International Journal of Language Studies
Volume 18, Number 2, April 2024, pp. 91-112 | DOI: 10.5281/zenodo.10475280

Linguistic diversity in institutional collections: Beyond preservation to valorisation

Ewan D. HANNAFORD, University of Glasgow, United Kingdom [1]
Marc ALEXANDER, University of Glasgow, United Kingdom

Language's capacity to shape and perpetuate ideologies, cultural values, and social conditions is well-established across linguistic theory. From this perspective, combatting linguistic prejudice and promoting language equity are key to contemporary cultural concerns around challenging prescriptivist worldviews and disrupting hegemonic historical perspectives. Institutional collections represent promising staging grounds for such efforts, with wide reach and accessibility, but are typically focused and curated in mainstream language varieties. This paper explores how institutional collections may correct this homogeneity, through connecting materials containing regional/social language varieties, including those of community archives, into collections more representative of diverse linguistic and cultural landscapes. Using an AHRC-funded project integrating community-generated content into the UK national collection as example, this paper addresses challenges and makes recommendations for effectively valorising language varieties in institutional collections. Consequently, this paper argues for the potential of linguistically diverse institutional collections as transformative tools for promoting language equity and reducing linguistic prejudice.

Keywords: Community Archives; Institutional Collections; Language Ideology; Linguistic Prejudice; Linguistic Variety

1. Introduction

It is widely established across linguistic theory that language, among its many other functions, serves as a powerful tool for the promotion and perpetuation of ideological and social values. This is perhaps most strongly recognised within the general discipline of sociolinguistics and, more specifically, in the field of critical discourse analysis; as Fairclough (1989) describes:

[1] Corresponding Author: ewan.hannaford@glasgow.ac.uk

ISSN: 2157-4898; eISSN: 2157-4901

> . . . ideological struggle takes place pre-eminently in language. We can think of such struggle as not only *in* language . . . but also *over* language . . . language itself is a *stake* in social struggle as well as a site of social struggle. (p. 88)

What such research centralises is the notion that social dynamics and power relations are integral to language use and attitudes, and that, conversely, language use and language attitudes establish and uphold power relations within society. In other words, dominant languages reflect dominant ideologies, and dominant ideologies are sustained by dominant languages. Such domination results in *linguistic prejudice*, describing both prejudice *towards* language (i.e., derogatory attitudes held towards speakers of specific languages or language varieties that results in social harms) and prejudice *through* language (i.e., language use that establishes, mediates, or perpetuates discrimination towards specific social groups) (Bourhis & Maass, 2005).

From an ideological perspective, such prejudice serves to reinforce the status quo, diminishing alternative perspectives by diminishing their language and their status. This is theoretically intertwined with notions of *prestige* and *non-prestige* language varieties—language varieties used by those in power are privileged greater cultural capital and status, often leading to the reification of these varieties (and their ideologies) as 'standard' (Millar, 2012), and the denigration of other language varieties and their speakers, as a result. As Milroy and Milroy (1991) note, in relation to English:

> . . . guardians of the language do not generally recommend the 'superior' systems of non-standard dialects: they confine their claims about superiority to aspects of *standard* English grammar . . .. It can be suggested therefore, that their real concerns are not wholly linguistic but largely social: they are in some way promoting the interest of the variety most widely considered to have prestige. (p. 15)

This fuels linguistic prejudice and its deleterious effects in a cyclical fashion: because a dominant group speaks a particular language variety, this variety is promoted over others as a 'prestiged' standard, leading to the further disparagement of groups already marginalised because of their use of non-standard language varieties. Consequently, contemporary attempts to disrupt dominant discourses and challenge hegemonic historical perspectives promote the mitigation of *linguistic prejudice* and advocate a push towards *language equity*—the recognition that all languages and linguistic varieties are equally valid forms of expression, with the same right to be recognised as such—in an attempt to shift away from the reinforcement of standard varieties and their contained values.

While many language users may be unfamiliar with these academic concepts, they will not be unfamiliar with their practice: all language speakers are, consciously or not, involved in the day-to-day struggle for dominance of languages and language varieties. It is the luxury of speakers of standard varieties that this struggle can often go unnoticed, since there are fewer contexts in which they have to modify their linguistic behaviours, but speakers of non-standard varieties are often forced to 'correct' their variational uses and code-switch to standard varieties to avoid negative social evaluation (Heller, 2007; Hughes et al., 2012). What such practices demonstrate is that language use is invariably bound to social identity (Edward, 2012; Fishman, 1989; Joseph, 2004), with linguistic attitudes used to delineate and reinforce societal distinctions between speakers. Linguistic prejudice is, therefore, often weaponised as a means of perpetuating broader discrimination: language attitudes become vehicles for the persecution of peoples and communities through the persecution of their language varieties, and language ideology is used to facilitate and 'justify' discriminatory behaviours—see, for example, Edward's discussion of the relationship between group identity and language purity/prescriptivism (Edward, 2012).

As a consequence, as well as reinforcing linguistically inaccurate perceptions that there are 'good' and 'bad' forms of language, linguistic prejudice also has significant impacts on the lived experience of individuals within society. Speakers who do not adhere to standardised linguistic forms or socialised language conventions experience repercussions across a wide range of settings, from poorer performance on language tests, limiting educational opportunities (Milroy & Milroy, 1991), to diminished credibility in legal contexts, impacting on the likelihood of being believed as a witness (Rickford & King, 2016), to inhibited access to health services, affecting the likelihood of receiving optimum treatment for illness (Sobo et al., 2005).

