Digital Scholarship in the Humanities, 2022, **00**, 1–14
https://doi.org/10.1093/llc/fqac072
Full Paper

Whose Language? Whose DH? Towards a taxonomy of definitional elusiveness in the digital humanities

Josh Brown [1]*

[1]School of Humanities Italian Studies, The University of Western Australia, Crawley, Australia
*Correspondence: Josh Brown, The University of Western Australia, 35 Stirling Hwy, Crawley WA 6009, Australia.
E-mail: joshbrownwa@gmail.com

Abstract

This article responds to the current interventions regarding spatio- and linguistic diversity in the digital humanities (DHs). Previous work has focused on the practitioners of DHs themselves, the diversity of projects, the geographical diversity of peoples and places which such projects represent, and others. Some literature has considered multilingual DH, whether a non-Anglophone DH is possible, or a DH 'accent'. This article pushes these boundaries further by considering forms of historical linguistic hybridity for languages, language varieties, and groups of people that are no longer extant. It considers one text in particular, the *Dictionnaire de la langue franque*, to show that, although 'mixed' languages are the norm in all societies, forms of hybridity are often left by the wayside in favour of increasing heterogeneity. This observation, in turn, leads to a taxonomy of definitional elusiveness.

1 Introduction[1]

This article responds to the current debate regarding spatio- and linguistic diversity in the digital humanities (DHs). Earlier attempts have gone some way in attempting to 'broaden out' the field, either by insisting on increased diversity in terms of the geographical location of DH projects, the projects themselves, or the people and places that such projects represent. Recently, terms such as 'multilingual DHs', or variations of this syntagm, have begun to appear in various fora including in scholarly literature (Horvath, 2021; Nilsson-Fernàndez and Dombrowski, forthcoming), as well as on various blogs, conferences, and professional organizations. Some scholars, such as Dombrowski (2022), have argued that 'Non-English DH is not a thing'. Usefully, she complicates the question of non-English DH, showing how power and scholarly communications are more problematic than is usually assumed across a variety of phenomena, including time zones, language of publication, journals in particular languages, etc. In a wide-ranging address, Dombrowski (2022) usefully frames multilingual DH in terms of broader questions relating to the field, articulating questions such as: 'who holds the keys to power? What languages are call for proposals in?' She notes that many of these issues are 'recursive problems'.

An international network for scholars using DH tools and methods on languages other than English has also been formed.[2] Other work has argued for greater language sensitivity and diversity in DH (Spence and Brandao, 2021), focused on the challenges of not using English as the dominant language in DH projects (Priani Saisó, 2020), or using non-Latin scripts (Wagner, 2020).[3] There also exists the much older work of a standing committee on 'Multi-lingualism and Multi-Culturalism' as part of the Alliance of Digital Humanities Organizations (ADHO), which is 'charged with developing and promoting policies in ADHO and its constituent organizations that will help them to become more linguistically and culturally inclusive in general terms, and especially in the areas where linguistic and cultural matters play a role'.[4] Another initiative includes Global Outlook::DHs (GO::DH), as a Special Interest Group of ADHO 'to help break down barriers that hinder communication and collaboration'.[5] And a recently created 'Section' of Modern Languages Open, entitled Digital Modern Languages, aims to 'bring together and expand research that engages with digital culture, media and technologies in relation to languages other than English' (Spence and Wells, 2021, p. 1). This article takes these forms of representation further, by arguing for a taxonomy of definitional elusiveness in DH. Efforts to increase linguistic diversity in DH have so far been made on various fronts, from scholars mainly located in the USA and Europe. Advocacy for increased linguistic diversity often comes in the form of arguing for specific languages, that is to say, standard languages (French, English, Italian, etc.). Less emphasis has been placed on forms of language that do not fit neatly into any one language category. This is particularly the case for contact languages.[6] One such instance is the so-called Mediterranean Lingua Franca (MLF), an extinct trade language, with a Romance lexical base used around the Mediterranean basin and discussed later in this article.

The article briefly canvasses some recent moves in DH to look at spatio- and linguistic diversity in DH, before considering just one specific cultural object from the past—a dictionary written by an anonymous author in 1830. This dictionary, entitled *Dictionnaire de la langue franque ou petit mauresque*, is an excellent example of a hybrid linguistic space which can be taken as a case-study, I argue, for historical forms of language and texts that do not fit into a particular taxonomy, or for whom no one in the contemporary world is advocating.[7] Towards the end of the article, I turn to questions of digital hybridity as a form of definitional elusiveness. A brief conclusion is provided.

So far, few projects in DH have attempted to apply digital methodologies to multilingual environments. One exception is Oberholzer and Kunzmann (2017), whose research project VerbaAlpina investigates the specific Alpine lexis in the multilingual Alpine region where different dialects and languages from the Germanic, Romance, and Slavic language families are spoken. In particular, they focus on the methodology used to transfer analogue data from traditional linguistic atlases and dictionaries into a digital format. This approach allows for representation and comparison of traditional geolinguistic data within a digital and multilingual research environment open to the public realm. Other projects have similarly focused on websites for the study of dialectal lexicon, such as Galician and Portuguese heritage lexicon (Negro Romero and Palmou, 2020). Both these projects focus on extant users of language. While Oberholzer and Kunzmann (2017) also make use of linguistic data from historical dictionaries, one also finds a corresponding modern community who continues to speak these languages, whether in the past or surviving. This is not the case with mixed languages which are no longer spoken, such as that contained in the *Dictionnaire*. Projects such as these are fundamental for representation of peoples and the languages they use. But the further application of DH in the area of multilingual contact has broader significance too, since it allows us to model the emergence of these very contact varieties (e.g. Tria *et al.*, 2015). It also provides a method for digital representation of hybrid linguistic data—and makes that data more available to researchers. There is still much work to be done in 'exposing' invisible collections of the many languages in the world. Indeed, Thieberger (2017) points to 7,000 such languages.

Similarly to the present paper, earlier studies have also argued for greater diversity in DH, and particularly with regard to modern data. Little work has been done on forms of historical representation. The result is that, in the case of certain paradigms, 'there may be other worlds, other DHs that we do not recognize' (Risam, 2017, p. 381), that generate a need to 'reshape power dynamics creating centers and peripheries' (p. 379). One repository is the *African Online Digital Library*. This resource houses multiple projects. It is facilitated by Michigan State University's DH centre. The institutional framework in Anglophone academia is also a focus in a paper by Pitman and Taylor. They consider how institutional frameworks of modern languages 'can contribute a much more diverse range of projects and materials conducted in languages other than English' (2017, para 2). In the case of the *Dictionnaire*, there is no extant group, university, library, or advocate, that is representing the historical people(s) who spoke and used the forms of language that the dictionary is purporting to describe—they have 'fallen through the cracks' of history. But this is precisely where DH tools in modern languages departments can benefit both the discipline and the object of study itself. In other words, modern languages' 'findings as regards digital content creation in various locales and communities around the globe can provide insights that would enrich DH, and contribute to its ongoing shaping of itself as a discipline' (Pitman and Taylor, 2017, para 24).

The lack of interest in 'mixed' forms of language and cultural hybridity may be due to historical and cultural reasons, at least in part. More and more, scholars are paying attention to 'issues related to ethnicity, gender, race, language, and class' (Galina Russell, 2014, p. 308). While questions of representation largely revolve in and around a US context, scholars have noted that the earlier debate 'was exclusive of other national contexts' (p. 308). Galina Russell calls for DH work to incorporate scholars 'from a broader range of countries and linguistic backgrounds than are currently represented' (p. 308). Nation-states are a decidedly contemporary construction, as are the languages that 'map' on to these states. It is perhaps unsurprising, then, that forms of hybridity (cultural, linguistic, whatever) are often less represented than other 'standard' languages, which more clearly embody the nation-states in which they are spoken.[8]

A rejection of these boundaries finds a parallel in literary studies. Recent work has mooted the possibility of discounting 'the literary-historical way of modeling culture altogether, in favor of a model that disregards periodization, or that is uninterested in historical change, or that denies the existence of "literature" as a category' (Flanders and Jannidis, 2019, p. 3). While that work is not directly concerned with language, the principle of rejecting particular taxonomies for a broader confluence of methods is one for which DH is well-primed to provide an answer. Indeed, the plurality of DH is one of its greatest strengths in developing new modes of scholarship and communication (see also Fitzpatrick, 2012). For example, Horvath (2021) introduces some topical collaborative projects that have sought to strengthen language diversity in DH.[9] In particular, she describes the preliminary insights from the *Disrupting Digital Knowledge Infrastructures* collective,[10] lessons from her pilot graduate course (DHs and East Asian Studies: Theory and Practice), as well as the role and significance of the DARIAH-EU supported OpenMethods platform. A useful coda to her paper deals with the challenge of teaching, and how 'showing students techniques and tools compatible with Asian characters can reveal a curious dual approach to think about the intersection between multilingualism and DH and their impact on each other, which is both realistic and empowering' (p. 12). Flanders and Jannidis' paper, and Horvath's examples, demonstrate practices that reject the current taxonomies. These include diverse aspects such as traditional periodizations, as in the case of Flanders and Jannidis' work, or by combining areas of practice such as multilingualism and DH, in order to create new approaches to data. In addition to showing recent initiatives carried out on multilingualism and DH, both papers also offer examples of new ways of either combining traditional approaches to produce new knowledge, or deal with different types of data (e.g. Asian characters) by new means. This is precisely the same framework being argued for here, in terms of historical, hybrid linguistic data, and digital methods.

