di Bari
Università di Bergamo
Università di Bologna
Università di Catania
Università di Genova
Università di Modena e di Reggio Emilia
Università di Napoli “Parthenope”
Università di Roma 2
Università di Torino
Università L’Orientale di Napoli
Università Vita-Salute San Raffaele
Université Bordeaux Montaigne
Université Catholique de Lille
Université de Montréal
Université de Paris 12
Université de Toulouse
Université Lumière Lyon II,
Université Mohammed V — Rabat
Université Paris Cité
Université Paris 8
University of Geneva
University of South Australia
Vrije Universiteit Brussel

GLOSSARY

Glossary

Algorithm: In computer science, a sequence of computations that makes it possible to solve a problem.

Artificial intelligence (AI): Science that proposes to develop intelligent computer systems by replicating human mental processes; by extension, the term refers more generally to computer models and material devices based on deep learning.

Artificial neural networks: Mathematical models that mimic the functions of the human brain. The networks are made up of elementary computing units called artificial neurons.

Bias: Forms of distortion of reality; errors that can lead to full-blown prejudices.

Computer model: Set of abstract mechanisms describing the structure of concrete knowledge. Factual reality is thus represented abstractly by means of formal (computer) languages.

Corpus (plural Corpora): A dataset selected in order to be searchable on the basis of specific criteria.

Deep learning: A field of AI research whereby artificial neurons process information to enable neural networks to learn.

Diatopic variation: The variation of languages across the geographical areas where they are spoken as an effect of sociolinguistic factors. Examples include the regional variation of Italian or of the French spoken in different areas such as France, Belgium, Switzerland, etc.

Language industry: The set of products, techniques, activities or services that require natural language processing. One example is the production of devices for machine translation, machine interpretation, etc.

Language model or linguistic model: A model where neural networks trained using self-supervised or semi-supervised learning employ probabilistic and statistical techniques to predict the use of one or more words in a sentence.

Large language model: Language models capable of automated unsupervised, self-supervised or semi-supervised deep learning based on enormous quantities of data. A LLM is thus an advanced artificial system that uses massive datasets to reproduce and generate human language, as in the case of ChatGPT.

Mathematical model: Set of abstraction mechanisms providing a quantitative representation of natural phenomena.

Minority language: Language used within a given territory of a State by nationals of that State who form a group numerically smaller than the rest of the State's population. While this definition has been taken from the *European Charter for Regional and Minority Languages* adopted by the Council of Europe in 1992, in this report, the term is also used—as Agresti specifies—to denote national languages that are used less than other languages in a given context.

Notes

Notes

Notes

Notes

Finished print in November 2023 by
Rotomail SpA - Vignate (MI)

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

How artificial intelligence can further European multilingualism

Strategic recommendations for European decision-makers

edited by
Rachele Raus

€ 29,00 |

ai4ei

Artificial Intelligence for European Integration
Jean Monnet Centre of Excellence

9 791256 000142