The colonial roots and continuities of AI language culture

Britta Schneider and Bettina Migge

Abstract *AI language technologies such as large language models (LLMs) and machine translation have become part of everyday life but we are rarely concerned with the cultural histories and epistemological backgrounds of these tools. In this chapter, we discuss the parallels between concepts of language in technology settings and discourses about language in the history of European colonialism. We compare and sketch historical links between colonial language ideologies and language ideologies found in contemporary AI language culture and study the socio-political and epistemological parallels in colonial times and the AI age. We base our discussion on previous linguistic anthropological work that has studied language and colonial discourse and compare it with discourses on language found in scholarly texts from computational disciplines, in texts published by commercial language technology companies (e.g., Microsoft, Meta AI, OpenAI, Google) and in what can be known about the design of computational devices and LLMs. Our discussion adds to an understanding that AI language technologies in many respects represent a continuation of colonial endeavours from the Global North. This also shows that the interplay of material technologies and language plays a decisive role in establishing and distributing human ideas, orders, and power.*

Keywords *Colonial or Missionary Linguistics; Language Ideologies; Standard Language Cultures; Language Technologies; Cultures of Computation; Commercialisation; Digitalisation; Self-learning Algorithms*

1. Introduction

AI language technologies, such as large language models (LLMs), have become part of everyday life, but we are rarely concerned with the cultural histories

and epistemological backgrounds of these tools. In this chapter, we discuss the parallels between concepts of language in technology settings and discourses about language in the history of European colonialism. As sociolinguistic and anthropological scholars interested in the role of language in social life—that is, in what we refer to as 'language culture' in our title—we both had no background in computation when we became involved in a project on *Language in the Human-Machine Era* (COST Action CA19102, n.d.).[1] We were keen to learn about the discourses and cultural concepts on language in the field of computational linguistics and machine learning technology and thus read scholarly articles and invited scholars and experts to give talks about these topics, which we jokingly referred to as our project of upskilling. From the very beginning, we were struck by the parallels between concepts of language in technology settings and discourses about language in the history of European colonialism. In computational texts and talks, for example, we frequently encountered the colonial trope that technologies will 'help' underprivileged communities by providing access to Western cultural practices. Clearly, the desire to include communities worldwide in digital spaces and AI technology practices is not (only) based on humanitarian goals but is part of global digital surveillance capitalism (Zuboff, 2019). Overall, there are similarities in how globally dominant actors, from historical European colonisers to current AI practitioners, exploit language to secure their position of dominance, while at the same time understanding this as a form of human progress. So, what is being packaged as new or revolutionary in AI is not necessarily all new but based on old models and motivations, and we believe that there is value in looking beyond our narrow current field of vision (also discussed in Keane, 2024).

In our exploration of the socio-political and epistemological parallels and historical links between language ideologies in colonial times and those in the AI age, we consider whether and how work on language in the digital and imminent human-machine era differs from earlier work in the colonial era, and ultimately whether such work has "left its colonial roots far behind" (Errington,

1 This research network, which ran from October 2020 to October 2024, focused on "the emergence of new types of language technology that mark a shift from the 'digital era' to the 'human-machine era'" and its aim was to facilitate a dialogue between commercial and academic technology designers, (socio-)linguists, and a wide range of practitioners using language technologies (https://lithme.eu/). The authors facilitated a Working Group dedicated to researching language ideologies, belief, and attitudes in this context.

2008, p. 150). Our aim is to raise awareness of the fact that we are currently confronted with a reordering of sociolinguistic realities, and we believe that it is useful to compare such reordering with another period of major change, which also shows that the digital turn follows a well-trodden and historically directed path. Given the increasing pressure in academia to provide 'solutions' and 'impact', which we have experienced not least in our interactions with computational sciences, we do not discuss concrete societal implications in this article but believe that critical knowledge of historical and current colonial forms of thought and action is in itself an important addition to academic and social debates.

The role of language in colonial enterprises is an established topic in linguistic anthropology referred to as colonial or missionary linguistics (Deumert & Storch, 2020, p. 3). It emerged following Fabian's (1986) monograph on linguistic description in Central Africa. The aim of this field is to critically examine the linguistic concepts and practices developed in the age of European colonialism that have crucially shaped our understanding of linguistic research and of what we understand as 'languages' – both in the colonies and in 'metropolitan' contexts. The history of colonial linguistics illustrates that linguistic epistemologies and practices are deeply intertwined with concepts of society, community, and personhood, and that constructions of language play a central role in legitimising political practices that legislate human difference (Errington, 2008). The study of such intertwined concepts of language and of society is based on the tradition of 'language ideology research' (Irvine & Gal, 2000; Woolard, 1998). In this tradition, the term 'language ideologies' does not refer to socially biased ideas about language. Instead, it is used to talk about epistemological concepts regarding language, the study of which involves, for example, the question of how the notion of 'languages'[2] as assumed 'things in the world' (and with it, ideas like words 'having meaning', see e.g., Silverstein, 2014) comes into being in culturally conditioned epistemologies. One traditional focus of this research field lies in critically reflecting epistemologies of Western linguistics, including in colonial histories (e.g., Deumert & Storch, 2020; Errington, 2001a, 2001b; Gal & Irvine, 2019). We follow this tradition in our own use of terminology.

What we observe in conceptualisations of language in cultures of computation—such as machine translation, the building of writing systems and key-

2 We use single quotation marks to indicate that these are assumptions or concepts that are controversial and to indicate that we do not align with these views.

boards, or data corpus construction for LLMs—has strong parallels to colonial traditions. Languages are conceptualised as given, object-like homogeneous entities that are understood as representing territorially-bound ethnic groups and thus serve to systematise and order human diversity. They are also conceptualised as tools that impact human thought and practices. In order to control them in the interests of the dominant groups (colonising, technology-enhanced, or technology-driven regimes), they are subjected to processes of shaping or standardisation. These processes are driven and justified by related specific moral discourses on appropriate language behaviour and on 'enlightening' and 'helping' subordinate 'others' in both settings.

Our discussion is based on linguistic anthropological work that has studied language and colonial discourse. It compares and contrasts it with discourses on language found in scholarly texts from computational disciplines, in texts published by commercial language technology companies (e.g., *Microsoft*, *Meta AI*, *OpenAI*, and *Google*) about the aims, functioning, benefits, and use of language technologies that they build. These include public and commercial machine translation tools, interpersonal communication tools such as *WhatsApp*, social networking tools such as *Facebook*, *Instagram*, keyboards, smartphone settings, chatbots, or voice assistants. We also consulted publications about what can be known about the design of computational devices and LLMs. This adds to an understanding that AI language technologies are not autonomous and agentive actors but part of cultural histories and practices in which the interplay of material technologies and language plays a decisive role in establishing and distributing human ideas, social orders, and power hierarchies.

2. The role of language in colonial and in digital culture

In this first section, we give a brief overview of the different approaches to language in the cultural contexts we focus on, that is, colonialism and AI. Colonialism has been defined as "the transformations wrought by high modern empire" through violence and displacement (Bayly, 2016, p. 2). It entails "a relationship of domination between an Indigenous (or forcibly imported) majority and a minority of foreign invaders" (Cheyfitz & Harmon, 2018, p. 271). Being convinced of their own cultural and moral superiority, the latter make decisions affecting the former in line with the interests of distant political centres. The term 'colonial' is commonly used to refer to Western empires. Colonialism

> produced unprecedented change and novelty, including massive and profoundly destructive material transformations and the constitution of a new kind of person: a colonial subject with a 'colonised mind', painfully if never fully subordinated by the coercions and 'othering' effects of the coloniser's power-knowledge. (Bayly, 2016, p. 2)

We treat colonialism as the time between the 15th century when Portuguese 'explorations' commenced until the 1960s, when most former colonies had gained formal political independence.

In the colonial era, an interest in language was bound to practices of economic interest, imperial conquest and religious conversion (Pennycook, 1998). Thus, the author of the first grammar book of a European vernacular language, Antonio Nebrija, stated in 1492, "la lengua fue siempre compañera del imperio" [language was always the companion of empire] (Cheyfitz & Harmon, 2018, p. 270; Nebrija, 1492). Later on, European colonial conquest developed into a form of mercantile capitalism, in which private financiers, that is, corporate companies employed or licensed by national states, took the initiative to establish trade and economic exploitation beyond European boundaries (Heller & McElhinny, 2017, p. 135). Already in this sense, there are interesting parallels to the current context where large technology companies pioneer global digitalisation, often financially supported by state actors (Crawford, 2021, Chapter 6). Historical colonial exploitation was legitimised by religious civilising arguments, namely by spreading the word of God to 'save' non-European souls. The control of communication to subordinate and coerce the 'other' was central in establishing European colonial power (Fabian, 1986).[3] The unified colonial vocabularies, texts, and language systems developed by Europeans created images of unified colonial subjects and territories that could be ruled and transformed in the image of the coloniser (Cheyfitz & Harmon, 2018, p. 272).