One form of linguistic hybridity currently underrepresented in the literature is contact languages, especially contact languages of the past. Typically, the domain of historical linguistics or corpus linguistics (e.g. Ebensgaard Jensen, 2014), DH has made steady progress in 'broadening out' to languages other than English. This also includes a broadening out to marginally represented groups. At least two reasons can be identified. The first is the (relative) cultural power of DH itself, and the so-called centres of attraction that are based in predominantly Anglophone countries, particularly the USA and parts of Europe (Risam, 2017). In this regard, the 'opening out' towards greater inclusivity is 'to some extent hampered by a lack of linguistic diversity' (Mahony, 2018, p. 371). Secondly, one must also realize the limitations of being able to work with multiple languages in multiple technologies, and the restrictions imposed by both language and the computer. To date, 'most available methods for computational text analysis are developed to operate on monolingual corpora' (Maier *et al.*, 2022, p. 19; see also Reber, 2019). Dombrowski (2020) has (usefully) complicated the process of preparing non-English texts for computational analysis, interrogating the very notion of what a 'word' is.[11] She shows how preparing texts for computational analysis is not a straightforward task, and that, at the current state of the field, one is only able to deal with a restricted number of languages—and a restricted number of tasks.

The next sections in this article look at the *Dictionnaire* as a hybrid linguistic space. They discuss forms of digital representation as it currently stands, and the implications of such representation for researchers. Not only does the issue become problematic of defining what a 'word' is in MLF research, but the appropriate descriptor for the variety itself has also been called into question. These issues can be seen to be a case-study of the representation of contact languages in DH more generally as well as examples of definitional elusiveness, as discussed later on.

2 The *Dictionnaire* as a hybrid linguistic space

One of the main issues in linguistic diversity in DH is precisely 'where the data come from' (Galina Russell, 2014, p. 308). The problem of 'diverse forms of data' is ongoing. It is not exclusive to DH. Fiormonte, for example, has noted that 'it is difficult to quantify the costs of monolingualism in DH' [difficile quantificare i costi del monolinguismo nelle DH] and that 'these costs are exacerbated by the specific disadvantage associated with the difficulty in translating into a lingua franca the results of research carried out on non-anglophone cultural objects (and often plurilingual, such as ancient texts, etc.)' (Fiormonte, 2017, p. 123).[12] While data constitute one side of the coin, another important aspect in multilingual approaches to text analysis is the tools and methods for language analysis on which this approach is based. It is true that 'for under-resourced languages the absence of preprocessing tools and corpora is a major obstacle' (Vanetik and Litvak, 2019, p. 20). In order to achieve good results, large high-quality annotated corpora are needed for training.

These questions become central when attempting to digitally (re)present forms of hybrid language in the past. One example of a hybrid language no longer used (written or spoken) is the so-called MLF—an extinct trade language, with a Romance lexical base, used around the Mediterranean basin, from anywhere between the late medieval to early modern periods. Therefore, part of the analysis which follows traces how researchers have aimed to discern the presence of MLF in historical documents.[13] The reality of Lingua Franca 'is not easily tangible, as has been remarked throughout the history of its study, analysis, and description'. Indeed, for over a century, 'the notion of Lingua Franca has been constantly constructed and deconstructed' (Selbach, 2017, p. 253). I suggest that what is missing from current analyses of MLF is not a disregard of the tools from historical linguistics, but that they have been applied incorrectly or inconsistently (Brown, 2022). Part of this article provides an analysis of one of the most well-known documents which purportedly records MLF phenomena—the *Dictionnaire de la langue franque* of 1830, now held in the Bibliothèque Nationale de France (BNF). The *Dictionnaire*, composed by an anonymous author, and published in Marseille, has been termed 'the only comprehensive source' [*die einzige umfassende Quelle*] by Schuchardt (quoted in Coates, 1971, p. 25). When these factors are taken into consideration—that is, when the text is placed within its proper sociolinguistic and historical context—the unremarkable linguistic phenomena recorded in the *Dictionnaire* are brought to light.

Selbach notes how certain researchers 'seemingly over-endorse the notion [of lingua franca] and have a tendency to provide a larger and highly inflated corpus (e.g. Rossetti, 2002); others deconstruct the notion of Lingua Franca at large, questioning the very idea of a specific language entity (e.g. Aslanov, 2002)'. It is (presumably) a similar language, if not the very same, to the one Burke (2004, p. 127) refers when he describes 'the original lingua franca' as being 'a mix of Venetian, Portuguese, and Arabic that was known in the Mediterranean world in the Middle Ages'. Even a study looking specifically at MLF in the Mediterranean in a historical framework, where we may hope to find traces of its presence, warns us not to confuse the term 'lingua franca' with '*Lingua Franca* proper' (Cremona, 1997, p. 53; his emphasis). Here, we are told that, during the seventeenth century, MLF was 'much in evidence during this period in spoken exchanges between Moslems and Christians' (p. 53).[14] Nolan has recently investigated the 'fact and fiction' of Lingua Franca studies, noting that 'the potential unreliability of some witness' accounts combines with the native language(s) bias of the writers and idiolectal variation, rendering their examples of Lingua Franca less credible' (2020, p. 83). In yet another case, Wansbrough's (1996) study conflates, in a purposeful way, both transfer mechanisms and the object of his analysis, which spans a period from 1500 BCE to 1500 CE. His focus is on the standard procedures of contact, noting that '*Lingua franca* refers to the several natural languages that served as vehicle in the transfer, but also to the format itself' (p. vii).

The question of naming the language variety described in the *Dictionnaire* is crucial for the way in which researchers have subsequently categorized the data they have worked with—be it from the *Dictionnaire*, or any other source. The problem is particularly acute when the goal has been to ascribe MLF phenomena to *other* linguistic varieties. These varieties have usually been national standards of Romance vernaculars (excluding some obvious influences of Arabic and Turkish), thereby disregarding the potential dialectal variation around the Mediterranean basin which may have contributed to the formation of MLF. In this regard, the glottonym *lingua franca* is no different to any other linguistic descriptor. As Pountain (2016, p. 638) has said, 'use of a language name does not imply the existence of a readily identifiable *Abstand* language'. Similarly, adoption of the term *lingua franca* by the anonymous author of the *Dictionnaire* does not imply a corresponding reality in either a diachronic or synchronic sense. The glottonym *lingua franca* is related to the broader historical context and linguistic historiography of late nineteenth-century philology in which Schuchardt was himself immersed. This, too, has had implications for perpetuating the MLF myth.[15] Let us briefly consider this context before moving to the previous taxonomies that have been proposed for MLF data, and how DH can help to shed light on contact languages of the past.