Indeed, the effects of linguistic prejudice and discrimination are so ubiquitous that there have been calls to enshrine in human rights law the right of speakers to learn, be taught, and use their native language and their right to learn and be taught all official languages and language varieties that enable full participation in the "cultural, economic, and political processes" of a country (Phillipson et al., 1995, p. 12). What this makes clear is that battles for language dominance are not only of theoretical or academic significance but processes which have direct, tangible impacts upon individual speakers, communities, and identities; initiatives towards language equity, and against linguistic prejudice, are responses that seek to rectify the damaging social aftermaths of these conflicts.

Though the potential detriment of language ideology is sizeable, the scale at which such conceptual grappling may take place linguistically can be large or

small: at a general level, it may occur between competing official languages within a particular society (e.g., between French and English in Quebec, or English and Scots in Scotland); less generally, it can be represented in contestations for social acceptance by different varieties of the same language (e.g., between regional varieties of English and RP in the UK); and, most closely, it appears in clashes between and within the range of discourses and discourse communities to which languages and language varieties play host (e.g., in differing styles and genres of academic English). A rough analogy can be drawn between these linguistic levels and Fairclough's (1989, pp. 28-29) "orders of discourse," with ideology operating at the level of *social orders*, between societies and their values (e.g., languages); *types of practice*, differing discourse practices within the same broader social order (e.g., varieties of the same language); and *actual practices*, different forms of actual discourse within a broader practice type (e.g., discourse communities and their discourse types). Across each of these theatres, it is social institutions that tend to serve as the staging grounds for linguistic conflicts; to draw again on Fairclough (1989, pp. 90-91), "The primary domains in which social struggle takes place are the social institutions . . .," and

> . . . if a discourse type so dominates an institution that dominated types are more or less entirely suppressed or contained, then it . . . will come to be seen as *natural* and legitimate because it is simply the way of conducting oneself.

It is within government, within education, within media, within all organisations providing public services that language ideology is most powerfully promoted, and in which there is greatest scope for the reorientation of linguistic values.

Yet, while there has been substantial research into language policy and ideology within government, education, and media settings (see, for example, Curdt-Christiansen & Weninger, 2015; Fowler, 1989; Johnson & Milani, 2010; Spolsky, 2012), there is less research into institutional collections as sites of linguistic contestation. This is surprising, given the widely acknowledged cultural influence of archives in the canonisation of knowledge and narratives (Assmann, 2010) and their consequent capacity to enable or disable access to, and the perpetuation of, languages and discourses and their contained ideological values. Wodak (1996, p. 9) identifies that:

> Power manifests itself in hierarchies, in access to specific discourses and information and most particularly, in the establishment of symbols. Which myths are considered relevant, or which ideologies, norms, and

values are posited, relates directly to the groups in power and their interests.

It is through this lens that this paper explores how institutional, and more broadly, national collections can contribute to efforts to combat linguistic prejudices, and promote language equity, given their status as both reservoirs of social values and as inherently hierarchical entities organised according to particular canonical ontologies.

Beginning by examining the extent to which language varieties have typically been accommodated within these settings, we then consider how recent technological developments may offer the potential for greater integration of linguistic diversity within established collections, while also interrogating the barriers and limitations of these approaches that may hinder more diversified linguistic landscapes within institutions. Advocating a nuanced approach towards linguistic preservation that takes into account these practicalities, we put forward suggestions for the most effective means of not merely *preserving* language varieties but *valorising* them within institutional frameworks, such that they are provided equal footing with traditional materials in these collections. Consequently, it is the fundamental argument advanced by this paper that, by promoting regional and social varieties through their broader, more effectual integration into institutional collections, these resources can become a significant force in combatting linguistic prejudice by widening the representation, and enhancing the social standing, of historically marginalised or underrepresented languages and communities.

2. Language varieties and institutional collections

Institutional collections may be contained within, and curated by, a range of different organisations and public bodies, including galleries, libraries, archives, and museums (collectively referred to as GLAM institutions) and universities. Institutional resources are made available digitally through digital archives, libraries, and repositories. While the focus of different institutions and collections necessarily varies depending on their remit, these different digital resources typically aim to serve similar functions: preserving and organising institutional materials, promoting their contained research, and widening access to their knowledge (Shreeves & Cragin, 2008; Tedd & Large, 2005).

From a linguistic perspective, these aims are well-aligned with attempts to promote language equality: digital libraries, archives, and repositories can facilitate the preservation of under-served linguistic varieties, and promote their status, by incorporating materials containing these varieties within collections maintained by long-standing, culturally valued institutions, whilst

also enabling wider access to these linguistic varieties beyond their original discourse communities. However, as institutional collections can only ever contain a selection of all possible material, they can also feed into linguistic prejudice by being more likely to incorporate materials that adhere to traditional language norms and more likely to elide materials that are not in standard forms and language varieties.

This process of selection can be detrimental in two ways. Firstly, it promotes standard languages and language varieties as favoured forms, by virtue of their inclusion within the collections of prestige institutions, and, as a result, the standard language and ideologies encountered in these prestige settings is reified in public opinion as the language of cultural influence. Secondly, as non-standard language varieties are *not* encountered by the public in these prestige settings—and, indeed, may not be encountered at all by individuals outside of the original discourse community of a particular language variety—public exposure to these language varieties is limited, and these varieties are positioned outside of the cultural mainstream, decreasing the likelihood of these varieties being accepted by general society.