Research on colonial linguistics often focuses on the contribution of missionaries to the fixing (transforming speech into writing) and dissemination of languages and their prime aim to convert people to Christianity (Deumert & Storch, 2020; Errington, 2008; Schmidt-Brücken et al., 2015; Warnke et al., 2016). There were also other actors such as scholars from other disciplines (e.g.,

3 Note that we do not distinguish between American and European colonial desires in this article, as the colonial ideologies of Anglo-U.S. and European discourses are not different in kind and emerged at roughly the same time. The current American dominance in technology may be regarded as colonial also towards European contexts, but we here do not focus on European specificities in that sense.

geographers, anthropologists, lawyers), administrators, travellers, adventurers, explorers, and "passionate autodidacts in philology" (Gal & Irvine, 2019, p. 247) who formed an ad hoc scientific community. It engaged in committing linguistic practices to paper, systematising them, and also disseminating them to interested audiences in the metropolitan centres (Gal & Irvine, 2019). Outputs like word lists, grammars, and dictionaries can also be described as the "[r]eduction of speaking to lines of text, inaccessible for speakers of the language and focussing on grammatical orders" (Deumert & Storch, 2020, p. 9). The work of missionaries and these others was enabled by commercial entities who brought them there and supplied them, and national administrations that were at quite a distance from where the work took place (Errington, 2008, p. 4). At the same time, these actors and their work also enabled the work of commercial and state actors and it was not easily possible to separate the three. According to Errington (2008, p. 14), academic comparative philology served as an additional midwife in the construction of languages by giving "ideological and intellectual support" to the project of creating print-literate forms of local languages in the colonies. The discourses and activities concerning language in colonial settings thus have to be understood against the background of religious, economic, and political aspirations, supported by conceptual academic ideas that predominated and interacted with non-academic discourses on language at the time.

Today's AI language culture is similarly based on the interest of commercial and state actors, interwoven with academic epistemologies and the desire to explore new cultural spheres.[4] AI language technologies are based on digitalisation, the application of machine learning and the availability of large masses of data through the Internet (see Katz, 2020, for a critical discussion of the term 'Artificial Intelligence' and its emergence). The original purpose of digitalisation and computation was to automate and simplify mathematical calculations and "to capture the knowledge expressed through individual and collective behaviours and encode it into algorithmic models" (Pasquinelli, 2023, p. 2). Until the mid-20th century, programming was primarily conceived of as a rather dull and therefore feminised activity, similar to the work of a secretary (Ensmenger, 2015). During the 1960s and 1970s, "male computer experts were able to successfully transform the 'routine and mechanical' (and

4 Traditional religious motivations play no role in contemporary AI discourses, even though it would be worthwhile to study the moral and transcendent underpinnings of these discourses in more depth (see Keane, 2024).

therefore feminised) activity of computer programming into a highly valued, well-paying, and professionally respectable discipline" (Ensmenger, 2015, p. 38). Home computers became available in the 1970s and 1980s (Ceruzzi, 2003), while connecting computers became possible in the late 1960s with the so-called ARPANET, a technological development that was co-created by the U.S. Ministry of Defence and U.S. American university research labs (Couldry & Hepp, 2017, p. 48). Due to its military origin, some refer to the internet as "weapon of empire" (Tarnoff, 2022, p. 12), which became an "electronic shopping mall" (Tarnoff, 2022, p. 18) during the 1990s. In 1991, the U.S. government handed over internet operations to commercial providers (Couldry & Hepp, 2017, p. 49). A lot of early computing and internet pioneers had a more playful and experimental approach to technologies, and many believed that the internet would allow for a more democratic, more liberal, and more just society (Bunz, 2012). Digital communication allowed for easy communication and the emergence of new forms of public space. However, digitalisation and online publics in the hands of monopolist private companies are today discussed as major threats to democracy (Noble, 2018; Pasquinelli, 2023, p. 251; Zuboff, 2019).

The search engine developed by *Google* was a core element in developing computer networks into a capitalist infrastructure in which money could be earned—*Google* became one of the most influential and successful companies worldwide by inventing digital and globally spread forms of advertising and marketing (Couldry & Hepp, 2017, p. 50). Once smartphones could access the internet, personalised tracking of individuals became possible (Couldry & Hepp, 2017, p. 51). The data collected is used for personalised advertising but can also be exploited for other purposes, by *Google* but also by other companies and governmental actors (Crawford, 2021, Chapter 6). Overall, the internet developed from a "closed, publicly funded and publicly oriented network for specialist communication into a deeply commercialized, increasingly banal *space for the conduct of social life itself*" (Couldry & Hepp, 2017, p. 50, italics in original). Digitalisation and the emergence of online culture can be understood as a development in which adventurous and curious individuals, the interests of capitalist actors, and governmental desires for establishing power by expanding and controlling markets came together in transforming the world—a cultural context that is not too dissimilar to colonial histories.

While the mathematical procedures to conduct machine-learning have existed for several decades (Katz, 2020), it was only in the 2010s that extremely large amounts of data, namely those that had been collected online via comput-

ers and smartphones, and processing units that were able to process them (so-called 'GPUs', Graphic Processing Units), allowed for a wider popularisation of such tools (Bommasani et al., 2021, p. 4). Even though the development of machine learning is not interested in language per se, language data has become a core focus—besides images, language is the kind of data that is mostly available on the web and is taken to represent human thought, desire, and culture. The publication of the machine-learning text generating language model *ChatGPT* in 2022 caused worldwide public debates, surrounding questions on the supposedly super-human abilities of the tool (Heaven, 2023), the end of academic education as we know it (Marche, 2022), or the possibly drastic changes to job markets (Toh, 2023).[5]

To build a large language model (LLM) like *ChatGPT*, a self-learning algorithm (a set of calculations, in the case of LLMs, matrix multiplications, see Castelle, 2023) analyses a very large text corpus to detect statistically likely word embeddings, a procedure referred to as 'training'. Once 'trained', the algorithm can make predictions about word sequences. The input of a large number of standardised texts—i.e., texts in which similar word sequences occur—is what makes prediction work well (Schneider, 2024; see Brown et al., 2020 on source and size of training data used by *OpenAI*, the *Microsoft* funded company that released *ChatGPT* in 2022). This means that the existence of standardised languages and centuries of producing standardised written text allow an algorithm to detect statistically likely word sequences. As we will discuss below, standard language cultures are embedded in histories and epistemologies of European modernity, colonialism, and literacy, but are also the foundation of the language culture of AI.

Artificial text generation is based on LLMs. These produce written text that is grammatically coherent and is often interpreted as being equal or even superior to human linguistic abilities. However, as illustrated above, LLMs are word prediction techniques. They are "systems which are trained on string prediction tasks: that is, predicting the likelihood of a token (character, word, or string) given either its preceding context or […] its surrounding context" (Bender et al., 2021, p. 611). LLMs have mostly been developed by computational scientists rather than linguists and have no access to grammatical structures or semantic meaning—still, the output is, at least on the grammatical side, often more convincing than the output of previous grammar-based efforts of

5 Note that there is also a critical counter-discourse to these grand narratives (see the 'AI Hype Wall of Shame' at https://criticalai.org/the-ai-hype-wall-of-shame/).

linguists to make computers 'understand' language (on linguistic approaches, see, e.g., McShane & Nirenburg, 2021). On the content level, the output of statistically likely strings of words is problematic: it can be (and often is) factually wrong, a phenomenon referred to as 'hallucination' (Bang et al., 2023). Despite the fact that LLMs were not developed per se for standardising or shaping language, they already have been shown to impact language practices, including structures, meanings, and understandings of language (see, e.g., Shaitarova et al., 2023; Vanmassenhove et al., 2019; Virtanen et al., 2019), and to lead to linguistic homogenisation (Liang et al., 2024).