Many of the MLF headwords appear numerous times for different French terms. For example, the MLF entry *adesso* is given as the corresponding element for both the French *maintenant* and *présent*. Similarly, MLF *cascar* is provided as the only entry for French *s'écrouler*, *glisser*, *tomber*, *couler*, *écouler*. This may be evidence for a lack of lexical diversification in MLF. Other studies of contact languages have shown that, after initial mixing takes place, we are likely to see wide polymorphy in the koine pool before levelling occurs (Britain, 2012, p. 224).[16] It is only in a subsequent period that multiple features may become levelled, but different semantic terms are likely to remain. In this sense, the entries recorded in the *Dictionnaire* do not appear to be representative of a separate variety, whether it has creolized or not.[17]

If one eliminates the duplicates in the MLF column from the 2,120 total number of entries in the *Dictionnaire*, 1,887 unique lexemes remain. Even a cursory glance at these entries, however, suggests a lack of complexity and diversity. As such, it is difficult to understand how they could be presented as evidence for a distinct variety of Romance. Of these 1,887 unique items, 697 are written the same as standard Italian.[18] Of these 697, a further 583 items may be subtracted whose orthography is clearly designed to mirror Italian pronunciation, but whose phonology is ultimately unknown. These include items such as *taska*, *soupa*, *pépé*, etc. Combining both categories (same lexeme and orthography), one arrives at a figure of 1,280 unique MLF items; in other words, a sizeable 68% of the total corpus can be said to be the same as standard Tuscan forms. This figure can be further increased, by considering items whose stem is exactly the same as standard Italian, but whose desinence shows a non-Italian element, for example, *dios*, *attention*, and *pensiéré*. These items bring the total count to 1,366, or 72% of the total corpus. Of the remaining elements, one finds a different lexical root (as in *boriqua* 'donkey'), or some feature indicative of phonetic change (*poder* 'to be able'). Nevertheless, these categories combined are relatively small when compared with the overall corpus, representing only 26% of all MLF entries.

At what point can we say this variety ever existed, particularly when the recorded forms show such a high degree of similarity, if not exactitude, with standard Italian, let alone other Romance varieties? This is not to deny that the author of the *Dictionnaire* recorded forms that historically existed, or that what has been passed down is a faithful attempt to record certain forms of this language by the author. But these words are not unexpected in an environment of language contact, and show no evidence of having creolized or having undergone any subsequent development as a contact variety. It may be typologically convenient to refer to mixed phenomena as a separate 'pidgin', 'lingua franca', etc. But the phenomena described are the expected results whenever any two (or more) varieties come into contact, especially when such varieties are in such a state of flux over a long period of time and are typologically related. How can DH help to account for such hybridity, and how might we make a space in DH work itself for hybrid data? Why is such work necessary?

Many existing catalogues or descriptors of hybrid languages do not fully reflect the actual linguistic variation contained in textual data from the past. In order to provide a more complete picture of this linguistic reality, and therefore cultural realities, of past persons, the starting point must be to allow for the creation of records that reflect such variety. Thieberger (2017, p. 433) notes that much remains to be done in order to extend the reach to digital language archives, such as 'assisting in locating legacy collections, describing and digitizing them, connecting with source communities/individuals, creating a means for online annotation (crowdsourcing), and of valuing the collections (both monetarily and academically)'.[19] Making a space for hybrid data is also important for other areas of investigation, especially data-driven approaches to studying semantic change and changing vocabularies. Hengchen *et al.* (2021) look at large datasets in historical newspaper archives in Dutch, Swedish, Finnish, and English, for example. Don and Knowles (2021) have also pointed to the possibility for DH to create new opportunities to scale up practical linguistic analysis using large datasets for under-represented varieties, such as indigenous languages. This work is also fundamental for digital representation of hybrid languages. But it has also remained elusive.

3 Digital hybridity as definitional elusiveness

What could the term 'definitional elusiveness' encompass? Rather than providing a definition of this term with watertight edges, I have in mind particular types of data, frameworks, theories, or infrastructures, which do not and cannot be categorized using currently existing terminologies. This is because these data either overlap across multiple taxonomies (multiple languages, datasets, etc.), or because the existing tools available for computational analysis are often designed to deal with the specifics of particular types of data in the first place. A notion of definitional elusiveness can also be useful in (re)thinking broader questions in DH itself. It prompts us to reframe Fiormonte's question posed above: What are the characteristics of multilingual DH? Definitions so far have been likely to beget more definitions, in an attempt to refine or incorporate forever more elements that may or may not be adequate for future work, leading to a 'kitchen sink' analysis.[20] In terms of the case-study being presented in this article, this is even the case for recent tools such as *The Atlas of Pidgin and Creole Language Structures Online*, where one might hope to find some trace of MLF.[21] The lack of representation of MLF here may be because the status of the variety still remains undefined—and will likely remain so. Questions are perennially being asked about whether it is indeed a pidgin, a creole, a koine, etc. Where exactly should MLF data be situated in a digital environment? It remains definitionally elusive.

One reason for the lack of focus on marginalized groups is the interdisciplinary nature of DH work. While scholarship has looked at a range of 'lost' voices from the past (including traditionally marginalized groups from history, linguistics, and so on), there exists an inverse relationship between powerful institutions and nations, and those groups themselves. In other words, if the centres of power are in the USA and

Europe, DH projects and funding bodies are much more likely to prioritize the peoples and languages which constitute those same centres. In the case of lost languages from the past, which do not find an easily transferable mapping on to contemporary nation states or institutional frameworks, more work must be done *across* borders in order to provide a more complete representation. This contemporary mapping also finds itself in a position of definitional elusiveness.[22]

A second reason for the lack of representation of forms of linguistic hybridity has been a lack of texts and/or access to such texts in digital forms and repositories. This is a problem with the data itself—or data organization, rather. Even recent projects that have a collaborative focus on 'big data' are often constrained by the extant tools and digital repositories made available by large institutions. For example, Sangiacomo *et al.*'s (2022) project presents a workflow for the optical character recognition (OCR) of digitized versions of early modern books freely available in the public domain, producing a final corpus consisting of 498 OCRed titles in three different geographical locations (Britain, France, and the Dutch Republic), in three different languages (English, French, and Latin).[23] The main providers of these works included Google Books, BNF, Munich Digitization Centre at the Bavarian State Library, Early English Books Online, and occasionally e-rara.ch and archive.org. The workflow they provide will certainly be useful for future scholars seeking to reduce the amount of time in improving OCR accuracy for computational text mining. This workflow also has real benefits for 'mixed' language data from the past that has yet to be subjected to such OCR workflows.

A third reason for the narrow focus on standard languages has to do with 'the standard processes that the discipline [DH] adopts' (McGillivray *et al.*, 2020). Given that much focus in DH and natural language processing (NLP) research has been on developing new computational systems or improving existing ones, researchers are inclined to adopt standard datasets 'using reproducible methods'. Consequently, the incentive for NLP researchers is to work on very restricted datasets and languages, 'leading to the development of tools which are optimized for those datasets and languages'. A corollary of this decision is that it 'drives research towards a very specific direction, away from the idiosyncratic features displayed by historical languages and DH data'. However, the 'idiosyncratic features' are precisely those elements which are likely to be the most interesting, the most useful, and the most diverse for humanities researchers seeking evidence of ways in which historical data can be represented in a digital space. Put more simply, there is a lack of computational models which can deal with mixed language textual data.[24] DH tools represent both an opportunity and a challenge for researchers in this sense. Previous work has laid the foundation for dealing with historical textual data. The challenge now is to exploit such tools to explore the full range of data available, as well as to provide a digital infrastructure that is suitable (enough) to represent the diversity inherent in the data itself.[25]

Definitional elusiveness is another reason for the lack of diversity in DH projects. Here, Galina Russell's question 'what's your DH accent?' is useful since it opens itself up to a variety of interpretations, from different types of interests, scholars, cultural backgrounds, linguistic diversity. Indeed, allowing for an 'openness' to this question in the form of a non-answer (or definitional elusiveness) permits a higher degree of flexibility rather than providing partial responses to questions that are intended to provoke precisely this form of fluidity. The question of definitions appears to be one such issue to which practitioners in the field of DH have constantly returned.[26] As Galina Russell (2014, p. 307) has stated:

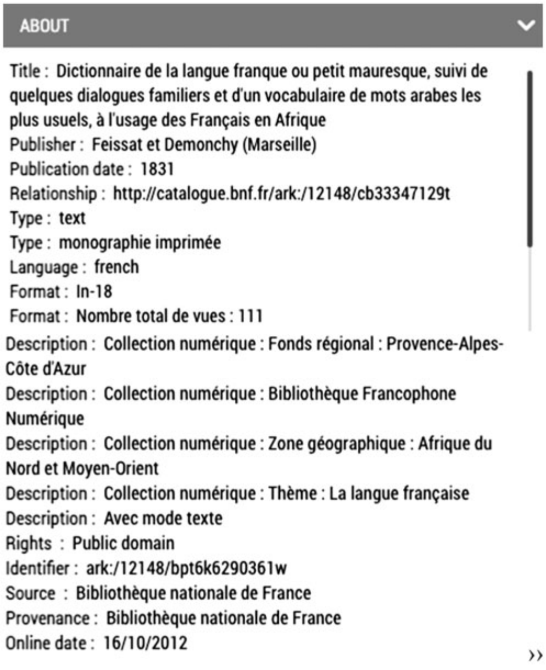
> Although historically the DH community has grappled with a definition, over the past few years there have been more vocal disagreements as we struggle to define what DH 'is' and what DH 'does'. As Gold writes in his introduction to 'Debates in the DH', this is 'a field in the midst of growing pains as its adherents expand from a small circle of like-minded scholars to a more heterogeneous set of practitioners who sometimes ask more disruptive questions. (Gold, 2012)

The possibility of leaving open the question of definitional elusiveness in scholarly enquiry means that no one discipline can lay claim to the remit of particular types of analysis, group, identity, language, and so on. These words recall Risam's useful terminology of naming the 'challenge of recognizing DH work', but also the way in which a 'DH accent' might help to 'negotiate some of the challenges to building a global DH community' (2017, p. 383). Again, it is precisely in DH work where such community and diverse accents can find a forum in which to flourish. The difficulties identified in DH are often the same as those of traditional disciplines: being able to cross-collaborate, identifying funding agencies for inter-disciplinary research, as well as practical questions of time zones and languages in which to work. Being able to problematize the issues at hand is a good first step in ensuring greater diversity in the field, but it is also one in which DH is apt to correct for. It has already been remarked that 'DH is uniquely positioned to take on such challenges by virtue of its combination of humanities knowledge and computational focus' (Galina Russell, 2013, cited in Risam, 2017, p. 383). In short, the shaping of the discipline is an ongoing venture. It is also one where modern languages academics are ‘fighting back’, pushing against the tide of ‘how heavily Anglophone a purportedly “global” DH really is’ (Pitman and Taylor, 2017, para 25). The data described above from the *Dictionnaire* represent one type of dataset which is not often the purview of enquiry from scholars outside of modern languages or linguistics. In this sense, ‘broadening out’ DH work to include even more plurilingual and plurifunctional methods of analysis can help to mould the discipline into the future.

The fact that the BNF and Gallica records only lists French as the language of the *Dictionnaire* obscures its true multilingual nature. If multiple entries were allowed, then it would signal the plurilingual nature of the document and of the archive itself. By allowing only one linguistic taxonomy, the multilingualism present in the archive remains ‘hidden’, out of view, and not immediately obvious to researchers looking for instances of language contact and therefore contact between peoples.

The current entry for the *Dictionnaire* and associated meta data as it appears in the Gallica interface is shown in Fig. 1.

The metadata shown above are themselves interesting, since the descriptor is given in English (e.g. *Description*) while the information is almost all in French (e.g. *Avec mode texte*, although some elements are in English, as in *Public domain*). Gallica provides two entries that provide metadata on the linguistic taxonomy for this particular object. The first is simply ‘Language’, which in this case is given as ‘french’. The second entry is the fourth ‘Description’ tag. Under ‘Collection numérique > Thème’, a further tag has been added as ‘La langue française’. While ‘thème’ undoubtedly corresponds to a sense of ‘subject matter’ or ‘language of subject matter’ in the Gallica entry, the fact that only one tag has been added, without further description of the hybrid nature of the document, means that the entry has been broad-brushed with the same standard language description. In short, the data expose what has been called an ‘invisible collection’. Thieberger has pointed out how digital language archives are invisible to aggregated searches, while other repositories (including many institutional repositories) ‘have language content that is not noted in the collection’s catalogue, so it is impossible to locate at all via search based on language names’ (2017, p. 423). In the case above, the other tags usefully make some reference also to the regional and geographical locations with which the document is concerned (e.g. first Description: ‘fonds régional’) and third Description > ‘Zone géographique: Afrique du Nord et Moyen-Orient’, although no ISO-639-3 (International Organization for Standardization) identifier is provided for language.

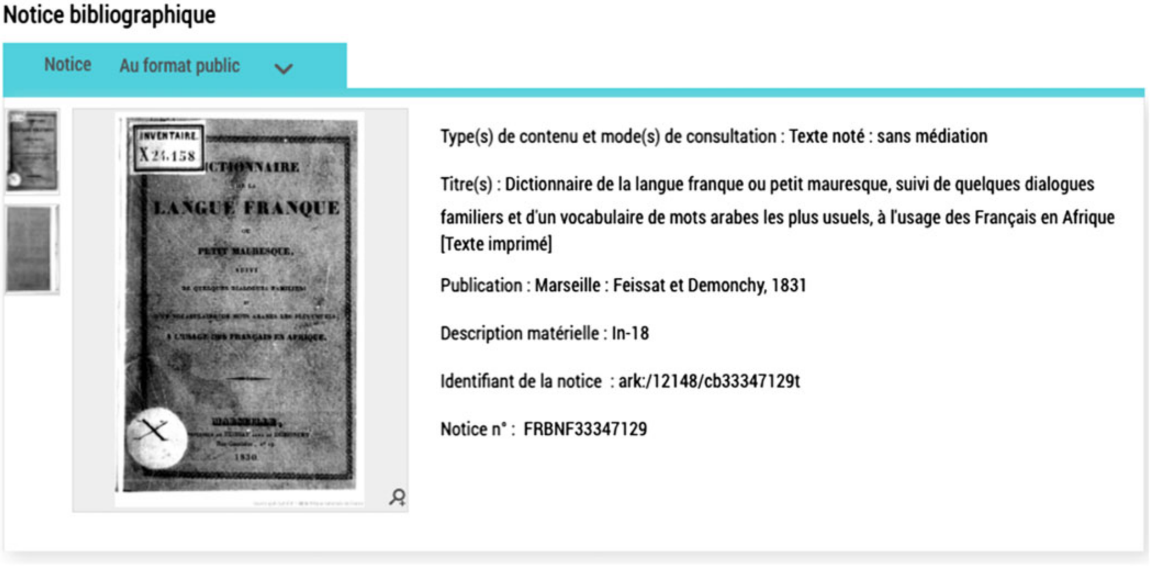


Title : Dictionnaire de la langue franque ou petit mauresque, suivi de quelques dialogues familiers et d'un vocabulaire de mots arabes les plus usuels, à l'usage des Français en Afrique
Publisher : Feissat et Demonchy (Marseille)
Publication date : 1831
Relationship : http://catalogue.bnf.fr/ark:/12148/cb33347129t
Type : text
Type : monographie imprimée
Language : french
Format : In-18
Format : Nombre total de vues : 111
Description : Collection numérique : Fonds régional : Provence-Alpes-Côte d'Azur
Description : Collection numérique : Bibliothèque Francophone Numérique
Description : Collection numérique : Zone géographique : Afrique du Nord et Moyen-Orient
Description : Collection numérique : Thème : La langue française
Description : Avec mode texte
Rights : Public domain
Identifier : ark:/12148/bpt6k6290361w
Source : Bibliothèque nationale de France
Provenance : Bibliothèque nationale de France
Online date : 16/10/2012

Figure 1. Metadata from Gallica for the *Dictionnaire*
Source: https://gallica.bnf.fr/ark:/12148/bpt6k6290361w.texteImage

Similarly, the corresponding entry for the *Dictionnaire* in the BNF catalogue does not include an entry for language (see Fig. 2).

