DOI: 10.1111/josl.12680

DIALOGUE: COMMENTARY

Journal of Sociolinguistics WILEY

Language is not a data set—Why overcoming ideologies of dataism is more important than ever in the age of AI

Iker Erdocia[1] | **Bettina Migge**[2] | **Britta Schneider**[3]

[1]School of Applied Language and Intercultural Studies, Dublin City University, Dublin, Ireland

[2]School of Languages, Cultures & Linguistics, University College Dublin, Dublin, Ireland

[3]The Faculty of Social and Cultural Sciences, European University Viadrina Frankfurt (Oder), Germany

Correspondence
Iker Erdocia, School of Applied Language and Intercultural Studies, Dublin City University, Dublin city, Ireland, Glasnevin 9, Dublin, Ireland.
Email: Iker.erdocia@dcu.ie

> *Meaning, then, is derived not through content or data or even theory in a Western context, which by nature is decontextualized knowledge, but through a compassionate web of interdependent relationships that are different and valuable because of difference.*
>
> (Simpson, 2014, p. 11)

Helen Kelly-Holmes' call to explore the implications for sociolinguistics arising from the increased commercially driven digitalization of society is very timely. Like Kelly-Holmes, we share the view that the growing prevalence of online and artificial intelligence (AI) technologies in all aspects of our lives requires a critical assessment of assumptions, approaches, and practices that have grounded sociolinguistic research since its inception. While our discussion confirms Helen's observations, we also urge the development of a general critical attitude toward understanding language as digital data. The starting point for our argument is Helen's claim that there is an erasure of "authentic" languages from public digital spaces, "making it more difficult to gather data on real usage because it would be necessary to rely on public areas and/or negotiate access to these private spaces" (p. 5). For us, her observation brings to the fore that treating language as data has always been problematic. We want to raise two issues: the general epistemological limitations of using digital user data as a representation of language and community, and the consequent need for methods that take seriously the study of language in its social, political, and technological context. We suggest ethnography as a

method for understanding what speakers actually do, and an opening of language research to also consider the workings and socio-political embeddings of digital and generative AI language technologies. Our discussion is in the spirit of a joint fruitful and constructive debate.

Let us start with a general critique of approaching language as "data" that correlates with social groups, which is so far a neglected aspect in the debates surrounding language, sociolinguistics, and AI. Historically, this discussion links to the colonial backgrounds of Western science and linguistics specifically. Colonial or missionary linguistic research (e.g., Deumert & Storch, 2020; Errington, 2008) demonstrates that dominant Western epistemologies of language and research methods in linguistics were shaped during the period of European colonialism. An important legacy of European colonialism is that it "sought to fundamentally change and reorganize the social and economic order of the societies it colonized, as opposed to satisfy itself with extracting tribute" (Couldry & Mejias, 2019, p. 70). Part of this endeavor involved language "development" activities aimed at the goal of Bible translation and turning the colonized into Christian disciples. This was based on constructions of language that are still dominant today. They developed on the grounds of "collecting data" (in colonial times, often from single speakers) and then transforming the human capacity of embodied, interactive and collaborative meaning-making into word lists, grammar books, or dictionaries (e.g., Deumert & Storch, 2020; Gal & Irvine, 2019, Chap. 9). Linguistic research is therefore grounded in what is now called the ideology of "dataism" (Bode & Goodlad, 2023). This is the belief in data as representing human behavior. In the age of digital technologies, this is coupled with the aim of tracking human behavior to predict and ultimately shape social life (Rushkoff, 2019). Dataism implies assumptions such as the belief in objectivity of quantification and trust in data processing agents. The resulting datafication of everyday life then consists of extracting information from the flow of social life, matching it to imagined social realities and categories and fixing such relationships. In the context of linguistics, the understanding of language as data, collected from oral practices transformed into writing, led to conceptualizing language as referential code and languages as "natural," given objects that are systematically and neatly structured (e.g., Pennycook, 2004). The outcomes of these activities are summarized in typologies, developmental hierarchies, and a canon of methods for optimal data extraction and analysis (for critical discussion, see Deumert & Storch, 2020).