The creation and design of LLMs is grounded in commercial capitalist motives and, as enormous computing resources are necessary to build a model, there are currently only few commercial actors who have the capacity to create LLMs from scratch (Bender et al., 2021; McIntosh, 2019). These are most notably *Meta*, *Google*, *OpenAI/Microsoft*, and several Chinese firms. At the same time, a large number of different actors participate but also counter developments of digital commercialisation and monopolisation. Computational scientists who work in academia and in smaller or larger companies are not necessarily actively supporting the capitalist endeavours of digital monopolies but their work may tacitly contribute to the better functioning of digital tools (see a myriad of papers dedicated to this topic).[6] Yet, critical work also abounds and there are large communities that support open source tools and conferences that discuss social biases and problems as digital and AI tools become more and more popular.[7]

Traditional linguists who focus on grammar description, the creation of balanced language corpora (i.e., corpora that consist of oral, written, formal, and informal language), and traditional fieldwork for data collection are mostly sidelined in this development, as it is above all computational scientists and computational linguists who contribute to the field, often with little training in other areas of linguistics, such as critical and socially oriented approaches. The commercially-driven interest to gain and maintain customers and thus to increase the performance of technological products and the number of languages they work in (e.g., keyboards, auto correction, machine translation, chatbots, etc.) has raised interest in sociolinguistics from the computational side (personal communication with technology developer; Nguyen et al., 2016). What are presented as insights from sociolinguistics are seen as

6 https://arxiv.org/

7 E.g., https://huggingface.co; https://facctconference.org/

helping to improve data quality and data modelling (Grieve et al., 2024). Overall, discourses and activities concerning language in computational settings are influenced by economic and political aspirations, embedded in specific cultures of value (mostly capitalist ones, in this case), and supported by conceptual academic ideas reminiscent of missionary linguistics traditions. The unifications, systematisations, and orders established in previous linguistic and colonial linguistic work are partly reproduced and partly reconfigured. We discuss several levels of links, similarities, and differences between colonial and AI language culture in the following.

3. Comparing colonial and AI language activities and theories

3.1 Language and power relationships

European colonialism was based on economic desires and on constructions of superiority, racial hierarchy, cultural hegemony, and the civilising and religious mission of the colonisers (Pennycook, 1998; Said, 1978). Colonialism was a desire to rule the world. Europeans imagined a "scale of human progress" (Gal & Irvine, 2019, p. 247) and saw themselves on top of that scale. The economic exploitation and brutal subjugation, involving the enslavement, carrying off, and killing of millions between the 14th and 18th centuries, were legitimised through narratives of 'civilising' via European culture and of 'saving of souls' by bringing Christianity to the colonies (see Section 2 above). In subsequent periods, European empires carved up, for example, the African continent at the Berlin Conference in 1884 and in the 19th and early 20th century governed large parts of the world.

In order to exploit resources in the colonies and to widen their markets, Europeans aimed to access and order the colonies, relying on their own cultural and linguistic models. These were shaped by modernist concepts that regard the world as ordered by natural laws and that approach social categories—among them nations, ethnicities, and languages—as quasi-natural and essentialist (Bauman & Briggs, 2003b; Williams, 1999, p. 11). Colonisation therefore not only meant a territorial, physical, and bodily subjugation and exploitation but also dominance on the cultural and conceptual level:

> For settlers to possess the lands which they fondly constructed as 'vacant' they needed to map them, to name them in their own language, to describe

> and define them, to anatomize the land and its fruits, for themselves and the mother country, to classify their inhabitants, to differentiate them from other 'natives,' to fictionalize them, to represent them visually, to civilize and to cure them. (Hall, 2000, pp. 24–25, as quoted in McElhinny & Heller, 2020, p. 135)

In terms of social order, Europeans "were unable to break from their ideological heritage" and "implicitly believed their concept of ethnicity to be the natural order and not merely one convention amongst others used to make sense of the world" (Harries, 1989, p. 90). They therefore relied on "their own system of ethnic classification and accepted without question that Africans [and other colonised people] should use the same distinctions and concepts" (Harries, 1989, p. 90). What is today referred to as 'methodological nationalism'—the assumption that nation-states, with bounded, homogenous cultural and linguistic groups are the 'natural' way of organising human difference (Schneider, 2019; Wimmer & Glick Schiller, 2002)—was conceptually transposed to all regions of the world. Language ideologies that understand language as an index of ethnic and national communities were similarly seen as a 'natural' way of approaching human difference. Transforming language practices into writing through data collection procedures and into 'languages' through the production of dictionaries and grammar books was a way to tame, reify, and regularise colonial worlds (Deumert & Storch, 2020; see also next section). The practical need of translating the Bible into indigenous languages (especially in evangelical contexts; Gilmour, 2007, p. 1763) for the purposes of religious legitimisation of colonial exploitation encouraged the imagination and creation of territories ordered along ethno-linguistic lines. Since, for purely practical reasons, one linguistic repertoire had to be decided on as the language for Bible translation, this repertoire then came to be understood as the language of the specific territory (see, e.g., Durston, 2007, who discusses this process in the case of Quechua).

The development of AI technologies, in its ideological underpinnings and constructions of culture, is clearly different from the colonial endeavour. The brutal histories of slavery and exploitation have no equivalence and the cultural context is not framed in open statements about racial superiority. Rather, technologies are described as supporting individuals and communities to become integrated into markets and public spaces, and to profit from technological progress in various ways (Bapna et al., 2022; Costa-jussà et al., 2022). The political-ideological framing, at least currently and in Anglophone publica-

tions, draws on democratic, (neo)liberal and egalitarian ideals, which can also be inferred, for example, from the many publications from non-commercial academic authors but also from commercial actors that discuss biases and stigmatisation of minorities as a problem (e.g., Bolukbasi et al., 2016; Crawford, 2017; Sun et al., 2019, among others). Note that it is not always easy for a newcomer to distinguish industry publications and academic publications and often researchers from both the industry and from academia work together, also because only industrial actors have access to the data, algorithms, and computational supports (e.g., cloud credits) of companies which are essential for carrying out research. This observation already hints at some of the social hierarchies and exclusive tendencies that, despite discourses that value democracy, are implied in the field. Commercially-funded, non-peer reviewed content that fuses with academic knowledge has been referred to as the "manipulation of academia to avoid regulation" and some criticise that "the majority of well-funded work on 'ethical AI' is aligned with the tech lobby's agenda" (Ochigame, 2022, p. 54). This overlap between academic researchers and industry is actively encouraged in academia because it promises association with major discoveries and financial support for institutions.

Without saying that this would be comparable to colonial racism, there are obvious constructions of superiority, evolutionary ideologies, and stark power hierarchies in digital societies. Digital technologies are culturally associated with modernisation and progress, with the apparent neutrality of mathematics (Golumbia, 2009; Svensson, 2022), and with specific constructions of masculinity (Ensmenger, 2015; Wajcman, 2010). The ability to design code and to build and understand technology is associated with social authority. Making technologies accessible to as many people as possible is now typically discussed in terms of 'helping' others and industrial publications present the distribution of technology as a welfare activity (e.g., Bapna et al., 2022; Costa-jussà et al., 2022). Discourses of 'help', 'harm', and 'philanthropy' are regularly directed at communities with a colonial history and construct hierarchical relationships between communities. Furthermore, and this is probably even more crucial, access to technologies and their distribution to communities worldwide is a double-edged sword. While it does allow for many opportunities such as entry to digital public spaces, entertainment or ease of communication, companies do not build technologies out of philanthropic intentions—even if they like to present it that way.

Technology development is, at least in the western world, embedded in capitalist markets, in which companies give priority to economic profits.