Given the ambiguous taxonomy of the language that the document describes, it is also not immediately obvious which other linguistic category should be included in the catalogue anyway. The *Dictionnaire* (purportedly) contains evidence of linguistic contact from a wide array of influences including Arabic, Catalan, Latin Luso-Arabic, French, Greek, Occitan Portuguese, Provençal, Spanish, and Turkish (Brown, 2022). The issue concerning the linguistic descriptor of the *Dictionnaire* leaves itself primed, therefore, as yet a further example of definitional elusiveness. The questions therefore arise: whose language? whose DH? It is difficult to provide precise reasons why further language categories are not provided for the *Dictionnaire*, or indeed for any cultural object which contains data in more than one language or variety. In this sense the hegemony of standard languages holds particular cultural dominance over other forms of linguistic (and non-linguistic) hybridity (and one wonders, in this case at least, about the cultural weight of French for language descriptors in library catalogues, in a French-speaking country). Gestures towards possible solutions could include offering search criteria in multiple languages, or even search criteria for languages whose status is



Notice bibliographique

Type(s) de contenu et mode(s) de consultation : Texte noté : sans médiation

Titre(s) : Dictionnaire de la langue franque ou petit mauresque, suivi de quelques dialogues familiers et d'un vocabulaire de mots arabes les plus usuels, à l'usage des Français en Afrique [Texte imprimé]

Publication : Marseille : Feissat et Demonchy, 1831

Description matérielle : In-18

Identifiant de la notice : ark:/12148/cb33347129t

Notice n° : FRBNF33347129

Figure 2. Metadata from the BNF catalogue for the *Dictionnaire*
Source: https://catalogue.bnf.fr/ark:/12148/cb33347129t

unknown. Describing this cultural object *only* in terms of French essentially subjugates the linguistic hybridity inherent in the *Dictionnaire*, thus rendering it invisible in an 'ecology of languages' where standards relegate other forms of language to an inferior status. Nevertheless, the linguistic taxonomy provided by the catalogue record of just one language—and a standard one—means that the hybrid nature of documents from the past cannot easily be found via the current search parameters. The optimum method for digitally representing such variation is one with which researchers are only now coming to terms.

What role could digital tools for corpus linguistics and computational approaches to language have for contact languages? While the analysis presented in Section 2 focused on word frequency and the issue of lexicalization in the *Dictionnaire*, DH can be useful in other areas of linguistic investigation, including modelling the emergence of contact languages and in helping to disambiguate derivational morphology. What would be useful in computational approaches to MLF data (or any contact variety, for that matter), would be further technical advances showing the hybrid nature of the data itself. For example, the MLF entry for 'shirt' is given as *camischia*. The corresponding French translation in the dictionary is given as *chemise*. However, it is by no means clear which standard language (if any) has contributed to the morphology of this particular lexeme.[27] The feminine singular morphological marker *-a* is also feminine singular in several Romance standards and dialects, while the lexical root *camisch-* resembles a number of other varieties, including French (*chemise*), Spanish and Portuguese (*camisa*), Italian (*camicia*), and to a lesser extent Romanian (*cămaşă*), not to speak of the question of how much weight might be assigned to dialectal lexicalization (e.g. Sicilian *cammisa*, pronounced in some varieties /kam'mi.sa/). How can computational tools help to 'pull apart', or disambiguate, such variation at the base of a variety like MLF? Ideally, one would like to be able to see which possible donor languages have contributed to the emergence of contact varieties even when such donor languages and the relative 'weight' of their contributions are unknown.[28] Even in the case of a tool like Voyant, for example, it would be useful to be able to perform analysis across multiple standard languages, rather than those currently available.

Scholars have already begun tackling these problems in computational text analysis. The current gaps in computational methods for social sciences research have recently been the subject of a paper by Baden *et al.* (2022). In it, they identify three issues that limit the utility and obstruct the progress of computational text analysis. The first is the prioritization of technological issues over validity concerns, giving rise to the operationalization of social scientific methods. The second refers to a mismatch between the focus of computational analysis on extracting specific contents and document-level patterns, and the need for researchers in the social sciences for multiple, often complex contents in the text. The third issue is more direct to our purposes here. It concerns the dominance of English language tools which depress comparative research and inclusivity towards scholarly communities examining languages other than English.

Another issue in the application of tools from computational linguistics and NLP to contact languages is scalability. Language independence 'is commonly presented as one of the advantages of modern, machine-learning approaches to NLP' (Bender, 2011, p. 1), with the aim of leveraging training data from a first language to a second language. However, Bender's study made it clear that 'we are not truly evaluating language independence with any systematicity' and that 'truly language-independent technology requires more linguistic sophistication than is the norm'. Related to scalability is the issue of working across multiple software, as researchers often do (or are required to do), in order to combine or concatenate different tools, 'leveling if not reversing the advantages of computational processing' (Baden *et al.*, 2022, p. 2). These authors identify a general theme of 'specialization before integration' as an umbrella term, pointing out how many available computational methods are 'built as standalone tools, using different coding languages and data standards and offering specialized interfaces that do not support concatenation within more complex pipelines' (p. 7). Related is the third 'gap' they identify in computational text methods, 'English before everything'. English reigns supreme. In part, they note, this is due to the simple morphology of English verbs and nouns and the fact that English has become naturalized in computational text analysis. This is 'hard to remove from existing technologies'. Consequently, they note that researchers regularly decide *against* investigating other, less studied language communities, 'defaulting to the same few research sites studied already by their anglophone colleagues' (p. 11). The comments they make are symptomatic of the kinds of issues with which any researcher finds themselves confronted when dealing with non-English data, let alone contact languages, since 'trying to force different languages into the corset of English-like language structure, researchers confront many open questions and preprocessing needs' (p. 10). Nevertheless, for some contact varieties, such as creoles, the application of certain techniques, such as modelling, has already begun to appear.

This is the case for Mauritian Creole. Researchers have described the methodological difficulties involved, such as obtaining detailed empirical data on the origin and structure of the creole (Jansson *et al.*, 2015). More recent work has shown how tools of biological evolution can help to model contact languages. For example, the emergence of creole languages has shown to be a dynamic process that mimics the process of creole formation in American and Caribbean plantation ecologies (Tria *et al.*, 2015; see also Lent *et al.*, 2021). Such models allow for further studies on the emergence of languages in contact, including in historical perspective, as well as the testing of specific hypotheses. This work is crucial for various activities which make use of computational models relating to contact phenomena across many cultural domains, including modelling and the emergence of dialects, translation, and digital mapping of the movement of peoples. The focus of this article is not to engage in the various advantages or disadvantages of any particular software or system. Digital tools for corpus linguistics may provide excellent avenues in understanding contact languages and linguistic hybridity better.

While the example presented in this article focuses on data from a historical dictionary in the past, the questions alluded to above are not just confined to situations from history. Indeed, Pitman and Taylor (2017, para 29) argue for the importance of incorporating new digital genres (such as hypermedia fiction or game art), and the necessity of ensuring they do not 'build upon an Anglophone heritage, but respond to and continue a rich tradition of cultural, literary and artistic experimentation (…) in many different languages and countries'. Forms of cultural hybridity and questioning of the traditional nation-state model have been on the agenda for many years in modern languages departments, including in both teaching and research.[29] To take one example, questioning what 'Italy' is, and what 'Italian' is, is often included in a variety of subject material in an undergraduate degree, including film studies, linguistics, literature, and more. It seems important, therefore, that DH also reflects the hybridity of these types of subject matter, particularly as more and more material becomes digitized from film, language, literature—and that models evolve in an appropriate way to reflect the hybrid nature of the subject matter itself.

4 Conclusion

In her presentation at the McGill 'Spectrums of DH' series, entitled 'What's a word? Multilingual DH & The English Default' (2020), Dombrowski shows how the question of what a 'word' is, does not have a straightforward answer. In the Anglophone bubble, the interrogative 'feels almost like a nonsense question'. Similarly, I have argued that the question of what 'language' a particular text is written in can have multiple solutions and sometimes no solution. Serious efforts to enhance the multilingual and multicultural nature of DH, including through the ADHO, have already been made and are ongoing. Other forms of representation and international networks are also beginning to take shape, providing new ways for language sensitivity and diversity in DH to emerge. Some of these focus on the use of English itself as the dominant language in DH projects, while others look at the use of non-Latin scripts in digital scholarship. Questions such as these are becoming more and more insistent, as more data become available and DH tools become increasingly more sophisticated in order to deal with the myriad issues that arise in processing such data. As Maier *et al.* (2022, p. 19) have recalled in relation to communication research, 'the necessity for analytic techniques suitable to explore large-scale multilingual text collections has never been more pressing'.