To draw on the language of another branch of linguistics—corpus linguistics—in the same way that institutional collections are gathered to represent a particular culture, or cultural topic/phenomenon, corpora are collections of texts that are gathered to investigate a particular type of language. Also as with institutional collections, a corpus can only ever contain a sample of all possible materials, due to logistical constraints of time, size, and comprehensibility. To produce corpora that most effectively demonstrate their given language, corpus linguists attempt to construct corpora from texts that maximise *representativeness*, how sufficiently a corpus incorporates the full variety of possible texts, speakers, genres, and other variables within the discourse being investigated, and *balance*, the degree to which materials are evenly distributed between the different text types, speakers, genres, eras, etc. that are relevant to the corpus topic (Biber, 1993; Sinclair, 2005). Though the 'corpora' of institutional collections intend to represent cultural, rather than linguistic, phenomena, without comprehensively representing linguistic diversity, cultural diversity cannot be fully represented either; any institutional collection that excessively focuses on mainstream language varieties and narratives is likely to become culturally *unrepresentative*, in omitting a diversity of relevant materials and contained cultural perspectives from non-standard linguistic varieties, and culturally *unbalanced*, in containing disproportionate amounts of material and cultural values from standard language varieties, at the expense of non-standard language varieties. Besides limiting the utility of such collections to linguists, collections that are unrepresentative and unbalanced in this way misrepresent the fundamental plurality of the culture and discourse around which they are constructed, in a

way that may bolster misunderstandings of linguistic diversity in different cultural contexts.

A growing awareness has emerged of the historical preference of institutional collections towards 'treasured' materials (Prescott & Hughes, 2018), i.e., those that have traditionally been culturally valued, with a concurrent push towards diversifying and decolonising institutional collections, in recognition of the limited and/or biased representations of the past that previous collections can provide (see, for example, Crilly & Everitt, 2021; Thylstrup et al., 2021). Though this paper focuses on the UK context, this has been a global concerted effort across the humanities, with specific research also exploring the role of linguistic diversity in institutional collections and the significance of broader linguistic representation within these settings (Seifart et al., 2018).

For example, Neumann (2019, p. 289) identifies the importance, as well as the challenges, of improving the linguistic representativeness of materials in the collections of the National Library of Australia, in order for it to serve its intended role of enabling ". . . *all* Australians, now *and in the future*, to be able to gain a comprehensive—and culturally and linguistically diverse—picture of Australian society, life and culture." Similarly, though in relation to a different form of institution, Meyers (2021) recognises the saliency of multilingual representation in theological libraries in the US, stating that, "A focus on the linguistic diversity of library collections—one that recognizes the value of collecting resources from all regions of the world—is an essential element in the decolonization of theological education" (pp. 11-12). Meanwhile, initiatives such as the CLARIN Knowledge Sharing Infrastructure seek to bring together resources for the promotion and preservation of linguistic diversity across global institutions, ". . . combining the expertise of language typologists, field linguists, sociolinguists, computer linguists, computer scientists, data curators as well as language archivists from institutions in several geographic locations into a single digital institution" (Hedeland et al., 2018, p. 2341).

Such efforts are generally focused on maintaining linguistic diversity at the level of languages and multilingualism, with attempts to preserve or promote languages in this way motivated by the recognition that each language can offer us something unique in understanding the world. Every language has significant cultural value, in addition to its linguistic functions. In *Language Death*, David Crystal (2000) lucidly summarises the key cultural contributions of language in relation to five areas. As he identifies, we should care about preserving languages because: maintaining linguistic diversity is important in maintaining cultural diversity, and the progressive potential enabled by such diversity of thought; cultural identity is intrinsically connected to language, with our understanding of communities and identities dependent on our understanding of their language; history is preserved in language, and

language is necessary for understanding our histories; each language contributes to the overall sum of human knowledge, and the expression of differing worldviews and forms of human experience; and languages themselves, and their evolution, are important cultural artifacts for examining social structures and human communication (2000). For these reasons, as well as fostering diverse materials in institutional collections, nurturing diverse languages within these settings is equally significant, in order to avoid losing the important cultural perspectives, ideologies, identities, and social values contained within them.

However, while languages can undoubtedly offer unique insights into cultural understanding, language varieties and dialects may equally do so: "dialects are just as complex as languages in their sounds, grammar, vocabulary, and other features" and therefore "dialect death *is* language death, albeit on a more localized scale" (Crystal, 2000, p. 38). Likewise, as mentioned earlier, contestation between languages is not the only stratum upon which ideological conflict and linguistic prejudice takes place. Just as languages are receptacles for unique worldviews and cultural and ideological perspectives, so too are language varieties, revealing and containing the social and cultural values of the discourse communities that produce them (cf. Hymes, 1972); as such, they are equally at risk of facing linguistic prejudice. Therefore, as well as recognising the need for greater equity between *languages* within institutional collections, it is also important that there is a focus on improving the broader representativeness of institutional collections in relation to varieties of the *same* language and to greater equity between linguistic varieties (and the communities that produce them), particularly those which have traditionally been overlooked by institutions. This includes appropriate representation of regional varieties, the language spoken by speakers from a particular geographical area (Hasan, 2004), e.g., Glaswegian, spoken by communities in and around the city of Glasgow, and social varieties, the language of a particular social group or class (Hasan, 2004), e.g., people with a shared vocation. To return to notions of language equity and linguistic prejudice, by bringing notions of linguistic diversification within institutions down to the level of linguistic varieties, it may be possible to challenge linguistic prejudice and linguistic stereotypes, and confront dominant ideologies, in a broader sense, through increasing the prestige of varieties and reducing derogatory attitudes towards speakers from particular regions or social groups. Equally, such efforts may serve to preserve language varieties, and their contained cultural knowledge, as they can languages; as Crystal notes, where languages are endangered, one of the primary methods of preserving their use is through increasing their prestige within dominant discourse communities and their visibility within the public domain (2000).