Sociolinguistics is part of this tradition. But unlike missionaries who aimed to impose their social imaginations on people through their linguistic activities, sociolinguists' goal is to discover and explain what people do with language, specifically variation, to deepen our understanding of language and its sociocultural embeddedness and to raise awareness and fight discrimination. Early work aggregated people's practices into externally defined, homogenizing macro social (e.g., age, class) and structural linguistic categories and derived meaning from statistical correlations between them. Over time, however, sociolinguistics has also successively questioned objectification. Subsequent research emphasized the construction of meaning as a local, interactional process that required understanding people's actions and views and advocated for detailed participant observation of people's everyday activities and semi-guided discussion. The aim is to discover locally relevant social and linguistic categories (e.g., practices, linguistic phenomena) and their local indexicalities and relationships in a holistic manner (Eckert, 2012). It is now widely accepted that social and linguistic categories are complex and dynamic across contexts and interactions. Language is a fuzzy network of social acts whose meanings emerge in context as it is fundamentally pragmatic and indexical, and meaning-making is a localized process (Eckert, 2008; Gal & Irvine, 2019; Silverstein, 2014). People are social agents who pick linguistic practices based on their identity, on the goals that they want to (temporarily) foreground, and on their current understanding of an interaction based on the indexicalities that they perceive. In addition, humans "dynamically reshape the context that provides organization for their actions within the interaction itself" (Duranti & Goodwin, 1992, p. 5), and in literate cultures, writing

and printing have co-constructed these contexts, in particular, normative ideas and epistemological approaches (Linell, 2005). Overall, sociolinguistic diversity is therefore rich and dynamic and defies neat correlative relationships.

What does all this suggest for the future of sociolinguistic research in a digital and AI technology enriched environment? The short answer is that we cannot limit our analyses to the study of digital language data, be it the effect of user interaction or the output of generative AI. On the one hand, we need to continue and enhance what we have been doing: observing people, understanding meaning-making in practice and considering the social embeddedness of language, while critically assessing, critiquing, and recalibrating our tools to avoid essentializing practices, contexts, and communities. The current social reality, ripe with unfamiliar tools, processes, and logics, requires us to upskill and engage. Ethnography's dedication to a multi-perspective and holistic understanding is well suited to grasp, for example, how, when, and where people actually use generative AI tools and how it impacts on language attitudes and language ideologies and therefore contributes to sociolinguistic realities. Instead of only examining the outputs of technological applications, that is, the linguistic data, we need to cast our net more widely. Given their intertwined nature, we have to explore people's offline and online experiences, activities and ideologies, the technological infrastructures and affordances, and their intersections in an integrated manner. Ethnography is here useful to study micro-practices, but at the same time, as it is limited to the observation of locally visible conditions, it needs to be complemented by other methodological approaches. The social, cultural, political, technological, and interactional situatedness and dynamic nature of any language activity need to be embraced in a multi-methodological fashion (e.g., Page et al., 2022). For example, young(er) people in the Global North often stay connected throughout most of their waking hours and frequently blend offline activities with simultaneous online activity, leading to at times intensely intertwined experiences because the technological and social contingencies of their public, private, and educational lives "allow" or even mandate a convergence of online and offline activities. For others, there might be a greater difference between online and offline worlds due to the lack of appropriate devices, data or network coverage, non-digitized contexts, or just a preference for offline interaction (Deumert, 2014, Chap. 3). Due to being involved in different communities of online and offline practice, individuals also develop, use, and learn to interpret linguistic and sociopragmatic indexicalities differently and develop different metapragmatic realities. Language is also not the only meaning-making resource. Type of technology, ways of using technologies (e.g., voice vs. written messages; multimodal vs. plain text) may also become contextualization cues and their indexicalities are not constant as different contexts have different affordances in terms of devices, literacy, and ideologies of language and media (Gershon, 2010). Without ethnographic observation and a consideration of the social and technological contexts, local meanings of language and the social indexicality of language and technology choices can easily be misinterpreted. This also applies to the linguistic output of AI tools, which is typically edited by users, according to their audiences and language ideologies, the latter increasingly influenced by the ascription of authority to data and algorithms, but possibly also by their rejection. The edited language feeds back into systems so that the whole AI arrangement becomes a complex socio-technical human–machine assemblage (Fester-Seeger et al., in preparation; Pennycook, 2024). In this, it is impossible to know what people do and why without engaging with people—the belief in objectified, decontextualized data as the sole source of knowledge creation has been problematic in the past and becomes even more so in an age of digital transnational interaction and AI interventions.

This also means that we need new conceptual tools, categories, and methodological approaches to study language in a society in which digital platforms, owned by a handful of American companies, make enormous profits with their data collection activities. They feed these into AI systems, which, in turn, impact language use, language ideologies, and the formation of communities worldwide. We

concur with Kelly-Holmes (2023) that we therefore cannot neglect the macro level in our research and need to put a greater focus on investigating and critically exploring sociopolitical structures and systems of commercialization and technology, and how they impact on language practices, language ideologies, linguistic research, and language policies. Our call for engagement with language in a holistic manner is thus not only a call for ethnography. We have to deepen our understanding of how language technologies are built and why. Understanding the ideological underpinnings of the market activity of the tech sector is of particular importance to fully capture the processes in which language technologies are embedded. The actual workings and motivations of digital technologies have received the least attention in linguistic research despite their impact on language practices (see however e.g., Jones et al., 2015). Critical sociological research (Couldry & Mejias, 2019) characterizes digital AI technologies as built on data colonialism, dominated by big tech companies' desire for maximization of profits through digital dispossession and data surveillance (Zuboff, 2019). Language data are of utmost centrality in making AI infrastructures a highly potent tool for structuring but also controlling humans and for commercially exploiting our life in the form of data. A research agenda that reacts to this could include, for example, exploring the implications for our field of recent critical accounts in the social sciences of the tech industry's global pillaging of human practices in the context of extractive capitalism (e.g., Couldry & Mejias, 2019; Zuboff, 2019). These critical insights can help inform a more nuanced understanding of the modus operandi of corporate language technologies, whose interests they serve and how they are monetized.