Historical linguistic data from a dictionary, the *Dictionnaire de la langue franque*, can be used as a case-study to show the hybrid nature of the language used in this text. The data discussed in this article, and my own background in a modern languages department, are also situated within the same framework of Pitman and Taylor's (2017, para 6) call for transdisciplinarity, or what they describe as 'disciplinary openness aimed at addressing complex research questions that do not sit neatly within any traditional discipline'. Put another way, we need to teach ourselves to be 'culturally multilingual, not just linguistically multilingual' (Dombrowski, 2021). Certain tools are already making access to data, and processing that data, more easily manageable. Language technologies and their associated infrastructures (e.g. CLARIN and ELG) can offer a wide variety of applications to 'discover, explore, exploit, annotate, analyse, or combine language data', but are still limited in being able to provide computational tools for hybrid languages.[30] Given the hybrid nature of all languages, the importance of considering contact languages (whether extant or extinct) is also important for the ongoing development of such tools since 'computational tool developers are likely to miss or inadequately model important textual properties, and are bound to laboriously reinvent the wheel through trial and error' (Baden *et al.*, 2022, p. 13).

DH still has a long way to go if a more robust answer can be provided to Fiormonte's (2012, p. 59) provocative question: 'Is there a non-Anglo-American DH, and if so, what are its characteristics?'. While I have focused on questions of linguistic hybridity, the arguments presented in this article surrounding definitional elusiveness could be easily applied to other forms of discipline hybridity as well. I am thinking in particular of texts, peoples, places, cultural objects, etc. which do not easily fall into 'categories' that are typically taken as focus for scholarly enquiry in the academy. I am not, however, arguing for an open-ended black box into which all forms of miscellany can or should be shoved. Given the ongoing tension in DH to 'define' both itself and 'others', one should leave open the possibility for definitional elusiveness.

I have argued for additional taxonomies to be provided in DH, particularly a taxonomy of definitional elusiveness. Part of the aim of the article has been to respond to Risam's (2017, p. 378) comment that the discipline 'needs ongoing theorization for navigating these differences and negotiating between local practices and global perceptions of DH'. I have also advanced reasons for the (as yet) lack of focus on marginalized groups, including the interdisciplinary nature of DH work, lack of access to texts, and access to digital repositories—but also the narrow focus on standard languages. In the case of the *Dictionnaire*, I have shown how very little linguistic metadata exists in two separate online catalogues, including the records in Gallica and the BNF. When linguistic information is provided, it is not in a standard format, it provides only a standard language, or it is inadequate to describe the kind of information the document contains. More work is needed, from both DH and linguistics, to provide a more complete picture of the heterogeneity that characterizes cultural objects of investigation and the disciplines themselves. Both disciplines are uniquely positioned to tackle this challenge. The work will be better if it is done together.

Notes

1. The author is grateful to John Kinder and Francesco De Toni for feedback on an earlier draft. Thanks also to two anonymous reviewers whose comments greatly helped to improve the article.
2. https://multilingualdh.org/en/.
3. See also the resources on 'Towards Multilingualism in Digital Humanities' hosted on Zenodo and available at: https://zenodo.org/communities/multilingual-dh/
4. https://adho.org/administration/multi-lingualism-multi-culturalism. See also the description provided in Mahony (2018, p. 384).
5. http://www.globaloutlookdh.org.
6. I use the term 'contact' in a broad sense, following Crystal (2008, p. 107): 'a term used in sociolinguistics to refer to a situation of geographical continuity or close social proximity (and thus of mutual influence) between languages or dialects'. On the following page, he notes that 'contact language or contact vernacular is also sometimes used to refer to a pidgin'. Since pidgin itself has a particular, codified meaning in sociolinguistics, and given that the MLF is not a pidgin, in this article I use the term contact in line with the first sense provided in Crystal. It is also worth noting recalling Wei's (2018) remarks that 'all languages are contact languages: have always borrowed from and mixed with other named languages'. For a recent article which discusses the notions of what constitute a 'language' and what a 'dialect', and which argues against such a distinction in order to reveal 'hidden' multilingualism in Italy, see Tamburelli (2014).
7. The full title as shown on the first page is *Dictionnaire de la langue franque ou petit mauresque, suivi de quelques dialogues familiers et d'un vocabulaire de mots arabes les plus usuels*; *à l'usage des Français en Afrique*.

[8] For further comments on the position of Modern Languages and nation-states in digital humanities, see Pitman and Taylor (2017, para 31).

[9] Also, see Dombrowski's useful Github repository of 'multilingual-dh-bibliography' (https://github.com/multilingual-dh/multilingual-dh-bibliography).

[10] A workshop, held in 2020, was hosted by the 'Language Acts and Worldmaking' project with the support of the 'Cross-Language Dynamics: Reshaping Community' project. Both projects were funded by the AHRC as part of the Open World Research Initiative. For further information, see https://languageacts.org/digital-mediations/event/disrupting-digital-monolingualism/ as well as the related report on multilingualism in digital theory and practice (https://zenodo.org/record/5743283#.Ym88XdNBx-1). Other initiatives include an online website, GitHub organization, and mailing list for Multilingual DH (https://multilingualdh.org/en/), a repository for Multilingual Research and Resources for DH hosted by McGill University (https://www.mcgill.ca/digital-humanities/article/multilingual-research-and-resources-dh), workshops on specific topics such as Multilingual Sentiment Analysis hosted at the University of Texas Library (https://guides.lib.utexas.edu/c.php?g=873758&p=6589171), and the recent conference *Digital Humanities Beyond Modern English: Computational Analysis of Premodern and Non-Western Literature* hosted at the University of Texas at Austin (https://quantitativecriticismlab.github.io/DHBME2/). See also the Australian Text Analytics Platform, which allows data-driven research by extracting and analysing machine-readable information from within unstructured text (https://www.atap.edu.au/).

[11] The question of what a *word* is can be seen as another example of definitional elusiveness, particularly so when it comes to linguistic hybridity but this is also the case for non-mixed data too. The Australian Text Analytics Platform includes a useful description on 'understanding text as data', where it is noted that 'the term *word* is problematic'. Marciniak (2016) reports on the assemblage of algorithms in mapping a public discourse around Big Data in science and politics.

[12] Difficile quantificare i costi del monolinguismo nelle DH, ma si può ipotizzare che questi siano addirittura acuiti dallo svantaggio specifico costituito dalla difficoltà di tradurre nella lingua franca i risultati di ricerche effettuate su oggetti culturali non anglofoni (e spesso plurilingui, come i testi antichi, ecc.).

[13] While the analysis presented below is limited to examples of lexical diversification, previous scholars have also investigated various aspects of MLF at other linguistic levels and from various points of view, including orthography, (Operstein, 2017a), morphology (e.g. Operstein, 2018), syntax (Operstein, 2017b), the question of its pidginization versus koineization, and language change (see related papers by Operstein, 2017c, 2021; Nolan, 2020). Additional investigations at the morphological level will help to further discern the hybrid nature of MLF. To quote but one case, the 1sg. pronoun *mi* 'I' has been described as Venetian (Cifoletti, 1989, pp. 64–65, 1991, p. 35; Burke, 2004, p. 127) but it is also a standard Italian form and present in many other Italo-Romance vernaculars (including the so-called 'dialects of Italy').

[14] This appears to be the same sense in which Migliorini (1960, p. 559) adopts the term as well: 'Along the coasts of the Mediterranean, especially in the east, Italian is still very well known in spoken usage in the simplified form of "lingua franca", and in written usage as a diplomatic language' [Sulle coste del Mediterraneo, specialmente orientale, l'italiano è ancora molto noto nell'uso parlato sotto la forma semplificata di 'lingua franca', e nell'uso scritto come lingua diplomatica]. The non-singularity which *lingua franca* encompasses has been noted by Tommasino (2017, p. 35), writing that it must include 'markedly different linguistic varieties'.

[15] Here, it is worth noting that Schuchardt makes reference to the French orientalist Jacques Auguste Cherbonneau's *Observations sur l'origine et la formation du langage arabe africain* of 1855. Cherbonneau described the *langage arabe africain* as 'a curious amalgamation of Spanish words, Italian terms, and French expressions' [un amalgame curieux de mots espagnols, de termes italiens et de tournures françaises]. In a footnote to her translation of Schuchardt's work, Venier (2012, p. 15, n. 1) comments on this description, noting that 'this alternation of nouns leads one to suppose a relationship *that, in reality, does not exist*' [questo alternarsi di sostantivi lascia supporre una relazione *che nella realtà non sussiste*] (my emphasis).

[16] Britain (2012, p. 224) describes levelling as 'the eradication of marked linguistic features, marked in the sense of being in a minority in the ambient linguistic environment after the contact "event," marked in the sense of being overtly stereotyped, or marked in the sense of being found rarely in the languages of the world and/or acquired late in first language acquisition'.