Access to the data of customers is today an asset on this market and data is often referred to as 'the new gold' as "[p]ersonal data create economic and social value at an increasing pace, and personal information is used today in many different situations for numerous purposes" (García-Gasco Romero, 2021, p. 171). The creation of AI technology is one of these purposes. Observers speak of a 'race' between the five U.S.-based Big Tech companies (*Meta*, *Microsoft*, *Amazon*, *Apple*, and *Google*) to dominate AI on a global scale (Weise & Metz, 2023). Domination here is not of a traditional political kind but is, first of all, based on economic desires—global commercial actors are interested in data as data analysis allows them to make predictions on consumer behaviour, for example, in customised advertising (Rushkoff, 2019, p. 68). Yet, access to human behaviour through data collection and surveillance (Zuboff, 2019) is obviously a very powerful tool and therefore also of political interest. Governmental actors have funded AI development from its very beginning, first and foremost the U.S. *DARPA* (*Defense Advanced Research Projects Agency*; see Crawford, 2021, p. 184), and algorithmic intervention on social media platforms has played a tacit (illegal) role in democratic elections (Meredith, 2018). In China, algorithmic surveillance prediction is already used to control and form human behaviour in public and private spaces (Deng, 2023; Pei, 2024). Crawford observes that, also in the U.S. context, state and commercial interests become increasingly merged and that digital technologies in the United States

> encompass all those parts of everyday life that can be tracked and scored, grounded in normative definitions of how good citizens should communicate, behave and spend. This shift brings with it a different vision of state sovereignty, modulated by corporate algorithmic governance, and it furthers profound imbalance of power between agents of the state and the people they are meant to serve. (Crawford, 2021, p. 209)

A discourse of 'helping' is entangled in this fusion of state and commercial actors—the U.S. department's algorithmic warfare programme, for example, is based on *Microsoft* technologies and its motto, depicted on its logo, is 'Our job is to help' (Crawford, 2021, p. 190).

In the western world, the financial realisation of AI language technology is therefore not only in the hands of 'the Big Five' but co-funded by public institutions, including universities, and by funders from the financial sector, the oil and pharmaceutical industries, real estate, and others (Katz, 2020). The inter-

est in AI language technologies by some of the most powerful economic and political actors shows their political and economic relevance. The race of very few powerful actors to rule the world, to exploit global markets, and to control and influence human thought and behaviour is indeed reminiscent of the colonial endeavour.

And again, language plays a central but often hidden role in gaining access to, ordering and governing the world, in this case, in the form of digitised language data. Modernist concepts of language developed in the age of colonialism prevail also in computational contexts—as mentioned above, language is typically understood as appearing in orderly categories and structures, and as indexing territorially-based national or ethnic and monolingual communities. There are, however, also important differences as the different technological affordances of writing/printing and of digital online media bring about different theoretical approaches to language and different practices of materialising language. These will be discussed in the following.

3.2 Language theories and epistemologies

3.2.1 Concepts of language in European colonialism and in missionary activities

Language theory in general is dominated by concepts that have been developed by Europeans. As Deumert and Storch observe, "[c]olonial ideologies about language are rooted in a longue durée" (Deumert & Storch, 2020, p. 12), in which, since at least the beginning of the fourteenth century, language has been constructed as "codifiable, structured, and bounded" (drawing on Bonfiglio, 2013). The ability to create shared meaning interactively via sign-making practices (as identified in theories of languaging, see, e.g., Love, 2017; Makoni et al., 2020) became increasingly understood as springing from bounded systems, tied to specific (homogenous) peoples and territories. This epistemological framing of language had various effects. In European contexts, the claim to have 'a language' was taken as a sign for a culture to be 'real' and as having roots in a distant past, which until today serves to legitimate political autonomy. Gal and Irvine discuss the case of German, where the construction of a unified German language played a crucial role in political emancipation and the formation of the German nation-state in the late 19th century (see also Barbour & Carmichael, 2000; Gal & Irvine, 2019, Chapter 9).

In colonial settings, the imagination and mapping of languages as 'natural objects' "out there to be discovered" was seen as a way for "plotting histo-

ries and relations among peoples" (Gal & Irvine, 2019, p. 248). The description of linguistic practices thus served to produce "colonial categories of social difference and […] models of and for ethnocultural identities" (Errington, 2001a, p. 23). Europeans projected their monolingual concepts of ethnic culture and territory to contexts which were often shaped by much more complex relationships of language and social affiliation, in which multilingual resources were, for example, understood as linked to social rank, religion or occupation (e.g., Irvine, 1989). European colonists and missionaries thus (mis-) construed the linguistic repertoires they observed as an expression of traditional monolingual cultures, "locations in a distant past, but also their relations to some perduring place" (Errington, 2001a, p. 27). Linguistic diversity was interpreted as a result of migration and/or conquest and multilingualism as a possible sign for the 'unruliness' of a speaker (Gal & Irvine, 2019, Chapter 9). At the same time, language was key to accessing the minds and thoughts of colonial subjects, and translation became a central strategy to influence and convert them. Through the documentation of linguistic repertoires and the subsequent translation of the Bible into the resulting languages, which for the first time appeared in Roman alphabetic script, missionaries in particular contributed to the construction (or invention) of languages as territorially and ethnically grounded entities. In doing so, they co-constructed new ethnolinguistic groupings and new language-based socio-economic stratifications in which literate converts had the highest status (Errington, 2001a, p. 24).

The understanding of languages as naturalised stable entities, emerging from stable and timeless cultures, also brought about the idea that linguistic differences express a scale of civilisation. During the 19th century, languages were increasingly described as 'organisms', which also contributed to understanding linguistics as a 'natural science' (Arens, 1969). The 'family tree of languages' was invented (Schleicher, 1869, as cited in Arens, 1969) and describes linguistic and cultural relationships in a framing of enduring and purist family relationships, with 'parents' and 'brothers' and 'sisters' (Irvine, 1995), ignoring processes of intercultural contact and colonial realities that had brought about creolisation (Irvine, 1995). In such naturalised imaginations of language and culture, European languages were placed high on an evolutionary scale. Particularly European written languages were described as 'rational' and therefore superior. A concept of language as ideally serving for context-free and logical discourse, linked to logocentric ideologies in which words have stable and definite meanings and are understood to refer to a non-linguistic outside, prevailed in intellectual circles in early European modernity (Bauman & Briggs,

2000; 2003a). Such imaginations of language as 'rational' excluded women, the working classes, and the colonial Other (Bauman & Briggs, 2003b) because rationality was conceptualised as a property of educated white men who were also the main agents in the public domain where rational thought and debate were conceptualised to take place.

On the grammatical level, it was morpho-syntactic differences that were seen as indexical of civilisational scales. Because Latin was taken as the reference model, and because all the terminology used to describe grammar derived from the description of Latin, the analysis of the morpho-syntax of other languages was biased and skewed towards Latin. Wilhelm von Humboldt, for example, studied typological differences of languages and interpreted more synthetic languages—languages in which grammatical information is expressed via morphological processes within a word—to be more expressive and complex than analytic languages in which grammatical information tends to be expressed in individual words (von Humboldt, 1836). Grammatical forms were also seen as 'window' into the human mind and, in contexts of colonial racism, specific grammatical forms, and particularly more analytic forms, were taken as sign of the inferior cognitive capacities of non-European speakers—"[i]t thus became customary to speak of primitive languages, in the same way some races were considered evolutionary inferior to others" (Mufwene, 2015, p. 453).

Although Latin was regarded as ideal and as an underlying reference for grammatical descriptions, the constructions of hierarchy in colonial language theory had an effect on the perception of other European languages. The fact that colonialism not only produced culture and language in the colonies but co-constructed imaginations of European culture is one of the important insights of postcolonial theory (Said, 1978). Thus:

> [i]f one of the central aspects of colonial discourse has been to construct the native Other as backward, dirty, primitive, depraved, childlike, feminine, and so forth, the other side of this discourse has been the construction of the colonizers, their language, culture and political structures as advanced, superior, modern, civilized, masculine, mature and so on. (Pennycook, 1998, p. 129)

European languages, and, over time, above all English, became markers of their speakers' "'progress', 'enlightenment' and 'enrichment'" (Gilmour, 2007, p. 1765).

These Euro-American developmentalist ideologies have had enduring effects on language policy in postcolonial nations. Overall, 20th century global language policy mostly reproduced ideas of territoriality and of language as referential, 'rational' tool. Corpus and status planning for non-European languages has also been carried out with ideals of homogeneity, efficiency, and simplicity in mind. An underlying teleological ideology often assumes that any language strives towards the ideal end form of an official and standard written language that can be used for academic purposes (Errington, 2001a, p. 34). This language should then fulfil the role of a 'neutral' "voice from nowhere" (Gal & Woolard, 2001) in the national or ethnic context where it is understood to originate.