In the overall context of changing socio-technological conditions of society, we must not forget the sociopolitical and economic context. The state has traditionally played a crucial role in the framing of sociolinguistic economies (Blommaert, 2010, p. 195). More recently, many governments of both the Global North and South appear to have adopted a techno-solutionist approach to AI, including language technologies. Public authorities have long delegated the development of digital technologies to the market, replacing government language technology policy with the strategies of the commercially driven private sector, in the belief that it would achieve social goods for all (Birhane, 2020; Morozov, 2013). This situation has resulted in an acute digital inequality among languages, where the technological readiness of populations (e.g., use of smartphones), the degree of language norming (e.g., uniform/roman scripts), the size of data sets, and/or the decision by companies to create artificial data sets (see, e.g., NLLB et al., 2022) impact on whether or not a language is provided with critical AI tools and thus becomes reified and visible in digital space. It has also created tension between the private sector of commercial providers of language technologies and the blurring role of public institutions as traditional regulators and exclusive holders of normative authority in language matters (Erdocia et al., under review). We have become utterly dependent on private technologies manufactured and controlled by a handful of opaque companies. Like the raw resource mining industries, they appear mostly indifferent to the social consequences of their activities and only invest minimally if obliged by government regulations to enhance their public image. It is expected that the state, also within supra-national organizations, regains a more active role as a guarantor of fundamental rights for users with regulatory and supervisory frameworks (see EU's *Digital Services Act*). In the language field, this includes public–private partnerships to develop accurate, ethical, and unbiased data sets and technologies for all (particularly "low-resource") languages in an attempt to reduce the technology gap between English and other languages (see "Language Equality in the Digital Age" resolution, European Parliament, 2018; Rehm & Way, 2023). In contexts like Spain, state-sponsored language academies are beginning to collaborate with tech corporations to extend their language authority to AI (see Erdocia et al., under review).

And yet, we overall still know little about how companies resource, compile, and turn language practices into data. This is due to big tech secrecy but probably also due to our own disciplinary

orientation and biases (where we often avoid interaction with computational linguistics). Academic and commercial computational publications that assess processes, models, and procedures such as data scraping, "curating," "cleaning," debiasing processes, or training and testing of language models are useful to get a deeper insight into the politics of language technologies. However, since they are oriented to computationally trained specialists and are typically framed in ideologies of dataism, they need to be triangulated with narrative explanations from different people involved in and affected by these processes. Ideally, this should be complemented with observing their activities. Thus, we suggest developing studies that use semi-guided interviews and observation to investigate professionals' practices (e.g., researchers, localizers, technology designers, annotators, data curators, and CEOs) and commercial, public, and lay users' experiences with these infrastructures. This would re-visibilize the "invisibilization of technology" that Helen Kelly-Holmes observes (p. 2).

Looking at sociolinguistic phenomena across macro-meso-micro levels may help us not only to take the pulse of traditional concepts in our field in the digital space, such as standard language, language authenticity or linguistic authority. It might also contribute to our understanding of people's value attributions to the legacies of modernist conceptions of language and nation-state in a technologized world of late modern communication. This would paint a rich picture of the language ideologies and social and commercial actors that co-construct sociolinguistic economies today; their social, political, financial, and linguistic dynamics; and their material affordances, practices, understandings, and the web of indexical relationships between them. Paying attention to the entire sociopolitical and technological structure that enables the existence and penetration into all spheres of life of AI—not just user's data, activities and views—will confront us with our own disciplinary assumptions, biases of dataism, categories, practices, and colonial ideologies and can only enhance our work. Sociolinguistic findings and expertise are increasingly sought out by the tech industry to help fine-tune the functioning of AI technologies. Comprehensive engagement with the intertwined online and offline context will put us in a better position to engage with this interest in our work, understand the role of language data in contemporary socio-political contexts, and, more broadly, how our work can contribute to critically engaged understanding of the sociolinguistics of AI.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

ORCID

Iker Erdocia https://orcid.org/0000-0003-2459-1346

REFERENCES

Birhane, A. (2020). Algorithmic Colonization of Africa. *Scripted: A Journal of Law, Technology & Society*, *17*, 2. https://script-ed.org/article/algorithmic-colonization-of-africa/

Blommaert, J. (2010). *The Sociolinguistics of Globalization*. Cambridge University Press.