[17] Derivational morphology also helps to explain other forms recorded in the *Dictionnaire*. But these endings similarly reflect patterns of already attested word-formation in other Romance varieties. In most cases, infinitives, always placed into the *-ar* category, are simply derivations from the form given for singular, masculine nouns (e.g. *dopio* > *dopiar; pranzo* > *prantzar*), with each item listed as a separate entry.

[18] In this count, I have included infinitives whose final vowel is subject to apocope (also considered a feature of standard Italian in certain phonetic environments, e.g. *parlar*). I have also included items showing diacritics, whose value is unknown in the *Dictionnaire*, as well as items whose orthography is ostensibly designed to mirror standard Italian (Tuscan), for example, *médiko*, *melio*, *bagniar*, and *picolo*.

[19] Digital methods can also help to reinterrogate current taxonomies of language boundaries and dialect classifications. For example, Tamburelli and Brasca (2018) recently developed an empirically based classification of Gallo-Italic through the use of dialectometry applied to atlas corpora.

[20] I borrow this term from analyses in statistical regression, in cases where long lists of independent variables are successively added to explain variance in a dependent variable. Analysts end up adding 'everything but the kitchen sink' to the regression, in the hope of finding statistical patterns.

[21] This database is available at: https://apics-online.info/contributions#2/30.3/10.0

[22] See also the recent comments in Spence (2021), who refers to 'una definición problemática'.

[23] For a study quantifying the impact of dirty OCR on historical text analysis in English, see Hill and Hengchen (2019).

24. Models are continuing to evolve rapidly, however. See Sarma *et al.* (2021).
25. A recent overview of the challenges of text analysis using deep neural networks in DH and information science can be found in Suissa *et al.* (2022).
26. On the question of whether digital humanities can be said to be a 'field' at all, see the useful comments in Horvath (2021, p. 2, foonote 1) and bibliography there.
27. For a study looking at word segmentation, unknown-word resolution, and morphological agreement in a Hebrew parsing system, see Goldberg and Elhadad (2013). For a paper applying automatic content analysis methods for political texts, see Grimmer and Stewart (2013). Maier *et al.* (2022) have recently evaluated machine a translation versus a multilingual dictionary approach for assessing strategies in topic modelling for multilingual text collections.
28. Similarly, Baden *et al.* (2022, p. 14) call on linguistics and discourse studies as disciplines which can offer 'a rich vocabulary for adding precision to operational choices, enabling their better translation into algorithmic procedures'.
29. Cf. Galina Russell (2014, p. 308) 'as the conversation veered towards aspects of race and gender, it was pointed out that the debate was mainly US-orientated and in itself was exclusive of other national contexts'. Galina Russell highlights the importance of incorporating 'scholars from a broader range of countries and linguistic backgrounds than are currently represented' (p. 308).
30. See for example, the currently available tools at: https://www.clarin.eu/.

References

Aslanov, C. (2002). Quand les langues romanes se confondent… La Romania va ailleurs. *Langage et Société*, **99** (1): 9–52.

Baden, C., Pipal, C., Schoonvelde, M., and van der Velden, M. A. C. G. (2022). The gaps in computational text analysis for social sciences: a research agenda. *Communication Methods and Measures*, **16**(1): 1–18.

Bender, E. M. (2011). On achieving and evaluating language—independence in NLP. *Linguistic Issues in Language Technology*, **6**(3): 1–26.

Britain, D. (2012). Koineization and cake baking: Reflections on methods in dialect contact research. In Wälchli, B., Leemann A., and Ender, A. (eds), *Methods in Contemporary Linguistics*. Berlin: De Gruyter Mouton, pp. 219–38.

Brown, J. (2022). On the existence of a Mediterranean lingua franca and the persistence of language myths. In Bennett, K. and Cattaneo, A. (eds), *Language Dynamics in the Early Modern Period*. Vol. **1**. London: Routledge, pp. 169–89.

Burke, P. (2004). *Languages and Communities in Early Modern Europe*. Cambridge: Cambridge University Press.

Cifoletti, G. (1989). *La Lingua Franca Mediterranea*. Padua: Unipress.

Cifoletti, G. (1991). L'influsso arabo sulla Lingua Franca. In Loprieno, A. (ed.), *Atti della Quinta Giornata Comparatistica*. Perugia: Dipartimento di Linguistica e Filologia Romanza, pp. 34–9.

Coates, W. A. (1971). The Lingua Franca. In Ingemann, F. (ed.), *Proceedings of the Fifth Annual Kansas Linguistics Conference*. Lawrence: University of Kansas, pp. 25–34.

Cremona, J. (1997). Acciocché ognuno le possa intendere: The use of Italian as a Lingua Franca on the Barbary Coast of the seventeenth century. Evidence from the English. *Journal of Anglo-Italian Studies*, **5**: 52–69.

Crystal, D. (2008). *A Dictionary of Linguistics and Phonetics*. Oxford: Basil Blackwell.

Dombrowski, Q. (2020). Preparing Non-English texts for computational analysis. *Modern Languages Open*, **45**(1): 1–9.

Dombrowski, Q. (2020). *What's a 'Word': Multilingual DH and the English Default*. https://quinndombrowski.com/blog/2020/10/15/whats-word-multilingual-dh-and-english-default/.

Dombrowski, Q. (2021). *Humanités numériques, цифровые гуманитарные науки, デジタル・ヒューマニティーズ: History and Future of DH Linguistic Diversity*. UCL Centre for Digital Humanities. https://www.ucl.ac.uk/digital-humanities/events/2021/apr/ucldh-online-histories-and-futures-linguistic-diversity-dh-rescheduled.

Dombrowski, Q. (2022). Non-English DH is not a thing. Talk presented at online conference *Digital Approaches to Multilingual Text Analysis*. https://metodhology.anu.edu.au/index.php/2022/01/20/digital-approaches-to-multilingual-text-analysis/

Don, Z. M. and Knowles, G. (2021). The digital humanities and re-imagined language description: A linguistic model of Malay with potential for other languages. *Digital Scholarship in the Humanities*. https://academic.oup.com/dsh/advance-articleabstract/doi/10.1093/llc/fqab101/6447178

Ebensgaard Jensen, K. (2014). Linguistics and the digital humanities: (Computational) corpus linguistics. *MedieKultur. Journal of Media and Communication Research*, **57**: 115–34.

Fiormonte, D. (2012). Towards a cultural critique of digital humanities. *Historical Social Research*, **37**(3): 59–76.

Fiormonte, D. (2017). Lingue, codici, rappresentanza. Margini delle Digital Humanities. In *Filologia Digitale: Problemi e Prospettive. Accademia Nazionale dei Lincei, Anno CDXIV—2017*. Contributi del Centro Linceo Interdisciplinare "Beniamino Segre", Vol. **135**. Rome: Bardi Edizioni, pp. 113–41.

Fitzpatrick, K. (2012). The humanities, done digitally. In Gold, M. K. (ed.), *Debates in the Digital Humanities*. Minneapolis: University of Minnesota Press, pp. 12–5.

Flanders, J. and Jannidis, F. (2019). Data modeling in a digital humanities context. In Flanders, J. and Jannidis, F. (eds), *The Shape of Data in Digital Humanities. Modeling Texts and Text-Based Resources*. London: Routledge, pp. 3–25.

Galina Russell, I. (2013). Las humanidades digitales globales. *Humanidades Digitales*. http://humanidadesdigitales.net/blog/2013/11/08/las-humanidades-digitales-globales/.

Galina Russell, I. (2014). Geographical and linguistic diversity in the digital humanities. *Literary and Linguistic Computing*, **29**(3): 307–16.

Gold, M. K. (2012). Introduction. In Gold, M. K. (ed.), *Debates in the Digital Humanities*. Minneapolis: University of Minnesota Press, pp. ix–xvi.

Goldberg, Y. and Elhadad, M. (2013). Word segmentation, unknown-word resolution, and morphological agreement in a Hebrew parsing system. *Computational Linguistics*, **39** (1): 121–60.

Grimmer, J. and Stewart, B M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, **21**(3): 267–97.

Haugh, M., and Schweinberger, M. *Australian Text Analytics Platform* [Online]. https://www.atap.edu.au/text_analysis/

Hengchen, S., Ros, R., Marjanen, J., and Tolonen, M. (2021). A data-driven approach to studying changing vocabularies in historical newspaper collections. *Digital Scholarship in the Humanities*, **36**: ii109–26.