Thus, the work of transforming languages into writing, which is then understood as an indexical representation of an ethnic and territorially-based group is still ongoing. And until today, missionary work that aims to spread Christian religious beliefs continues to play an important role in this context. Crucial here is the *Summer Institute of Linguistics* (*SIL*)[8] which was established in the United States in 1934. It provides "linguistic, anthropological and sociolinguistic expertise to aspiring Bible translators" (Kamusella, 2012, p. 71) and is closely connected to *Wycliffe Bible Translators*,[9] whose goal is to disseminate Christianity through translation of the Bible into as many languages as possible. *SIL* shapes conceptualisations of language particularly in the Global South in several key ways. It offers training in the different activities that are part of this process, ranging from language and culture description, literacy development, academic publishing, translation practices, Bible study to publishing and dissemination of its products.[10] Its impact is non negligible in that it has trained and supported over 5000 missionary linguists from the Global North and the South and impacted more than a 1000 languages, mostly in the Global South (Errington, 2008). According to its own figures, its current activities are impacting more than "855+ million people, 1341 communities and 98 countries".[11] *SIL* is also active in the technology-enhanced development of written codes out of oral language practices: "SIL software supports fieldwork in linguistics and anthropology by streamlining collection, analysis and archiv-

8 http://www.sil.org
9 https://www.wycliffe.org/
10 https://www.sil.org/about/discover
11 http://www.sil.org/

ing of language and culture data".[12] Finally, *SIL* also has a leading role in systematising linguistic diversity through its coordination, editing, and publishing of *Ethnologue*, "the single largest, most widely cited compendium of knowledge of global linguistic diversity" (Errington, 2008, p. 153).

Although *Ethnologue* was initially developed by *SIL* to guide and provide background for its own Bible translation activities, the *International Organization for Standardization* first "invited SIL to develop an ISO 639–3 standard to cover all the world's languages" and subsequently made *SIL* its code registration and allocation agency:[13]

> This code allocation is the actual uniformized world-wide registration of languages that amounts to their de facto international recognition. It also appears to be de jure (though not overtly articulated as) recognition in light of international law, insofar as the International Organization for Standardization creates and maintains elements and procedures of this law. (Kamusella, 2012, p. 72)

ISO-codes are used to implement different languages into computing systems. This means that the colonial, missionary activities of transforming interactive practices of meaning-making into writing are brought into a further reification in digital culture.

3.2.2 Concepts of language in AI culture

The above observations show that there is a direct link between colonial linguistics and digital language culture (although openly racist language theories have been abandoned). In technology settings, as in European colonialism, language is imagined as deriving from ethnically homogenous groups that can be orderly mapped in territorial space. Accounts of language entail the idea that a 'fully developed' language profits from a standard writing system and efforts are made to create writing systems where these do not exist yet. Languages for which a certain degree of normalisation cannot be achieved are usually denied inclusion into AI processes such as machine translation (Costa-jussà et al., 2022, pp. 12–18). In contrast to colonial times, there is no discourse that describes distinct languages or grammatical forms as 'more or less developed', instead linguistic diversity is generally described as 'wealth' (van Esch et al.,

12 http://www.sil.org/about/discover/technology-language-development

13 http://www.iso.org/iso-639-language-code

2019). However, evolutionary ideologies and teleological concepts are still common, but they are reconfigured in relation to the affordances of digital systems.

Languages that are embedded in computational systems are typically referred to with ISO-Codes. It is common in digital culture to equal language (as a general human phenomenon) with datafied language, that is, language that has been rendered into machine-readable text.[14] Only datafied language can be used for documenting and predicting user behaviour or for training AI systems. An understanding of language as digital data is common-sense so that the fact that language is based on meaning-making practices that humans produce can be easily forgotten. This can be shown in the following short passage from a text by *Stanford's Center for Research on Foundation Models* (*CRFM*): "[…] we highlight the role of people as the ultimate source of data" (Bommasani et al., 2021, p. 7). The wording displays the theoretical grounding of widespread discourses on language in contemporary AI culture. Readers are reminded that 'people' are the 'source of data'. It is typical in computational publications that language is not discussed as human interactive praxis but as a data source.

In contrast to colonial times, languages are not distinguished based on their typological characteristics but based on the size of their data sets. These become a central marker of differentiation and 'development'. Depending on the size of written and datafied text corpus, languages are therefore approached on a scale from 'high-resource' to 'low-resource' (e.g., Bommasani et al., 2021; Costa-jussà et al., 2022). There are long lists of languages, ordered according to the number of words that have been datafied (Bapna et al., 2022). Unsurprisingly, 'low-resource languages' are typically languages with colonial histories (but also other minority languages or those simply not aligned with administrative units such as states). These languages are described as in need of 'help'. A prominent paper by *Meta*, the company that owns *Facebook*, *Instagram*, and *WhatsApp*, carries the title *No Language Left Behind* (*NLLB*; Costa-jussà et al., 2022) and discusses ways to include 'low-resource languages' into

14 Datafication consists of extracting information from the flow of social life, identifying it with imagined social categories and fixing such relationships. It is part of the ideology of 'dataism' (Bode & Goodlad, 2023), which broadly assumes that data represents human behaviour and that quantification and its agents are objective. In linguistics, datafication has involved collecting of oral practices via recording and transforming it into writing. This process has been instrumental in conceptualising "language as referential code and languages as 'natural', given objects that are systematically and neatly structured (e.g., Pennycook 2004)" (Erdocia et al., 2024).

digital technologies. Promotional videos of the inclusion of these languages into *Meta*'s technologies celebrate linguistic diversity—and its datafication as welfare and 'support' (for a more detailed discussion, see Schneider, in press). Such discourses reproduce missionary relationships but also colonial language ideologies that understand languages as systemic entities, ordered according to ethnic communities. They also construct new hierarchical social relationships based on access to technology, in which large U.S.-based technology companies are the unquestioned leaders.[15]

In the provision of language technologies for 'low-resource languages', there is an important emphasis on classifying and ordering languages for validating those that are targeted for inclusion in AI, which reproduces the colonial idea of language as worthy if written and adds the criterion of 'written in Wikipedia'. To obtain a list of 200 lesser-used languages for developing machine translation, the above-mentioned *Meta* (*Facebook*) consortium (Costa-jussà et al., 2022, p. 12) created "a preliminary list of over 250 possible language candidates" based on the following considerations:

> First, we considered all languages with a Wikipedia presence. [...] Next, we solicited lists of languages spoken in various regions by native speakers, focusing particularly on African languages—a category of languages that have historically been underrepresented in translation efforts [...]. We then examined language coverage in multiple existing datasets in the natural language processing community, paying focused attention on training datasets without accompanying evaluation datasets. Finally, we considered [...] the approximate number of native speakers and other community-level variables relevant to our work. Next, for each of the language candidates, we partnered with linguists from various specialized language service providers to understand if each of these languages has a standardized written form [...] because having a reliable, high-quality evaluation dataset is critical to accelerated experimental progress. (Costa-jussà et al., 2022, pp. 12, 16)

Once the languages were selected for intervention, *NLLB* cross-referenced, that is, compared, the information with existing language regimes from the Global North such as the ISO codes and those from *Glottolog*, which are developed and administered by a group of typological linguists.[16] Both sets of codes reproduce a colonial enumerative and differentiation approach, focusing

15 At least in western settings and here not discussing the case of China.

16 https://glottolog.org/

on identifying (inventing?) bounded languages, and mapping them in space and time. Technology companies also apply additional metrics. For example, *Meta* classifies languages by recording their web support—whether they are supported by *Google Translate* and *Microsoft Translate*, the types of scripts in use, their resource-level, i.e., "if there are fewer than 1M publicly available, de-duplicated bitext samples with any other language within our set of 200 languages" (Costa-jussà et al., 2022, p. 17), and by assigning them a unique name "due to formatting limitations".

In these techno-colonial discourses and enumerative logics, it is no longer Latin that serves as reference model but English plays a hegemonic role. The data set for English is by far the largest for all languages in the world. Currently, more than half of the language data appearing on the web is in English. As machine learning is above all based on web-scraped online data, machine learning language technologies work best in Standard English (which is also true for voice technology, Markl, 2022). Languages typologically different from English are discussed as a technical problem (e.g., Costa-jussà et al., 2022, p. 79). Monolingual data that entails stylistic and genre variation is treated as 'clean' and 'rich', while non-monolingual data is referred to with the metaphor 'dirty' and it thus requires 'cleaning' (see also, e.g., Kreutzer et al., 2022). The hierarchical metaphorical framing of English is depicted vividly in the following quote:

> English data is not only orders of magnitude more abundant than that of lower-resource languages, but it is often cleaner, broader, and contains examples showcasing more linguistic depth and complexity. (Bommasani et al., 2021, p. 25)

Besides the very large English language dataset, the perceived 'cleanliness' of English is based on the fact that it is the unmarked medium of communication for global, expert and academic (human to human) interaction in technology and academia. It thus has been criticised that English is often equated with 'language' in general in digital culture. The critical computational linguist Bender therefore proposed the so-called 'Bender Rule', which suggests that speakers at computational sciences conferences should explicitly mention the language they are talking about, based on the observation that the equation of English with 'language' often goes unnoticed (Bender, 2019; Schneider, 2022). Technological innovation is typically first developed for products that serve English-speaking markets, which means that users whose language has not

been integrated into the systems often use English-language products, but also users who are keen to always use the newest update of a tool (Leblebici & Schneider, 2025). This means that the dataset for English is expanding even more. Finally, English terms are used in almost all programming languages. The dominance of English in the world of computation has the effect that languages other than English are sometimes discursively constructed as one category—the above-mentioned Stanford research paper, for example, uses the term 'non-English languages' (Bommasani et al., 2021) and thus constructs an English-non-English binary.

