LINGUISTIC DIVERSITY AND AI IN THE EU: A CHALLENGE AND A DRIVER FOR LEARNING OPPORTUNITIES: THE CASE OF FRENCH LANGUAGE

Assoc. Prof. Alida Maria Silletti, PhD
Department of Political Science,
Universita degli studi di Bari Aldo Moro

Abstract:

This proposal follows from a classroom activity involving students learning French as a foreign language. It concerns the use of YouTube's speech-to-text tool in order to transcribe automatically oral speech and to revise it in the post-editing process with students. Speech-to-text tools will be considered as a driver for developing new innovative learning opportunities for foreign languages in the EU at a university level. Nevertheless, this opportunity could only be experimented if a language is written and gathers a linguistic repository for all levels of communication. In the EU, all the official languages are recognised and have the same rights in the official communication, but the same does not concern regional and/or minority languages, nor other languages spread in the EU. Hence, after drawing the status of the EU languages and the functioning of speech-to-text tools, it will be shown that the most represented language in AI is English, despite many other official or non-official languages. The discussion that will be presented deals with the respect of linguistic rights and diversity in the EU, and with the way in which EU initiatives may contribute to it, also by developing proper AI technologies.

Keywords: Artificial Intelligence; speech-to-text tool; linguistic diversity; linguistic rights; French language

Introduction

The didactical and scientific activity which will be represented in this paper is the result of two European projects in which the author participates as a member. In particular, this is the case for the project *Artificial Intelligence for*

European Integration (AI4EI)[1] of the Jean Monnet Centre of Excellence of the University of Turin (ended in 2023), whose responsible person was Rachele Raus (University of Bologne-Forlì) and for the more recent and ongoing project (2024-2026) of the Jean Monnet Module entitled *Communicating EU for Participating* (COMEU4PAR)[2], whose responsible person is Angela Maria Romito (University of Bari, Department of Political Science). As a participant in these European projects funded by the European Commission and aimed at allowing citizenship to know and to be in touch with the EU policies, in this paper some ideas will be developed in order to look at the potentialities of (non-generative) AI tools for didactical activities related to foreign language learning at university level, to make some reflections on EU linguistic patrimony and to think about the possibility for the EU of developing proper made in Europe AI technologies addressed to its citizens and institutions (Raus 2023).

The outline of this paper will consist of the presentation of a didactical experience in Italy involving AI at university level in French as a foreign language and as a language for specific purposes. This section will be followed by some insights dealing with the functioning of a speech-to-text tool, in the aim of introducing the status of languages (official, and regional or minority languages) in the EU. Finally, some remarks will be presented about the respect of linguistic rights and diversity in the EU, and about the need of made in Europe AI technologies.

A university didactical experience in Italy involving AI in French as a foreign language and as a language for specific purposes

Since the pandemic period, during the teaching activity of French language – advanced, addressed to MA degree students in International Relations and in Administrative Science at the Department of Political Science of the University of Bari, many activities are conducted for perfectioning students' knowledge of French language. Among them, the possibility of using AI tools applied to institutional communication of the French President Emmanuel Macron (official discourses, official messages, press conferences, and interviews).

This group is primarily composed, per year, of 30 Italian native language students who already have at least a B1 level of French language knowledge and who normally use AI tools in their everyday life. They are allowed to reflect on both AI tools as a driver and as a challenge for language development and the features which characterise institutional discourse (Oger 2005) and an oral communication which is not entirely spontaneous before it is performed by the speaker. This material allows not only to better concentrate on the

[1] https://www.jmcoe.unito.it/about_us

[2] *Communicating EU for participating* (COMEU4PAR - Pr. n. 101175902).

automatic transcription of this oral speech, but also to deal with the way a speech-to-text tool performs. By looking at YouTube videos, it is possible to activate both subtitles and automatic transcription, as it is shown in Figure 1, taken from a declaration by Emmanuel Macron held on 22nd March 2022:

Figure 1 - YouTube's subtitles and automatic transcription
(https://www.youtube.com/watch?v=BpRLBy8EIYw&t=25s)