3. Documenting and democratising linguistic variety

While it may be clear that increasing the diversity of linguistic varieties in institutional collections would be beneficial to attempts to combat linguistic prejudice against these varieties, the answer to how to go about this process is less obvious. If the language of underrepresented social groups and regions is not already preserved and recorded in the existing materials of large institutional collections and repositories, where and how can they be found? Attempts to gather such material by external bodies—e.g., researchers or governmental institutions—face significant challenges, both in finding time and resources to conduct this collection and in producing collections that are accurately representative of the communities and linguistic varieties in question. Dorian (2014, p. 24) expounds several such issues in relation to the preservation of endangered languages, concluding that:

> Because what gets recorded is affected by so very many factors—e.g., how well source(s) and researchers know one another or like one another, how many people are present and listening, how much factionalism or leadership competition exist within the population the researcher would like to record, what the recording medium is, how forthcoming the sources are (unknowns such as the behaviour or even aptitude of previous researchers can be major influences; see Grinevald 2005), the race, gender, and/or age of the researchers and the sources, even the season of the year and the weather in some settings—it seems inevitable that any documentary record we produce will be skewed in ways and directions that we recognize poorly or fail to recognize at all.

In these ways, and others besides, though attempts to diversify linguistic varieties present in institutional collections made by these institutions themselves are welcome, they are likely to face many obstacles.

An alternative approach to this issue lies in community archives: archives that centre around particular geographical or social communities and are run by members of that community, rather than by mainstream heritage institutions or governmental organisations (Gilliland & Flinn, 2013). In the UK, a 2007 report estimated that there may be over 3000 of such organisations (Flinn, 2007), spread across a wide diversity of different regions and topics. Central to the concept of such archives is their status as community-managed entities, being constructed by communities for the preservation of their own cultural heritage. As such, they can constitute a more honest reflection of the cultural items and forms which are significant to particular regional or social groups, free from the influence of external organisations or pressures of broader societal norms. Linguistically, this means that many materials within the collections of community archives—e.g., local records, stories, narratives,

poetry, oral recordings, and audiovisual pieces—may be collected in the social and regional linguistic varieties that are typically used by these communities, rather than being stored in, or translated into, a standard linguistic variety for dissemination outside of their original community. As a result, the materials within community archives can be an invaluable resource for providing access to language varieties that might otherwise be inaccessible and democratising the representation of language and culture beyond traditional cultural institutions.

From this perspective, facilitating wider access to community archives, and preserving their materials, represents a powerful opportunity to increase public understanding of the language varieties of these communities and their cultures, diminishing prejudicial attitudes; in their ideal utilisation, community materials may serve as a means of disrupting or inverting linguistic hierarchies, by showcasing the value of alternative language varieties over standard language in particular discourse contexts. The wider presentation of these language varieties and their contained narratives, histories, and social values, could also enable them to push back against established ideological perspectives represented in the standard language materials of particular collections, contesting canonical or 'master' narratives through the presentation of alternative narratives (Black, 2014; Hyvärinen, 2020) that are produced by communities in their own language varieties. Additionally, the incorporation, integration, or linking of such materials into institutional collections presents a chance to improve equality of access to these varieties and increase the standing of these varieties, and their values, in public conceptions, by drawing regional and social language varieties of previously marginalised groups onto the level of prestige materials held by such institutions.

Recognising the significance of community archives to goals of broader representation and diversity, linguistic and otherwise, there has been increasing research into ways of incorporating or representing community archives in institutional settings (Bastian & Flinn, 2018). This has been particularly pronounced with the growing emergence of, and growing public familiarity with, digital resources and tools that facilitate more effective and efficient digital archiving by community groups.

In the UK, ongoing work on the *Our Heritage, Our Stories* project (UKRI, 2023) is attempting to widen access to the digital materials of UK community archives through a new public-facing resource, hosted by The National Archives (TNA) (the UK's official government archives) and made compatible with existing materials within their collections for searching, linking, and comparing. To do so, this project is leveraging expertise from across the humanities and combining this with computer science expertise, adopting a combination of Natural Language Processing techniques—including Named Entity Recognition,

Entity Linking, and Relation Extraction—to allow for digital materials in community archives to be linked and integrated with existing archival materials in TNA in an automated fashion. Automated approaches such as these overcome challenges of scale and have been successfully applied in other subject areas, including in relation to legal (Sleimi et al., 2021) and historical (Humbel et al., 2021) texts, but not, until now, to digital archives. The promise of such approaches is the widened accessibility and promotion of community archives at the national level, facilitating easier public access to a wider range of linguistic varieties and their contained cultural knowledge and values. This institutional recognition is, in turn, anticipated to lead to shifts in public conceptions of these linguistic varieties and reduced linguistic prejudice towards speakers, in the ways outlined above.

However, while there is much potential in community archives for improving access to, and representation of, linguistic varieties, and while automated approaches can facilitate this progress, there remain many challenges and questions in the practice of integrating linguistic varieties into institutional collections without flattening, misrepresenting, or misinterpreting these materials. There is also a real danger that any overly simplistic incorporation of linguistic varieties in institutional collections may not merely hinder movements towards language equity but become actively counterproductive to this aim. Approaches which reductively 'accept' linguistic varieties into existing, standardised frameworks—instead of attempting to address alternative language varieties as equally valid forms of expression and rectify historical power imbalances—are likely to lead to a reification of the minority status of these varieties and result in their struggle for proper recognition being subsumed into an inadequate relationship of qualified tolerance: "Where dominated discourses are oppositional, there will be pressure for them to be suppressed or eliminated; whereas containment credits them with a certain legitimacy and protection—with strings attached!" (Fairclough, 1989, p. 91).