4. Conclusion: Global tech colonialism and how to overcome it

Although AI-driven language technologies emerged in a context that is different from colonial times, this paper argues that they reproduce colonialist legacies at several levels. While couched in a discourse of 'helping' and 'change for a better future', big tech companies exploit language data in order to make monetary profits: technologies are first and foremost commercial products. Like their colonial forefathers, big tech companies are also not interested in understanding language and its use or their speakers per se. In fact, human interactive practices and linguistic traditions are not seen as an expression of culture or identity but as a data source. These data sources are seen as a resource for developing powerful and profit-making tools to manage people. This is done through refashioning everyday language practices into data, which is enabled by processes of ordering, homogenisation, fixing, and alignment according to models based on European cultural concepts of language and practices that emerged during colonial times. Local visions of language and culture are marginalised or even completely stamped out through processes of alignment with digitally powerful languages, English first and foremost.[17]

The social hierarchies that existed during colonial times also continue, though in a partially reconfigured manner. As in colonial times, languages from the Global South are positioned as needing 'development' to enable the 'development' for their speakers, in this case their integration into the commercial and knowledge economy championed by companies of the Global

17 Due to space constraints, we have not discussed these in detail here. For further insights into non-European language concepts, see, for example, Schneider (2021) and Migge (2020).

North. Decisions to 'technologically develop' a language generally do not originate with speaker communities and their needs. Instead, languages are 'selected' by companies based on technological and commercial logics such as easy access to sufficiently large corpora, the existence of technologically literate speaker communities that have sufficient purchasing power and that are sufficiently unified in their use of orthographies.

Access to language technologies such as automatic machine translation has the potential to give people access to a wider range of knowledge sources, experiences, and ways of connecting with people. However, unlike the development of writing and literacy practices in the colonial era, language technologies of the digital era are much more heavily dependent on the infrastructures of commercial companies. Their development and maintenance require costly tools that are owned only by a small number of companies and who can grant and refuse access at will. They also do not share their designs and practices, and the use of their infrastructures usually requires giving up control of existing tools developed by the grassroots and subscribing, at a cost, to their tools. At the same time, the global collection and surveillance of data in ever more languages, in the hands of very few actors, is the continuation of the desire to construct a universal world order. We observe an evolutionary teleology that strives for a data set of 'N=all' and, as media sociologists critically argue: "What it exploits is our lives as human beings" (Mejias & Couldry, 2024, p. 33). Overall, the dominance of U.S. corporations in global technology provision and data politics means that "[o]ur era is attempting to bring back into fashion the old myth that the West alone has a monopoly on the future" (Mbembe, 2021; see also Birhane, 2020).

At the time of writing this text, there is still little public awareness of the political and economic relevance of language as data, and of the colonial ideologies embedded in digital and AI infrastructures. Colleagues from or working on the Global South are taking the lead here and critically discuss the power of language data (e.g., Birhane, 2020; Markl et al., 2023; Miceli & Posada 2022). Several initiatives aim to put data collection in public hands, with Maori activists being a prime example. The Maori community has refused to allow global companies to collect their language data and has produced its own data sets. They argue that "Our data would be used by the very same people that beat that language out of our mouths to sell it back to us as a service [...]. It's

just like taking our land and selling it back to us" (Hao, 2022).[18] All in all, this does not mean that we should stop using technology or that the development of language technology should come to an end. It means that, in a democratic society, the infrastructures that frame and shape public life and discourse should be in the hands of those who use them, not in the hands of a monopoly of capitalist corporations.

References

Arens, H. (1969). *Sprachwissenschaft: Der Gang ihrer Entwicklung von der Antike bis zur Gegenwart* (Band 1: Von der Antike bis zum Ausgang des 19. Jahrhunderts). Athenäum Fischer Taschenbuchverlag.

Bang, Y., Cahyawijaya, S., Lee, N., Dai, W., Su, D., Wilie, B., Lovenia, H., Ji, Z., Yu, T., Chung, W., Do, Q. V., Xu, Y., & Fung, P. (2023). A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. *arXiv*. https://doi.org/10.48550/arXiv.2302.04023

Bapna, A., Caswell, I., Kreutzer, J., Firat, O., van Esch, D., Siddhant, A., Niu, M., Baljekar, P., Garcia, X., Macherey, W., Breiner, T., Axelrod, V., Riesa, J., Cao, Y., Chen, M. X., Macherey, K., Krikun, M., Wang, P., Gutkin, A., … Hughes, M. (2022). Building machine translation systems for the next thousand languages. *arXiv*. https://arxiv.org/abs/2205.03983

Barbour, S., & Carmichael, C. (Eds.) (2000). *Language and nationalism in Europe*. Oxford University Press. https://doi.org/10.1093/oso/9780198236719.001.0001

Bauman, R., & Briggs, C. L. (2000). Language philosophy as language ideology: John Locke and Johann Gottfried Herder. In P. V. Kroskrity (Ed.), *Regimes of language: Ideologies, polities, and identities* (pp. 139–204). School of American Research Press.

Bauman, R., & Briggs, C. L. (2003a). Making language and making it safe for science and society: From Francis Bacon to John Locke (Chapter 2). In *Voices of modernity: Language ideologies and the politics of inequality* (pp. 19–69). Cambridge University Press. https://doi.org/10.1017/CBO9780511486647.003

18 See also Coffey (2021); for more initiatives, see https://rising.globalvoices.org, https://commonvoice.mozilla.org/ca

Bauman, R., & Briggs, C. L. (2003b). *Voices of modernity: Language ideologies and the politics of inequality*. Cambridge University Press. https://doi.org/10.1017/CBO9780511486647

Bayly, S. (2016). Colonialism / postcolonialism. In F. Stein (Ed.), *The Open Encyclopedia of Anthropology.* http://doi.org/10.29164/16colonialism

Bender, E. M. (2019, September 14). The #Bender Rule: On naming the languages we study and why it matters. *The Gradient.* https://thegradient.pub/the-benderrule-on-naming-the-languages-we-study-and-why-it-matters/

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. https://doi.org/10.1145/3442188.3445922

Birhane, A. (2020). Algorithmic colonization of Africa. *SCRIPTed, 17*(2), 389–409. https://doi.org/10.2966/scrip.170220.389

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

Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V., & Kalai, A. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. *NIPS'16: Proceedings of the 30th Conference on Neural Information Processing Systems*, 4356–4364. https://dl.acm.org/doi/10.5555/3157382.3157584

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N. S., Chen, A. S., Creel, K. A., Davis, J., Demszky, D., … Liang, P. (2021). On the opportunities and risks of foundation models. *arXiv*. https://doi.org/10.48550/arXiv.2108.07258

Bonfiglio, T. P. (2013). Inventing the native speaker. *Critical Multilingualism Studies, 1*(2), 29–58.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). Language models are few-shot learners. *arXiv*. https://doi.org/10.48550/arXiv.2005.14165

Bunz, M. (2012). *Die stille Revolution*. Suhrkamp.

Castelle, M. (2023, June 30). *Transformer models and their language ideological implications* [Video]. Talk given at COST Language in the Human-Machine Era Working Group 'Ideologies, Beliefs, Attitudes'. Available upon request.

Ceruzzi, P. E. (2003). *A history of modern computing* (2nd ed.). MIT Press.

Cheyfitz, E., & Harmon, A. (2018). Translation and colonialism. In J. Evans & F. Fernandez (Eds.), *The Routledge handbook of translation and politics* (pp.270–286). Routledge.