After a general outlook at this transcription, which is freely accessible and whose results can be downloaded by copying it manually and pasting it in a word file – as it is done for students, by creating a table with two columns, the left one containing the automatic transcription and the right one with an empty space because it will contain the revised transcription which each student has to perform –, attention is paid to aspects dealing with oral French language and with French syntax (Le Goffic 2001, 2005). Indeed, in YouTube's automatic transcriptions punctuation markers are completely lacking, as well as, partially, graphic markers. The linguistic revision students will carry out during this atelier and by themselves, at home, takes into account punctuation markers and grammatical problems coming from the peculiarities of French language. Indeed, this language contains many homophones which correspond to different written words. Hence, the aim is to focus on a grammar integrating both oral and written French (Silletti 2025). By looking at the performances of YouTube's AI tool, it is possible to consider, as also previous studies have demonstrated as far as French language is concerned (Tancoigne *et al.* 2020), that this software is not the most performant one, but it is free. It tries to reproduce at least 80-85% of what it perceives, even though in this effort it also reproduces inexistent words or non-senses. This is the reason these results have always to be checked during a human post-editing activity for each AI tool. AI speech-to-text tool's performances depend on several factors (Tancoigne *et al.* 2020): external noises, speaker's pronunciation, overlapping – in case of multiple voices –, speaker's accent, but also internal biases of the AI tool, which could be probably reduced if the amount of AI texts for training and

performing the system was higher. As a consequence, it is important to look at the functioning of an AI speech-to-text tool.

The functioning of YouTube's speech-to-text tool

Speech-to-text tools have known diverse and more and more performed evolutions since their creation during the Cold war (Kneubühler 2022). Nowadays, an AI speech-to-text tool results from the combination of different factors which interact in order to offer the best result (Dufraux 2022). They are composed of an acoustic signal which has to be decoded by the machine for obtaining the speech-to-text transcription. During the decoding process, several models interact, that is an acoustic model, a language model, and a lexicon. Each model is based on a previous corpus which serves as reference corpus for comparing the decoding of the acoustic signal, and the acoustic and language model, to which the existing vocabulary has to be added, as it is presented in Figure 2, taken from Dufraux (2022, p. 17):

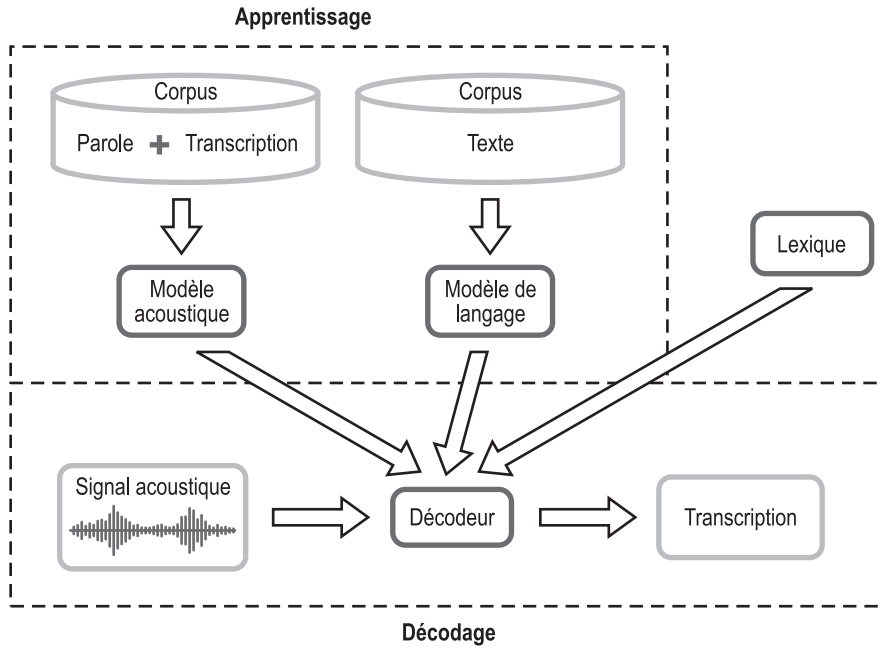

Figure 2 - Functioning of a speech-to-text tool

These features show that the further an oral speech is aligned to these models, which are trained on French standard language (variety spoken in Paris), the more the results of the AI tool are well-performed. Moreover, the system underlying an AI speech-to-text tool is trained on an enormous amount of data, which is big data. As a consequence, it is possible to infer that, in terms of languages, the best automatic transcriptions may come from the most

represented languages, and, on the contrary, that the worst ones depend on a minor representativity of a language – even though also other factors influence the results. Furthermore, it is also possible that any AI tool can be created for a single language if this language is not represented in a written form but only in an oral one. Hence, the main question which might be asked starting from this atelier on French language consists of considering whether this activity could be conducted also for other EU languages and for EU less represented languages.