Hill, M. J. and Hengchen, S. (2019). Quantifying the impact of dirty OCR on historical text analysis: Eighteenth Century Collections Online as a case study. *Digital Scholarship in the Humanities*, **34**(4): 825–43.

Horvath, A. (2021). Enhancing language inclusivity in digital humanities: Towards sensitivity and multilingualism. *Modern Languages Open*, **1**(26): 1–21.

Jansson, F., Parkvall, M., and Strimling, P. (2015) Modeling the evolution of creoles. *Language Dynamics and Change*, **5** (1): 1–51.

Lent, H., Bugliarello, E., de Lhoneux, M., Qiu, C., and Søgaard, A. (2021). On language models for creoles. In *Proceedings of the 25th Conference on Computational Natural Language Learning (CoNLL)*, November 10–11, Association for Computational Linguistics, pp. 58–71.

Mahony, S. (2018). Cultural diversity and the digital humanities. *Fudan Journal of Humanities and Social Sciences*, **11**: 371–88.

Maier, D., Baden, C., Stoltenberg, D., De Vries-Kedem, M., and Waldherr, A. (2022). Machine translation vs. multilingual dictionaries assessing two strategies for the topic modeling of multilingual text collections. *Communication Methods and Measures*, **16**(1): 19–38.

Marciniak, D. (2016). Computational text analysis: Thoughts on the contingencies of an evolving method. *Big Data and Society*, July–December: 1–5.

McGillivray, B., Poibeau, T., and Fabo, P. R. (2020). Digital humanities and natural language processing: "Je t'aime … Moi non plus". *Digital Humanities Quarterly*, **14**(2). http://www.digitalhumanities.org/dhq/vol/14/2/000454/000454.html.

Migliorini, B. (1960). *Storia della Lingua Italiana*. Florence: Sansoni.

Negro Romero, M. and Palmou, X. S. (2020). Le Trésor du lexique patrimonial galicien et portugais: un portail pour l'étude du lexique dialectal. *ILCEA [En ligne]*, **39**. https://journals.openedition.org/ilcea/9933.

Nilsson-Fernàndez, P. and Dombrowski, Q. (Forthcoming). Multilingual digital humanities. In O'Sullivan, J. (ed.), *The Bloomsbury Handbook to the Digital Humanities*. London: Bloomsbury.

Nolan, J. (2020). *The Elusive Case of Lingua Franca. Fact and Fiction*. London: Palgrave Macmillan.

Oberholzer, S. and Kunzmann, M. (2017). Geolinguistic documentation of multilingual areas. VerbaAlpina and the challenges of digital humanities (DH). In Buchstaller, I. and Siebenhaar, B. (eds), *Language Variation—European Perspectives VI*. Amsterdam: John Benjamins, pp. 199–213.

Operstein, N. (2017a). The orthography of the Dictionnaire de la Langue Franque. *Mediterranean Language Review*, **24**: 1–33.

Operstein, N. (2017b). The syntactic structures of Lingua Franca in the *Dictionnaire de la Langue Franque*. *Italian Journal of Linguistics*, **29**(2): 87–130.

Operstein, N. (2017c). The Spanish component in Lingua Franca. *Language Ecology*, **1**(2): 105–36.

Operstein, N. (2018). Inflection in Lingua Franca: From Haedo's *Topographia* to the *Dictionnaire de la langue franque*. *Morphology*, **28**(2): 145–85.

Operstein, N. (2021). *The Lingua Franca: Contact-Induced Language Change in the Mediterranean*. Cambridge: Cambridge University Press.

Pitman, T. and Taylor, C. (2017). Where's the ML in DH? And where's the DH in ML? The relationship between modern languages and digital humanities, and an argument for a critical DHML. *Digital Humanities Quarterly*, **11**(1). http://www.digitalhumanities.org/dhq/vol/11/1/000287/000287.html.

Pountain, C. J. (2016). Standardization. In Ledgeway, A. and Maiden, M. (eds), *The Oxford Guide to the Romance Languages*. Oxford: Oxford University Press, pp. 634–43.

Priani Saisó, E. (2020). Challenges of not using English as the dominant language in DH international projects. *Disrupting Digital Monolingualism Workshop*. https://languageacts.org/digital-mediations/event/disrupting-digital-monolingualism/ddm-workshop-videos/.

Reber, U. (2019). Overcoming language barriers: Assessing the potential of machine translation and topic modeling for the comparative analysis of multilingual text corpora. *Communication Methods and Measures*, **13**(2): 102–25.

Risam, R. (2017). Other worlds, other DHs: Notes towards a DH accent. *Digital Scholarship in the Humanities*, **32**(2): 377–84.

Rossetti, R. (2002). La Lingua Franca: Una langue méditerranéenne à travers les siècles. *Université de Nantes, 22/23 Avril 2002, Chaire Du Bellay de l'Académie de la Méditerranée*, 1–15.

Sangiacomo, A., Hogenbirk, H., Tanasescu, R., Karaisl, A., and White, N. (2022). Reading in the mist: high-quality optical character recognition based on freely available early modern digitized books. *Digital Scholarship in the Humanities*, 1–13.

Sarma, N., Singh, R. S., and Goswami, D. (2021). SwitchNet: Learning to switch for word-level language identification in code-mixed social media text. *Natural Language Engineering*, **28**(3): 337–59.

Selbach, R. (2017). On a famous lacuna: Lingua Franca the Mediterranean trade pidgin? In Wagner, E.-M., Beinhoff, B., and Outhwaite, B. (eds), *Merchants of Innovation. The Languages of Traders*. Berlin: De Gruyter Mouton, pp. 252–71.

Spence, P. (2021). Las humanidades digitales en 2021. *Alcance*, **10**(25): 370–86.

Spence, P. and Brandao, R. F. (2021). Towards language sensitivity and diversity in the digital humanities. *Digital Studies/le champ numérique (DSCN)*, **11**(1): 1–29. https://www.digitalstudies.org/article/id/8098/.

Spence, P. and Wells, N. (2021). Introduction: Digital modern languages launch issue. *Modern Languages Open*, **1**(17): 1–4.

Suissa, O., Elmalech, A., and Zhitomirsky-Geffet, M. (2022). Text analysis using deep neural networks in digital humanities and information science. *Journal of the Association of Information Science and Technology*, **73**: 268–87.

Tamburelli, M. (2014). Uncovering the 'hidden' multilingualism of Europe: an Italian case study. *Journal of Multilingual and Multicultural Development*, **35**(3): 252–70.

Tamburelli, M. and Brasca, L. (2018). Revisiting the classification of Gallo-Italic: a dialectometric approach. *Digital Scholarship in the Humanities*, **33**(2): 442–55.

Thieberger, N. (2017). What remains to be done—exposing the invisible collections in the other 7,000 languages and why it is a DH enterprise. *Digital Scholarship in the Humanities*, **32** (2): 423–34.

Tommasino, P. M. (2017). Travelling East, writing in Italian. *Philological Encounters*, **2**: 28–51.

Tria, F., Servedio, V. D. P., Sangol Mufwene, S., and Loreto, V. (2015). Modeling the emergence of contact languages. *PLoS ONE*, **10**(4): 1–11.

Vanetik, N. and Litvak, M. (2019). Multilingual text analysis: History, tasks, and challenges. In Litvak, M. and Vanetik, N. (eds), *Multilingual Text Analysis. Challenges, Models, and Approaches*. New Jersey: World Scientific, pp. 1–29.

Venier, F. (2012). *La corrente di Humboldt. Una lettura di "La Lingua Franca" di Hugo Schuchardt*. Rome: Carocci.

Wagner, C. (2020). Challenging research infrastructures from a multilingual DH point of view—impulses from two workshops on non-Latin scripts. *Disrupting Digital Monolingualism Workshop*. https://languageacts.org/digital-mediations/event/disrupting-digital-monolingualism/ddm-workshop-videos/.

Wansbrough, J. E. (1996). *Lingua Franca in the Mediterranean*. London and New York: Routledge.

Wei, L. (2018). *Linguistic Innovation and Change: A Translanguaging View*. Keynote address presented at *Sociolinguistics Symposium* 22, University of Auckland, New Zealand, June 27–30.