Coffey, D. (2021, April 28). Māori are trying to save their language from Big Tech. *Wired*. https://www.wired.co.uk/article/maori-language-tech

COST Action CA19102. (n.d.). *Language in the human-machine era (LITHME)*. Retrieved September 28, 2024, from https://lithme.eu/

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., Akula, B., Barrault, L., Mejia Gonzalez, G., Hansanti, P., … Wang, J. (2022). No Language Left Behind: Scaling Human-Centered Machine Translation. *ArXiv*. https://doi.org/10.48550/arXiv.2207.04672

Couldry, N., & Hepp, A. (2017). *The mediated construction of reality*. Polity.

Crawford, K. (2017, December 5). *The trouble with bias – NIPS 2017 Keynote – Kate Crawford* [Video]. YouTube. The Artificial Intelligence Channel. https://youtu.be/fMym_BKWQzk?si=SMyhYTAynPUR-5G9

Crawford, K. (2021). *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. Yale University Press.

Deng, B. (2023, May 6). In China, AI cameras alert police when a banner is unfurled. *Radio Free Asia*. https://www.rfa.org/english/news/china/surveillance-06052023142155.html

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 Academic. https://doi.org/10.1093/oso/9780198793205.003.0001

Durston, A. (2007). *Pastoral Quechua: The history of Christian translation in colonial Peru, 1550–1654*. University of Notre Dame Press. https://doi.org/10.2307/j.ctvpg8689

Ensmenger, N. (2015). "Beards, sandals, and other signs of rugged individualism": Masculine culture within the computing professions. *Osiris, 30*(1), 38–65. https://doi.org/10.1086/682955

Erdocia, I., Migge, 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*, *28*(4), 20–25. https://doi.org/10.1111/josl.12680

Errington, J. (2001a). Colonial linguistics. *Annual Review of Anthropology, 30*, 19–39. https://doi.org/10.1146/annurev.anthro.30.1.19

Errington, J. (2001b). Ideology. In A. Duranti (Ed.), *Key terms in language and culture* (pp. 110–112). Blackwell.

Errington, J. (2008). *Linguistics in a colonial world. A story of language, meaning and power*. Blackwell. https://doi.org/10.1002/9780470690765

Fabian, J. (1986). *Language and colonial power: The appropriation of Swahili in the former Belgian Congo 1880–1938*. Cambridge University Press.

Gal, S., & Irvine, J. T. (2019). *Signs of difference: Language and ideology in social life*. Cambridge University Press. https://doi.org/10.1017/9781108649209

Gal, S., & Woolard, K. A. (2001). Constructing languages and publics: Authority and representation. In S. Gal & K. A. Woolard (Eds.), *Languages and publics: The making of authority* (pp. 1–12). Routledge.

García-Gasco Romero, M. (2021). Personal data: The new black gold. In J. M. Ramírez, & B. Bauzá-Abril (Eds.), *Security in the global commons and beyond* (pp. 171–182). Springer.

Gilmour, R. (2007). Missionaries, colonialism and language in nineteenth-century South Africa. *History Compass*, *5*(6), 1761–1777. https://doi.org/10.1111/j.1478-0542.2007.00472.x

Golumbia, D. (2009). *The cultural logic of computation*. Harvard University Press. https://doi.org/10.4159/9780674053885

Grieve, J., Bartl, S., Fuoli, M., Grafmiller, J., Huang, W., Jawerbaum, A., Murakami, A., Perlman, M., Roemling, D, Winter, B. (2024). The sociolinguistics foundations of language modeling. *arXiv*. https://doi.org/10.48550/arXiv.2407.09241

Hall, C. (2000). Introduction: Thinking the postcolonial, thinking the empire. In C. Hall (Ed.), Cultures of empire: A reader. Colonizers in Britain and the empire in the nineteenth and twentieth centuries (pp. 1–37). Manchester University Press.

Hao, K. (2022, April 22). A new vision of artificial intelligence for the people: ln a remote rural town in New Zealand, an Indigenous couple is challenging what AI could be and who it should serve. *MIT Technology Review*. https://www.technologyreview.com/2022/04/22/1050394/artificial-intelligence-for-the-people/

Harries, P. (1989). Exclusion, classification and internal colonialism: The emergence of ethnicity among the Tsonga-Speakers of South Africa. In L. Vail (Ed.), *The creation of tribalism in Southern Africa* (pp. 82–117). University of California Press.

Heaven, W. D. (2023, August 30). AI hype is built on high test scores: Those tests are flawed. *MIT Technology Review*. https://www.technologyreview.co

m/2023/08/30/1078670/large-language-models-arent-people-lets-stop-testing-them-like-they-were/

Heller, M., & McElhinny, B. (2017). *Language, capitalism, colonialism: Toward a critical history*. University of Toronto Press.

Irvine, J. T. (1989). When talk isn't cheap: Language and political economy. *American Ethnologist*, *16*(2), 248–267. https://doi.org/10.1525/ae.1989.16.2.02a00040

Irvine, J. T. (1995). The family romance of colonial linguistics: Gender and family in nineteenth-century representations of African languages. *Pragmatics*, *5*(2), 139–153. https://doi.org/10.1075/prag.5.2.02irv

Irvine, J. T., & Gal, S. (2000). Language ideology and linguistic differentiation. In P. V. Kroskrity (Ed.), *Regimes of Language: Ideologies, polities and identities* (pp. 35–83). School of American Research Press.

Kamusella, T. (2012). The global regime of language recognition. *International Journal of the Sociology of Language*, *2012*(218), 59–86. https://doi.org/10.1515/ijsl-2012-0059

Katz, Y. (2020). *Artificial whiteness: Politics and ideology in artificial intelligence*. Columbia University Press. https://doi.org/10.7312/katz19490

Keane, W. (2024). *Animals, robots, gods: Adventures in the moral imagination*. Allen Lane.

Kreutzer, J., Caswell, I., Wang, L., Wahab, A., van Esch, D., Ulzii-Orshikh, N., Tapo, A., Subramani, N., Sokolov, A., Sikasote, C., Setyawan, M., Sarin, S., Samb, S., Sagot, B., Rivera, C., Rios, A., Papadimitriou, I., Osei, S., Ortiz Suarez, P., … Adeyemi, M. (2022). Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, *10*, 50–72. https://doi.org/10.1162/tacl_a_00447

Leblebici, D., & Schneider, B. (2025). Digital assemblages and their English entanglements: Digital design, voice assistant use and smartphone setting choices of translingual speakers in Berlin. In J. Won Lee & S. Rüdiger (Eds.), *Entangled Englishes*. Routledge.

Liang, W., Izzo, Z., Zhang, Y., Lepp, H., Cao, H., Zhao, X., Chen, L., Ye, H., Liu, S., Huang, Z., McFarland, D., & Zou, J. Y. (2024). Monitoring AI-modified content at scale: A case study on the impact of ChatGPT on AI conference peer reviews. *Proceedings of the 41st International Conference on Machine Learning, Proceedings of Machine Learning Research*, *235*, 29575–29620. https://proceedings.mlr.press/v235/liang24b.html

Love, N. (2017). On languaging and languages. *Language Sciences, 61*, 113–147. https://doi.org/10.1016/j.langsci.2017.04.001

Makoni, S. B., Severo, C. G., & Abdelhay, A. (2020). Colonial linguistics and the invention of language. *Language planning and policy: Ideologies, ethnicities, and semiotic spaces of power*, 211–228.

Marche, S. (2022, December, 6). The college essay is dead: Nobody is prepared for how AI will transform academia. *The Atlantic.* https://www.theatlantic.com/technology/archive/2022/12/chatgpt-ai-writing-college-student-essays/672371/.

Markl, N. (2022). Mind the data gap(s): Investigating power in speech and language datasets. In B. R. Chakravarthi, B. Bharathi, J. P. McCrae, M. Zarrouk, K. Bali, & P. Buitelaar (Eds.), *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion* (1–12). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.ltedi-1.1

Markl, N., Wallington, E., Klejch, O., Reitmaier, T., Bailey, G., Pearson, J., Jones, M., Robinson, S., & Bell, P. (2023). Automatic transcription and (de)standardisation. In Proceedings – SIGUL 2023, 2nd Annual meeting of the Special Interest Group on Under-resourced Languages: A satellite workshop of Interspeech 2023 (pp. 93–97). International Speech Communication Association. https://doi.org/10.21437/SIGUL.2023-20

Mbembe, A. (2021). *Out of the dark night: Essays on decolonization*. Columbia University Press.