Bode, K., & Goodlad, L. M. E. (2023). Data worlds: An introduction. *Critical AI*, *1*, 1–2. https://doi.org/10.1215/2834703X-10734026

Couldry, N., & Mejias, U. (2019). *The cost of connection: How data is colonizing human life and appropriating it for capitalism*. Stanford University Press.

Deumert, A. (2014). *Sociolinguistics and mobile communication*. Edinburgh University Press.

Deumert, A., & Storch, A. (2020). Introduction: Colonial Linguistics—Then and now. In A. Deumert, A. Storch, & N. Shepherd (Eds.), *Colonial and decolonial linguistics: Knowledges and epistemes* (pp. 1–21). Oxford University Press. https://doi.org/10.1093/oso/9780198793205.003.0001

Duranti, A., & Goodwin, C. (Eds.). (1992). Rethinking context: An introduction. *Rethinking context: Language as an interactive phenomenon* (pp. 1–42). Cambridge University Press.

Eckert, P. (2008). Variation and the indexical field. *Journal of Sociolinguistics*, *12*(4), 453–476. https://doi.org/10.1111/j.1467-9841.2008.00374.x

Eckert, P. (2012). Three Waves of Variation Study: The emergence of meaning in the Study of Sociolinguistic Variation. *Annual Review of Anthropology*, *41*, 87–100. https://dx.doi.org/10.1146/annurev-anthro-092611-145828

Erdocia, I., Migge, B., & Schneider, B. (under review). Language in the age of AI technology—From human to non-human authenticity, from public governance to privatisation. *Language in Society*, 00–00.

Errington, J. (2008). *Linguistics in a colonial world: A story of language, meaning, and power*. Blackwell.

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

Fester-Seeger, M., Migge, B., Purschke, C., & Schneider, B. (under review). Special issue: AI technology as human interaction—Linguistic and cultural framings of AI as communicative infrastructure. *AI & SOCIETY*.

Gal, S., & Irvine, J. T. (2019). *Signs of difference: Language and ideology in social life*. Cambridge University Press.

Gershon, I. (2010). Media ideologies: An introduction. *Journal of Linguistic Anthropology*, *20*, 283–293. https://doi.org/10.1111/j.1548-1395.2010.01070.x

Jones, R. H., Chik, A., & Hafner, C. A. (2015). *Discourse and Digital Practices*. Routledge.

Kelly-Holmes, H. (2023). *Language policy 4.0: Agency, readiness and relevance in an increasingly automated future* (*Working Papers in Urban Language and Literacies*, 309). University of Limerick.

Linell, P. (2005). *The written language bias in linguistics: Its nature, origins and transformations*. Routledge.

Morozov, E. (2013). *To save everything, click here: Technology, solutionism, and the urge to fix problems that don't exist*. Public Affairs.

NLLB Team. Costa-jussà, M. R., Cross, J., Çelebi, O., Elbayad, M., Heafield, K., Heffernan, K., Kalbassi, E., Lam, J., Licht, D., Maillard, J., Sun, A., Wang, S., Wenzek, G., Youngblood, A. I., Akula, B., Barrault, L., Gonzalez, G. M., … Wang, J. (2022). *No language left behind: Scaling human-centered machine translation*. arXivabs/2207.04672

Page, R., Barton, D., Unger, J. W., & Zappavigna, M. (2022). *Researching language and social media: A student guide*. Routledge.

Pennycook, A. (2004). Performativity and language studies. *Critical Inquiry in Language Studies*, *1*, 1–19. https://doi.org/10.1207/s15427595cils0101_1

Pennycook, A. (2024). *Language assemblages*. Cambridge University Press.

Rehm, G., & Way, A. (2023). *European language equality: A strategic agenda for digital language equality*. Springer Nature.

Rushkoff, D. (2019). *Team human*. W.W. Norton & Company.

Silverstein, M. (2014). Denotation and the Pragmatics of language. In N. J. Enfield, P. Kockelman, & J. Sidnell (Eds.), *The cambridge handbook of linguistic anthropology* (pp. 128–157). Cambridge University Press.

Simpson, L. B. (2014). Land as Pedagogy: Nishnaabeg intelligence and rebellious transformation. *Decolonization: Indigeneity, Education & Society*, *3*, 1–25.

Zuboff, S. (2019). *The age of surveillance capitalism: The fight for a human future at the new frontier of power*. Public Affairs.

How to cite this article: Erdocia, I., Migge, P. B., & Schneider, B. (2024). Language is not a data set—Why overcoming ideologies of dataism is more important than ever in the age of AI. *Journal of Sociolinguistics*, 1–6. https://doi.org/10.1111/josl.12680