Given the historic marginalisation of minority communities by institutions, there is also likely be a degree of hesitancy or reluctance towards attempts to now accommodate their materials by external parties, "… especially if they are members of the society that threatened the community in the first place" (Crystal, 2000, p. 148). It is therefore imperative that attempts to preserve and promote linguistic varieties in institutional settings, including through use of community archives, take a conscientious approach to this action, moving beyond language *preservation* to language *valorisation*.

4. Valorising linguistic varieties

The scope and influence of institutional collections means that these settings can do far more than merely preserve language; indeed, we would be doing a disservice to efforts to reduce linguistic prejudice and promote language equity

if we limited our conception of the role of institutional collections to the fields of documentation and preservation. To quote Crystal again, "no language has ever been saved just by being documented' (2000, p. 149); if all we are doing is collecting languages to be recorded in galleries, libraries, museums, and archives, what are these languages being preserved *for*? Equally, while contemporary digitisation efforts in the humanities are useful for the preservation of linguistic varieties, these can only ever be part of a solution (and do not address linguistic inequalities or prejudicial attitudes).

A push towards improving the social standing of linguistic varieties, reducing linguistic prejudice, and disseminating the value of linguistic diversity is necessary to establishing a societal appreciation of the cultural significance of language varieties and their potential for reorienting ideological values, and it is here that institutions can exert their most powerful influence. It is worth highlighting here that the playing field is not level: promotion of standardised varieties has been the historic norm, and the diversification of linguistic varieties in institutional collections is a movement against decades of travel in the other direction. Consequently, attempts to valorise, and not just preserve, linguistic varieties must recognise and attempt to rectify historical power imbalances as far as possible, rather than simply being a matter of accommodation. To further this conversation, drawing on prior approaches and ongoing experiences on the *Our Heritage, Our Stories* project, the following non-sequential principles are suggested as key practical considerations in the diversification of linguistic varieties in institutional settings.

4.1. Representativeness and balance

Efforts to incorporate linguistic varieties into institutional collections should take into account what would constitute the most representative selection of materials for a particular cultural topic. If the focus of a collection is broad, such as a general national archive, the range of language and language varieties should be similarly broad, attempting to represent as diverse a selection of linguistic varieties as are present in the discourse population. However, if a collection's focus is specific, such as on a particular cultural movement or population, the linguistic diversities represented in its materials should be equally specific to those used by members of that specific discourse community. In the same vein, if the discourse surrounding a particular topic is predominantly conducted in a particular linguistic variety, the collection should be balanced to reflect this as far as possible, while maintaining a representative selection of materials. Community archives can offer a key resource in providing this balance to institutional collections, facilitating the incorporation of diverse linguistic varieties that are otherwise unattainable or restrictedly available. The key rationale of this principle, in relation to linguistic prejudice and language equity, is that by enabling the public to

engage with institutional collections that are accurately representative and balanced linguistically, the public will be able to appreciate the significance of linguistic varieties culturally, with institutional prestige serving to reorient what linguistic varieties are viewed as valued.

4.2. Considered language levels

Tied to the principle of representativeness and balance is consideration of the layers at which language ideology operates, and their interaction with institutions. As discussed, linguistic prejudice and language ideology can operate at the level of language, language variety, and discourse, respectively mapping, approximately, onto the concepts of social order, types of practice, and actual practices (Fairclough, 1989). In order to comprehensively consider the linguistic diversity of institutional collections, the representativeness and balance of collections must be evaluated across each of these linguistic and/or discursive levels, moving beyond considering just the potential diversity of languages in a collection to the diversification of language varieties and discourse types appropriate to a particular cultural collection. Incorporating linguistic diversity in institutional collections without considering language levels may lead to the reproduction of linguistic inequalities at lower orders of discourse: for example, incorporating Scots materials that only represents one variety of Scots, has the potential to diversify a collection on the level of languages but homogenise perceptions of the language at the level of language varieties. In a community archives context, this may involve ensuring the inclusion of regionally or socially diverse archives focused on the same topic, to adequately represent the cultures and localities involved. The intention of this principle is that representation should reflect linguistic diversity at each level of ideological conflict, moving beyond just thinking about linguistic prejudice and language equality at the level of languages, as these often operate at more specific language levels.

4.3. Maintained complexity

An inherent tension exists between diversifying the linguistic varieties contained in institutional collections and the standardised frameworks in which these collections are held; likewise, there are significant challenges in attempting to situate non-standardised, diverse materials within community archives (which may contain alternative language varieties) into the standardised models of institutions. However, any attempts to overcome these difficulties must, to the greatest degree feasible, avoid temptations to simplify or 'translate' materials to fit into pre-existing models, as such approaches inevitably result in a reduction in the value of these varieties; as Mithun (1998, p. 189) identifies:

> The loss of languages is tragic precisely because they are not interchangeable, precisely because they represent the distillation of the thoughts and communication of a people over their entire history. Language instruction and documentation that is limited to translations of English words or even English sentences misses the point entirely.

Instead, a conscious effort must be made to maintain complexity wherever possible, to avoid assimilating materials in ways that counterproductively flatten their diversity, and to prevent suggestions that regional and social language varieties are subordinate to those used in standardised structures. While concessions may inevitably need to be made in certain circumstances, such compromises should be resisted as far as possible, in order to maintain as full a picture of represented varieties as possible and to reassert their linguistic parity. Approaches which link materials into institutional settings rather than subsuming them should be favoured, which may involve incorporation of language-specific resources that allow for linguistic diversity to be maintained whilst also supporting compatibility with institutional standards. For example, *Our Heritage, Our Stories* is exploring use of the digitised *Scottish National Dictionary* (DSL Online, 2022) as a means of improving computational interpretation of Scots materials, in order for these to be linked with materials in other language varieties for searching and comparison. Where such resources do not exist for a language variety, the expertise of the community will be crucial to properly representing varieties without deprecating their complexity, as discussed below.