McElhinny, B., & Heller, M. (2020). The linguistic intimacy of five continents: Racializing language in empire. In H. S. Alim, A. Reyes, & P. V. Kroskrity (Eds.), *The Oxford handbook of language and race* (pp. 130–152). Oxford Academic. https://doi.org/10.1093/oxfordhb/9780190845995.013.8

McIntosh, D. (2019). We need to talk about data: How digital monopolies arise and why they have power and influence. *Journal of Technology Law & Policy*, *23*(2), 185–213.

McShane, M., & Nirenburg, S. (2021). *Linguistics for the age of AI*. MIT Press. https://doi.org/10.7551/mitpress/13618.001.0001

Mejias, U. & Couldry, N. (2024). *Data grab: The new colonialism of big tech and how to fight back*. University of Chicago Press.

Meredith, S. (2018, April 10). *Facebook-Cambridge Analytica: A timeline of the data hijacking scandal*. CNBC. https://www.cnbc.com/2018/04/10/facebook-cambridge-analytica-a-timeline-of-the-data-hijacking-scandal.html

Miceli, M., & Posada, J. (2022). The data-production dispositif. *Proceedings of the ACM Human-Computer Interaction*, *6*(CSCW2), Article 460, 1–37. https://doi.org/10.1145/3555561

Migge, B. (2020). Broadening creole studies: From grammar towards discourse. *Journal of Pidgin and Creole Languages*, *35*(1), 160–177. https://doi.org/10.1075/jpcl.00050.mig

Mufwene, S. (2015). Race, racialism, and the study of language evolution in America. In M. D. Picone & C. E. Davies (Eds.), *New perspectives on language variety in the south: Historical and contemporary perspectives* (pp. 449–474). University of Alabama Press.

Nebrija, A. (1492). *Gramática de la lengua castellana*. https://archive.org/details/A336029/page/n143/mode/2up

Nguyen, D., Doğruöz, A. S., Rosé, C. P., & de Jong, F. (2016). Computational sociolinguistics: A survey. *Computational Linguistics*, *42*(3), 537–593. https://doi.org/10.1162/COLI_a_00258

Noble, S. U. (2018). *Algorithms of oppression: How search engines reinforce racism*. NYU Press. https://doi.org/10.2307/j.ctt1pwt9w5

Ochigame, R. (2022). The invention of 'ethical AI': How big tech manipulates academia to avoid regulation. In T. Phan, J. Goldfein, D. Kuch, & M. Mann (Eds.), *Economics of virtue: The circulation of 'ethics in AI'* (pp. 49–56). Institute of Network Cultures.

Pasquinelli, M. (2023). *The eye of the master: A social history of artificial intelligence*. Verso Books.

Pei, M. (2024). *The sentinel state: Surveillance and the survival of dictatorship in China*. Harvard University Press. https://doi.org/10.2307/jj.10860939

Pennycook, A. (1998). *English and the discourses of colonialism*. Routledge.

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

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

Said, E. W. (1978). *Orientalism*. Routledge.

Schleicher, A. (1869). *Darwinism tested by the science of language*. John Camden Hotten.

Schmidt-Brücken, D., Schuster, S., Stolz, T., Warnke, I. H., & Wienberg, M. (Eds.). (2015). *Koloniallinguistik. Sprache in kolonialen Kontexten*. de Gruyter. https://doi.org/10.1515/9783110424799

Schneider, B. (2019). Methodological nationalism in linguistics. *Language Sciences*, *76* https://doi.org/10.1016/j.langsci.2018.05.006

Schneider, B. (2021). Creole prestige beyond modernism and methodological nationalism: Multiplex patterns, simultaneity and non-closure in the sociolinguistic economy of a Belizean village. *Journal of Pidgin and Creole Languages*, *36*(1), 12–45. https://doi.org/10.1075/jpcl.00068.sch

Schneider, B. (2022). Multilingualism and AI: The regimentation of language in the age of digital capitalism. *Signs and Society*, *10*(3), 362–387. https://doi.org/10.1086/721757

Schneider, B. (2024). A sociolinguist's look at the "language" in Large Language Models. *Critical AI*, *2*(1). https://doi.org/10.1215/2834703X-11205168

Schneider, B. (in press). Transnational voices from nowhere leave no one behind: Hierarchical chronotopes in AI language culture. In S.-Y. Park & B. Bolander (Eds.), *Language and Transnationalism*.

Shaitarova, A., Göhring, A., & Volk, M. (2023). Machine vs. human: Exploring syntax and lexicon in German translations, with a spotlight on anglicisms. In T. Alumäe & M. Fishel (Eds.), *Proceeding of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)* (pp. 215–227). University of Tartu Library. https://aclanthology.org/2023.nodalida-1.22

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. https://doi.org/10.1017/CBO9781139342872.007

Sun, T., Gaut, A., Tang, S., Huang, Y., ElSherief, M., Zhao, J., Mirza, D., Belding, E., Chang, K.-W., & Wang, W. Y. (2019). Mitigating gender bias in natural language processing: Literature review. In A. Korhonen, D. Traum, & L. Màrquez (Eds.), Proceedings of the 57th annual meeting of the Association for Computational Linguistics (pp. 1630–1640). Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1159

Svensson, J. (2022). Modern mathemagics: Values and biases in tech culture. In M. Filimowicz (Ed.), *Systemic bias: Algorithms and society* (pp. 21–39). Routledge. https://doi.org/10.4324/9781003173373-2

Tarnoff, B. (2022). *Internet for the people: The fight for our digital future*. Verso.

Toh, M. (2023, March 29). 300 million jobs could be affected by latest wave of AI, says Goldman Sachs. *CNN Business*. https://edition.cnn.com/2023/03/29/tech/chatgpt-ai-automation-jobs-impact-intl-hnk/index.html

van Esch, D., Sarbar, E., Lucassen, T., O'Brien, J., Breiner, T., Prasad, M., Crew, E., Nguyen, C., & Beaufays, F. (2019). Writing across the world's languages: Deep internationalization for Gboard, the Google keyboard. *arXiv*. https://doi.org/10.48550/arXiv.1912.01218

Vanmassenhove, E., Shterionov, D., & Way, A. (2019). Lost in translation: Loss and decay of linguistic richness in machine translation. In M. Forcada, A. Way, B. Haddow, R. Sennrich (Eds.), *Proceedings of machine translation*

Summit XVII: Research Track (222–232). European Association for Machine Translation. https://aclanthology.org/W19-6622

Virtanen, A., Kanerva, J., Ilo, R., Luoma, J., Luotolahti, J., Salakoski, T., Ginter, F., & Pyysalo, S. (2019). Multilingual is not enough: BERT for Finnish. *arXiv*. https://doi.org/10.48550/arXiv.1912.07076

von Humboldt, W. (1836). *Über die Verschiedenheit des menschlichen Sprachbaus und ihren Einfluß auf die geistige Entwicklung des Menschengeschlechts*. Königliche Akademie der Wissenschaften.

Wajcman, J. (2010). Feminist theories of technology. *Cambridge Journal of Economics*, *34*(1), 143–152. https://doi.org/10.1093/cje/ben057

Warnke, I. H., Stolz, T., & Schmidt-Brücken, D. (2016). Perspektiven der Postcolonial Language Studies. In T. Stolz, I. H. Warnke, D. Schmidt-Brücken (Eds.), *Sprache und Kolonialismus: Eine interdisziplinäre Einführung zu Sprache und Kommunikation in kolonialen Kontexten* (pp. 1–26). de Gruyter. https://doi.org/10.1515/9783110370904-001

Weise, K., & Metz, C. (2023, December 8). The race to dominate A.I. *The New York Times*. https://www.nytimes.com/2023/12/08/briefing/ai-dominance.html

Williams, G. (1999). *French discourse analysis: The method of post-structuralism*. Routledge.

Wimmer, A., & Glick Schiller, N. G. (2002). Methodological nationalism and beyond: Nation-state building, migration and the social sciences. *Global Networks*, *2*(4), 301–334. https://doi.org/10.1111/1471-0374.00043

Woolard, K. A. (1998). Introduction. Language ideology as field of inquiry. In B. B. Schieffelin, K. A. Woolard, & P. V. Kroskrity (Eds.), *Language ideologies: Practice and theory* (pp. 3–48). Oxford University Press.

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