Institutional Language Futures in Times of Hyper-Collective Epistemic Goods

Back in 2017 to 2019, I was working on a text mining technique and transformer technologies for language extraction in a cross-linguistic comparison of William Faulkner's *The Sound and The Fury* and its translations. It proved to be paramount to address the practical and ideological challenges of balancing human and machine labour in text computing. I wrote on the implications of removing human labour from digital research practices in several articles, including the one published in the 2021 issue of the Hong Kong journal dedicated to Digital Humanities and Translation. These observations went unnoticed in my surrounding, including the department where I continue working. With the rise of Generative Artificial Intelligence (GenAI), the concerns around capturing human authenticity, creativity, and diversity in language studies and textual scholarship remain relevant. Yet today we are in a different place. It is hard to defend the utility of the proposal to have less of machines and more of human labour to model texts and languages in rich yet slow ways when machine productivity is often mistaken for or strategically framed as creativity.

Some of the current digital challenges pertinent to language studies and textual scholarship go far beyond the rise of GenAI and into early digital or even pre-digital periods. These challenges are both economic and ideological by nature. Trained predominantly on English-language data, Large Language Models (LLM), for example, propagate the already widespread linguistic hegemony of English. Like the English language itself, its data and LLMs are hyper-collective epistemic goods. Both globally and locally, they tend to have the greater use value than other languages and research objects created around them. It is customarily easier to embed one's work written in English and focused on English-language data or English authors within the local and global systems of research evaluation. Similarly, some research areas or topics like big data, AI or code vibing are currently considered more significant, valuable, and desirable than research focused on small data or technologies other than AI or not linked explicitly to AI. We tend to gravitate towards these goods in the production and dissemination of our research and education in exchange for greater support, recognition, and visibility in academia.

This academic economy of scale tends to produce monocultures, to use the analogy with natural ecologies. In colonial farming, for example, economy of scale always kills biodiversity. Human intelligence is diverse; its diversity comes from different cultures, multilingualism, and intercultural communication. Yet with no linguistic diversity in digital tools like English-focused LLMs and practices, the ways we think and express ourselves will be mainstreamed even more in the long-run. While we are not short of critical views of hyper-collective epistemic goods and of institutional alignments that promote them, research culture and pedagogies based on critical practice that would prioritize alternative R&D solutions and other languages are plausible futures at best.