The status of languages in the EU

The EU officially recognises the same rights to its 24 official languages, but their representation in the EU is unequal, also in the European institutions. Indeed, even if each European deputy has the right to express themselves in the official language of their country inside the European Parliament, even if the official language of the EU Court judgements is French, and even if work languages in the EU are mainly English, French, and German, it is well-known that the most used language in the EU remains the English language, also after the Brexit. This overrepresentation for English language – which becomes the *lingua franca* of more and more contexts, like also in international conferences or in international publications, like the present one, for instance – is a great advantage for developing AI tools aimed at both automatic transcription and automatic translation. Furthermore, AI translations in languages different from English often continue to pass through a *pivot* language, which is the English one, only because AI translations in and from English language are the widest and the most developed (Raus *et al.* 2023). This inequality and discrimination among official languages is worsened for less represented languages in the EU, which is regional and/or minority languages, or other language „minorized" in the EU (Agresti 2023), whose knowledge, learning, and written corpora are not equal nor, in some cases, not existing. Even though there is a European Charter for regional or minority languages[3] since 1992, held by the Council of Europe for preserving and promoting these languages as part of the EU cultural patrimony, its application continues to be discontinuous due to the disagreement of some of its member States. Twenty-five of them (belonging to the Council of Europe) approved and ratified it in their domestic legislations – this is the case for Romania, for instance –, while some others – like, among others, France – signed this Charter but did not ratify it; finally, some others did not approve nor ratify it – this is the case for Italy, for example. This document aims at building a Europe based on democracy and cultural diversity, and at using these languages in the public and private life. Nevertheless, as already mentioned, the recognition of this Charter is not homogeneous in Europe nor in the EU member States. This Charter (art. 1) considers a language as „regional or minority" one, hence without clearly distinguishing them. Indeed, the Council of Europe allows

[3] https://www.coe.int/en/web/european-charter-regional-or-minority-languages

each member State to distinguish or to identify its minority languages or its regional languages. Some confusions persist on these categories, for each language, which could have different status from a member State to another, and for each member State. Moreover, a language which could be official for an EU member State could be recognised as a minority one for another EU member State. Nevertheless, the most important problem concerning minority, sometimes also „ultra-minority", and regional languages is the fact that their representation depends on the economic value of them (Agresti 2023), which is often perceived more important than their real existence. Another threat deals with the responsibility of this promotion and preservation, which generally depends on local associations or on voluntary private citizens rather than by a State or a Government. Just to give an example, it is possible to look at Basque language, which is a difficult regional language to learn, which is also the only one in the EU which does not belong to the Indo-European family of languages. Its complexity and unicity show that efforts dealing with this language may be not fruitful in terms of economic gains as this language is difficult to be learnt and not widely used. These remarks influence the development of AI tools – even though AI tools in Basque language, namely for automatic translation, exist (Sarasola *et al.* 2023) –, because their functioning is based on big data and not all the languages present a linguistic repository for all levels of communication to be exploited for AI tools.

Conclusions: the respect of linguistic rights and diversity in the EU and the aim and need of made in Europe AI technologies

Internet represents a challenge and an opportunity for linguistic diversity, but all the languages have to be digitalised. Moreover, AI implies language standardisation – as its functioning is based on probabilities and statistics – and a large amount of data to be analysed, which depend on the contexts in which a language is used. Instead of promoting the EU linguistic diversification, this standardisation is only beneficial for English language, whose linguistic patrimony is the widest one and the most powerful in terms of AI tools (Vetere 2023). On the contrary, the further a language is underrepresented in the web the more its digitalisation is difficult, with the risk of a digital extinction of it. Hence, AI may be an opportunity for all the EU languages to be technologically equipped, to normalise them and their uses, to use them in the public sphere, but nowadays it only represents a voluntarily praxis depending on speakers and promoters of these languages, like local or regional associations (Agresti 2023). A more general action should be carried out by the EU to promote a linguistic planification for protecting its minority and regional languages, also coupled with AI tools (Raus 2023). It is fundamental that the EU develops proper technologies for promoting linguistic diversity through artificial intelligence, by AI speech-to-text tools, whereas the European AI Strategy[4],

[4] https://digital-strategy.ec.europa.eu/en/policies/european-approach-artificial-intelligence

developed between July and August 2024 and reinforced in April 2025, aims at excellence and trust through concrete rules and actions, but any specific information is presented about the languages which could contribute to it. In other words, the main language to reach this objective is English, that is the most standardised one.