4.4. Community inclusivity

Any attempts to incorporate linguistic varieties into institutional settings should involve the inclusion of their communities of speakers in this process. While there are certainly challenges involved in defining who qualifies as a speaker for a given language or variety (Dorian, 2014), any diversification of linguistic varieties in institutions without consultation with its community of speakers is likely to lead to reductions and simplifications that reduce the value of this incorporation; full understanding, and, consequently, proper representation, of the cultural significance of the forms and features of languages or varieties is impossible to attain without such a dialogue with their discourse community. Furthermore, without inclusion of the community in such efforts, attempts to promote linguistic diversity are doomed to fail; "Languages need community in order to live" (Crystal, 2000, p. 154) and, consequently, 'Institutions cannot replace individuals" (Crystal, 2000, p. 118) if languages and language varieties are to be reconceptualised, revitalised, and valorised, and not merely preserved.

It is here that the potential of community archives becomes particularly apparent, allowing for communities of speakers to be brought into the institutional fold alongside their materials. As well as ensuring a community of invested speakers are involved in the promotion of linguistic varieties, this is also significant from the perspective of linguistic prejudice in beginning to redress historical inequalities and possible marginalisation of these communities by institutions, serving to reorient the ideological centres of these settings. Furthermore, as experienced in work on the *Our Heritage, Our Stories* project, the expertise of communities is a vital source of further context and cultural experience that can serve to counterbalance shortcomings of purely automated approaches at interpreting non-standard materials, which can fail to properly recognise the linguistic varieties present and their cultural significance.

4.5. Accessibility and visibility

While community input is crucial to the integration of diverse materials into institutions, once these materials are connected to institutional collections, it is equally fundamental that these materials are made accessible and visible outside of their original communities. If linguistically diverse materials can only be searched for and identified by speakers of these varieties, then, in practice, their representation in institutions may be of little more use than if they had not been included at all; such a relationship would constitute an unfortunate manifestation of Fairclough's (1989, p. 91) concerns regarding alternative discourses becoming "suppressed" via "containment" by the dominant discourse. Speaking in relation to the diversification of the National Library of Australia, Neumann (2019) identifies similar concerns, stating, ". . . the Library's 'collecting and preserving' of material will in itself not enable users to understand Australians' diverse histories. The collected and preserved material needs to be discoverable and accessible" (p. 291).

Valorising linguistic varieties requires the capacity for these resources to be encountered by all individuals exploring an institution's collections, and for these materials to be made equally as visible as materials in standard varieties. Practically, this requires material to be linked across linguistic varieties, with searches made in one language or language variety identifying relevant materials from other language varieties, via synonymous terms, semantic categories, or sociolinguistic features. To avoid reductive approaches that translate linguistic varieties into standard usage, this means that' "Discoverability depends on the quality of the metadata attached . . . ." (Neumann, 2019, p. 291).

It is in this process that the expertise of community speakers, and the principle of *community inclusivity*, is again salient (and where the incorporation of

language-specific resources into automated methods, discussed earlier, may also prove fruitful), enabling institutions to categorise and link materials across linguistic varieties. Linguistically diverse materials, from community archives or otherwise, should also be located at the same level as standard materials and not relegated to sub-sections of collections that might suggest inferior status, i.e., searches for material should reveal all relevant materials in results, regardless of linguistic variety; by promoting the visibility of materials in this way, and enabling their exploration, the legitimacy of alternative varieties can be communicated through their co-location with historically prestige materials within collections.

6. Discussion and conclusion

Though this paper has focused on the UK context, the push for improving linguistic equity in institutional collections is of global concern, as examples from US and Australian institutional practices show. The principles described here for more efficacious inclusion of linguistic varieties in institutional collection are designed to be broad, so as to be applicable across cultural contexts, though of course individual collections, institutions, and countries will each have their own idiosyncrasies and cultural histories to take into consideration in this process.[1] Incorporating these principles, such work should improve the effectiveness of linguistic diversification in institutional collections and enable them to become key tools in combatting linguistic prejudice and promoting language equity. Likewise, community archives are equally a global phenomenon, and their potential contribution to the process of diversifying collections should be equally valued in all institutional contexts.

Institutional collections can serve as powerful tools to reduce linguistic prejudice and promote language equity but are currently an underappreciated avenue for such efforts. While existing collections are typically focused on standard languages and language varieties, increased diversification of materials in these venues can be leveraged to improve social perceptions of historically marginalised languages, peoples, and cultures, and reduce discriminatory attitudes by promoting these diverse perspectives in prestige settings. Community archives should be recognised as an important prospective source of such materials and their contained varieties, as well as a crucial source of expertise for their interpretation and appropriate representation. However, the inclusion of materials representing non-standard linguistic varieties, including community materials, cannot be approached inconsiderately, and care must be taken to avoid any simplistic accommodation of these materials that maintains their subordinate status and equates to a caveated diversification of institutional collections that preserves the status quo.

As such, efforts should move beyond attempts at *preservation* to the *valorisation* of linguistic varieties, and their containing materials, in institutional collections. These attempts should be guided by the key principles suggested in this paper (drawn from prior and ongoing efforts in this area), striving to achieve *representativeness and balance*, *considered language levels*, *maintained complexity*, *community inclusivity*, and *accessibility and visibility*. Adopting these priorities, institutional collections can become significant positive influences on public attitudes towards languages and language varieties and disrupt dominant discourses regarding prestige standard varieties, promoting greater language equity, diversifying cultural and ideological perspectives, and improving societal treatment for speakers of alternative language varieties.

Notes:

1. Though outside the focus of this paper, the linguistic landscapes of the Global South may hold particular nuance in this regard. For instance, conceptions of *representativeness and balance* are likely to be especially acutely impacted by colonial legacies in some regions of the Global South and their institutions, as recognised by contemporary linguistic decolonisation efforts (Agyekum, 2018). Similarly, consideration of *language levels* should recognise that, in some contexts, competition between regional languages and varieties may be equally, or more, central to concerns around linguistic prejudice and language ideology than competition between these so-called 'minority' languages and globally dominant languages (e.g., English)—Simpson's (2008) edited volume *Language and National Identity in Africa* provides numerous such examples.

Authors' Statement

Both authors have contributed to all sections of this paper. Ewan Hannaford conceptualized and prepared the article draft, while Marc Alexander provided advice and revisions on each section. For the purpose of open access, the authors have applied a Creative Commons Attribution (CC-BY) licence to any Author Accepted Manuscript version arising from this submission

Acknowledgments

The authors are grateful for the support of the wider project team on *Our Heritage, Our Stories*, funded by the UK Arts and Humanities Research Council (project no. AH/W00321X/1).

The Authors

Ewan D. Hannaford (Email: ewan.hannaford@glasgow.ac.uk) is a post-doctoral Research Assistant at the University of Glasgow, currently working on the *Our Heritage*, *Our Stories* project. His research examines how language may influence attitudes and behaviours, incorporating investigations of language change, stigma, media and health discourses, and large-scale linguistic analysis.

Marc Alexander (Email: marc.alexander@glasgow.ac.uk) is Professor of English Linguistics at the University of Glasgow, Director of the Historical Thesaurus of English, and Deputy Principal Investigator for the *Our Heritage, Our Stories* project. His research focuses on the study of meaning and effect in English. He has published on historical lexicology, digital humanities, political discourse, medical discourse, metaphor, astronomical names, the linguistics of colour, the history of Parliament, cognitive linguistics, and detective fiction. He is a Fellow of the Royal Society of Arts and the Royal Historical Society and Chair of the Board of Directors of Dictionaries of the Scots Language.

References

Agyekum, K. (2018). Linguistic imperialism and language decolonisation in Africa through documentation and preservation. In J. Kandybowicz, T. Major, H. Torrence & P. T. Duncan (Eds.), *African linguistics on the prairie: Selected papers from the 45th Annual Conference on African Linguistics* (pp. 87-104). Language Science Press. https://dx.doi.org/10.5281/zenodo.1251718

Assmann, A. (2010). Canon and archive. In A. Erll & A. Nünning (Eds.), *Media and cultural memory: An international and interdisciplinary handbook* (pp. 97-107). De Gruyter. https://doi.org/10.1515/9783110207262

Bastian, J., & Flinn, A. (2018). *Community archives, community spaces: Heritage, memory and identity*. Facet Publishing. https://doi.org/10.29085/9781783303526

Biber, D. (1993). Representativeness in corpus design. *Literary and linguistic computing, 8*(4), 243-257.

Black, J. (2014). *Contesting history: Narratives of public history*. Bloomsbury. http://doi.org/10.5040/9781350249714

Bourhis, R., & Maass, A. (2005). Linguistic prejudice and stereotypes. In U. Ammon, N. Dittmar, K. J. Mattheier & P. Trudgill (Eds.), *Sociolinguistics: An international handbook of the science of language and society* (pp. 1587-1601). De Gruyter.

Crilly, J., & Everitt, R. (Eds.). (2021). *Narrative expansions: Interpreting decolonisation in academic libraries*. Routledge. https://doi.org/10.29085/9781783304998

Crystal, D. (2000). *Language death*. Cambridge University Press.

Curdt-Christiansen, X., & Weninger, C. (Eds.). (2015). *Language, ideology and education: The politics of textbooks in language education*. Routledge. https://doi.org/10.4324/9781315814223

Dorian, N. (2014). Introduction. In C. Dorian (Ed.), *Small-language fates and prospects: Lessons of persistence and change from endangered languages—collected essays*. (pp. 1-29). Brill. https://doi.org/10.1163/9789004261938_002

DSL Online. (2022). *Dictionaries of the Scots language online* (Version 3.0) [Computer software]. https://dsl.ac.uk

Edward, J. (2012). *Multilingualism: Understanding linguistic diversity*. Bloomsbury.

Fairclough, N. (1989). *Language and power*. Longman.

Fishman, J. (1989). *Language and ethnicity in minority sociolinguistic perspective*. Clevedon.

Flinn, A. (2007). Community histories, community archives: Some opportunities and challenges. *Journal of the Society of Archivists, 28*(2), https://doi.org/10.1080/00379810701611936

Fowler, R. (1989). *Language in the news: Discourse and ideology in the press*. Routledge. http://doi.org/10.4324/9781315002057

Gilliland, A., & Flinn, A. (2013). Community archives: What are we really talking about? *CIRN Prato Community Informatics Conference 2013.* [Keynote]

Hasan, R. (2004). Code, register and social dialect. In B. Bernstein (Ed.), *Class, codes and control* (pp. 224-254). Routledge.

Hedeland, H., Lehmberg, T., Rau, F., Salffner, S., Seyfeddinipur, M., & Witt, A. (2018). Introducing the CLARIN Knowledge Centre for Linguistic Diversity and Language Documentation. In *Proceedings of the eleventh international conference on language resources and evaluation.* LREC 2018.