In conclusion, in order to also promote other languages in this strategy focused on AI, teaching personnel has to be formed at school and university level in order to learn and to teach a minority or regional language of the EU, as it was pointed out in one of the proposals (Silletti 2023) included in the Guidelines addressed to the EU decision makers (Raus 2023), inside the above mentioned European project AI4EI. From another perspective, the goal of fighting against linguistic standardisation and of promoting language diversity and multilingualism in the EU (Raus *et al.* 2023) could also be intended as a way for reinforcing EU citizenship, because language protection and diversity allow member State citizens to participate in the EU activities, by investing on their shared linguistic and cultural patrimony and by combining it with the need of defending it. This is the reason this objective is also directly linked to the above-mentioned Jean Monnet module COMEU4PAR.

BIBLIOGRAPHY

- Agresti, G. (2023), „Intelligence artificielle et langues minoritaires: du bon ménage? Quelques pistes de réflexion", in Raus, R. *et al.* (eds), *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*, Turin, Milan, Ledizioni LediPublishing, p. 47-68, https://www.collane.unito.it/oa/items/show/132
- Dufraux, A. (2022), *Exploitation de transcriptions bruitées pour la reconnaissance automatique de la parole*, thèse de doctorat, Université de Lorraine, http://docnum.univ-lorraine.fr/public/DDOC_T_2022_0032_DUFRAUX.pdf
- Kneubühler, M. (2022), „Qu'advient-il de la sémantique et de l'interaction dans les transcriptions automatiques d'un logiciel ? Regard praxéologique sur le speech-to-text", *Revue d'anthropologie des connaissances*, 16/2, https://journals.openedition.org/rac/27069
- Le Goffic, P. (2001), „La phrase revisitée", *Le français aujourd'hui*, 148/1, p. 55-64, https://www.cairn.info/revue-le-francais-aujourd-hui-2001-4-page-96.htm?try_download=1
- Le Goffic, P. (2005), *Grammaire de la phrase française*, Paris, Hachette.
- Oger, C. (2005), „L'analyse du discours institutionnel entre formations discursives et problématiques socio-anthropologiques", *Langage et société*, 114/4, p. 113-128.
- Raus, R. (ed) (2023), *Per un'intelligenza artificiale a favore del multilinguismo europeo. Raccomandazioni strategiche rivolte ai decisori europei*, https://www.collane.unito.it/oa/items/show/153#?c=0&m=0&s=0&cv=0.

- Raus, R., Silletti, A.M., Zollo, S.D., Humbley, J. (2023), „Multilinguisme et variétés linguistiques en Europe à l'aune de l'intelligence artificielle", in Raus, R. *et al*. (eds), *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*. Turin, Milan, Ledizioni LediPublishing, https://www.collane.unito.it/oa/items/show/132
- Sarasola, K., Aldabe, I., Aranberri, N. (2023), „Enabling additional official languages in the EU for 2025 with language-centred AI", in Raus, R. (eds), *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*, Turin, Milan, Ledizioni LediPublishing, p. 91-105, https://www.collane.unito.it/oa/itemsshow/132
- Silletti, A.M. (2023), „Investire in tecnologie „made in UE"", in Raus, R. (ed), *Per un'intelligenza artificiale a favore del multilinguismo europeo*. *Raccomandazioni strategiche rivolte ai decisori europei*, 78-81, https://www.collane.unito.it/oa/items/show/153#?c=0&m=0&s=0&cv=0
- Silletti, A.M. (2025), „La transcription automatique au service de l'apprentissage des langues: quelques réflexions sur la phrase complexe en français", *LANGAGES*, 237, p. 131-152.
- Tancoigne, E., Corbellini, J.-P., Deletraz, G., Gayraud, L., Ollinger, S., *et al*. (2020), *La transcription automatique : un rêve enfin accessible? Analyse et comparaison d'outils pour les SHS*. *Nouvelle méthodologie et résultats*, Rapport de recherche, MATE-SHS, halshs-02917916v2
- Vetere, G. (2023), „Elaborazione automatica dei linguaggi diversi dall'inglese: introduzione, stato dell'arte e prospettive", in Raus, R. *et al*. (eds), *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*, Turin, Milan, Ledizioni LediPublishing, p. 69-87, https://www.collane.unito.it/oa/itemsshow/132