Heller, M. (2007). Code-switching and the politics of language. In W. Li (Ed.), *The bilingualism reader* (2nd ed.). Routledge. https://doi.org/10.4324/9781003060406

Hughes, A., Trudgill, P., & Watt, D. (2012). *English accents & dialects: An introduction to social and regional varieties of English in the British Isles* (5th ed.). Routledge. https://doi.org/10.4324/9780203784440

Humbel, M., Nyhan, J., Vlachidis, A., Sloan, K., & Ortolja-Baird, A. (2021). Named-entity recognition for early modern textual documents: A review of capabilities and challenges with strategies for the future. *Journal of Documentation, 77*(6). https://doi.org/10.1108/jd-02-2021-0032

Hymes, D. (1972). Models of the interaction of language and social life. In J. J. Gumperz & D. Hymes (Eds.), *Directions in sociolinguistics: The ethnography of communication* (pp. 35-71). Holt, Rinehart & Winston.

Hyvärinen, M. (2020). Toward a theory of counter-narratives: Narrative contestation, cultural canonicity, and tellability. In K. Lueg & M. Lundholt (Eds.), *Routledge handbook of counter-narratives* (pp. 17-29). Routledge. https://doi.org/10.4324/9780429279713-3

Johnson, S., & Milani, T. (Eds.). (2010). *Language ideologies and media discourse: Texts, practices, politics*. Bloomsbury.

Joseph, J. (2004). *Language and identity: National, ethnic, religious*. Palgrave Macmillan.

Meyers, J. (2021). The importance of linguistically diverse collections: Decolonizing the theological library. *Theological Librarianship, 14*(2). https://doi.org/10.31046/tl.v14i2.2889

Millar, R. (2012). Social history and the sociology of language. In H. Hernández-Campoy & J. Conde-Silvestre (Eds.), *The handbook of historical sociolinguistics* (pp. 41-59). Wiley-Blackwell. https://doi.org/10.1002/9781118257227

Milroy, J., & Milroy, F. (1991). *Authority in language: Investigating language prescription & standardisation* (2nd ed.). Routledge.

Mithun, M. (1998). The significance of diversity in language endangerment and preservation. In L. Grenoble & L. Whaley (Eds.), *Endangered languages: Language loss and community response* (pp. 163-191). http://dx.doi.org/10.1017/CBO9781139166959.008

Nuemann, K. (2019). In search of "Australia and the Australian people": The National Library of Australia and the representations of cultural and linguistic diversity. In K. Darian-Smith & P. Hamilton (Eds.), *Remembering migration* (pp. 285-299). Palgrave Macmillan. https://doi.org/10.1007/978-3-030-17751-5_19

Philipson, R., Rannut, M., & Skutnabb-Kangas, T. (1995). Introduction. In R. Philipson, M. Rannut & T. Skutnabb-Kangas (Eds.), *Linguistic human rights: Overcoming linguistic discrimination* (pp. 347-370). Mouton de Gruyter.

Prescott, A., & Hughes, L. M. (2018). *Why do we digitize? The case for slow digitization*. Archive Journal. https://www.archivejournal.net/essays/why-do-we-digitize-the-case-for-slow-digitization

Rickford, J., & King, S. (2016). Language and linguistics on trial: Hearing Rachel Jeantel (and other vernacular speakers) in the courtroom and beyond. *Language, 92*(4), 948-988. https://doi.org/10.1353/lan.2016.0078

Seifart, F., Evans, N., Hammarström, H., & Levinson, S. (2018). Language documentation twenty-five years on. *Language*, *94*(4), e324-e345. https://doi.org/10.1353/lan.2018.0070

Shreeves, S., & Cragin, M. (2008). Institutional repositories: Current state and future. *Library Trends, 57*(2), 89-97.

Simpson, A. (2008). *Language and national identity in Africa*. Oxford University Press.

Sinclair, J. (2005). Corpus and text: Basic principles. In M. Wynne (Ed.), *Developing linguistic corpora: A guide to good practice* (pp. 17-30). Oxbow books.

Sleimi, A., Sannier, N., Sabetzadeh, M., Briand, L., Ceci, M., & Dann, J. (2021). An automated framework for the extraction of semantic legal metadata from legal texts. *Empirical Software Engineering, 26*(3). https://doi.org/10.1007/s10664-020-09933-5

Sobo, E., Seid, M., & Gelhard, L. (2005). Parent-identified barriers to pediatric health care: A process-oriented model. *Health Services Research, 41*(1), 148-172. https://doi.org/10.1111/j.1475-6773.2005.00455.x

Spolsky, B. (Ed.). (2012). *The Cambridge handbook of language policy.* Cambridge University Press. https://doi.org/10.1017/cbo9780511979026

Tedd, L., & Large, A. (2005). *Digital libraries: Principles and practice in a global environment*. K. G. Saur.

Thylstrup, N. B., Agostinho, D. Ring, A. D'Ignazio, C., & Veel, K. (2021). Big data as uncertain archives. In N. B. Thylstrup, D. Agostinho, A. Ring, C. D'Ignazio & K. Veel (Eds.), *Uncertain archives: Critical keywords for big data* (pp. 1-27). MIT Press. https://doi.org/10.7551/mitpress/12236.001.0001

UKRI. (2023). *Our Heritage, Our Stories: Linking and searching community-generated digital content to develop the people's national collection.* UK Research and Innovation. https://gtr.ukri.org/projects?ref=AH%2FW00321X%2F1

Wodak, R. (1996). *Disorders of discourse*. Longman.