13 Digital Approaches to Multilingual Text Analysis

The *Dictionnaire de la langue franque* and Its Morphology as Hybrid Data in the Past

Josh Brown

13.1. Introduction[1]

Text analysis can be seen to mark the birth of digital humanities (DH). A widely conventionalised date fixes its origins to the beginning of 1946, with Roberto Busa's plans for the *Index Thomisticus*—a massive attempt to encode nearly 11 million words of Thomas Aquinas' writings on IBM punch cards (Sula and Hill 2019; also Terras and Nyhan 2016). From these early origins, text analysis has now become a very wide research area, involving knowledge discovery in written text. Before data can be processed in any meaningful way, text must be structured through the application of natural language processing (NLP). Multilingual texts pose a particular problem in this regard, since tools from DH often need to apply analysis that is not related to a linguistic structure of text in one specific language but rather relies on methodologies common to many languages (Vanetik and Litvak 2019, 1). Much of the early focus in text analytics has been on English language material. Since, researchers have become even more adventurous in attempting to analyse corpora that are multilingual in nature. The precise nature of what 'multilingual' and 'text' mean is far from obvious. This chapter attempts to respond to the question: How is it best to proceed when we try to analyse digitally varieties of languages that do not belong to any one linguistic taxonomy?

Some focus has gone into attempts to 'broaden out' DH scholarship across multiple languages in recent years. Terms such as 'multilingual digital humanities', or variations of this syntagm, have begun to appear in various fora, both in scholarly literature (Horvath 2021; Nilsson-Fernàndez and Dombrowski 2022) and on blogs and other online spaces. Others have argued that 'Non-English DH is not a thing' (Dombrowski 2022). Dombrowski shows how power and scholarly communications are more problematic than is usually assumed across a variety of phenomena, including time zones, language of publication, and journals in particular languages. In a wide-ranging address, she usefully frames multilingual DH in terms of broader questions relating to the field, such as 'who holds the keys to power? what languages are call for proposals in?' Many of these issues are 'recursive problems'. Part of the aim of making terms such as these more visible has been to highlight the

DOI: 10.4324/9781003393696-18

geographical and linguistic diversity already present in DH (Galina Russell 2014; Spence 2014; Mahony 2018; Dombrowski 2021).

This chapter responds to the ongoing development of multilingual digital humanities as it relates to multilingual texts. Specifically, it complicates the question of what 'multilingual' and 'text' mean in situations of language contact that render both terms ambiguous. The chapter conceptualises this broader framework to one specific document from the past: the *Dictionnaire de la langue franque* of 1830. This dictionary, written by an anonymous author in Marseille, has been described as containing the most comprehensive and complete lexical entries for a Mediterranean trade language used throughout the early modern period. The chapter then turns to look at standard language models and mixed language data by considering the underlying assumptions of particular stemmers.

13.2. 'Multilingual' Texts in DH

What might the term 'multilingual' mean when we talk about 'multilingual texts'? Given the polysemy of both these terms (multilingual and texts), their applications in situations of language contact may end up rendering them ambiguous and ultimately toothless when applying them to situations in the past.

Plainly, multilingual texts can mean texts written in multiple languages. Historical texts that contain evidence of multilingualism often vary widely, including in terms of the degree of contact between languages, number of languages, linguistic typology, and others. A traditional descriptor has been to characterise the degree of multilingualism at either the inter-sentential or intra-sentential level or both. In other words, we may be dealing with textual data that combines languages at the sentence level or that code-switch languages at the word level (e.g. *sea casa—sea house*). In many cases of language contact, it is difficult (if not impossible) to ascribe a particular 'word' to any one linguistic variety in a categorical way: code-mixing between Italian and English, for example, leads to data being produced such as *fensa* 'fence', with an English nominal stem (*fens-*) but Italian morphological inflection (*-a*). In other cases still, influences may be present at an orthographical level. This is the case in a variant such as *hogy* 'oggi' [today], where *h* represents a purely orthographical convention (i.e. without any phonological value), taken as evidence for a Latinising script (Brown 2015, 689). The use of grapheme *-y* here shows the author's uncertainty in representing word-final vowels in a situation where no standard orthography exists. This uncertainty is one aspect which modern scholars must also confront when attempting to represent digital forms of linguistic heterogeneity. Since words often combine multiple varieties in ways that make it difficult to distinguish one variety from the other, there is no assured way to disambiguate which language a particular word belongs to.

The examples *sea casa*, *fensa*, and *hogy* discussed earlier, with multiple variants discernible within the same 'word', do not belong to any contemporary, standard language. Given that much historical language contact takes place in situations where no standard exists, variants which freely combine elements from two (or more) languages can be said to occupy a code-intermediate space.[2] Part of the

reason for a lack of interest in digital representations of heterogeneous linguistic variants is due to the fact that historical linguistics itself has had a strong tradition of ideology of the standard (Watts 2015). Developing tools such as annotated corpora or applying part-of-speech (POS) tags for historical language poses non-trivial problems even for languages that are historically recognised and which continue to be spoken by modern communities. This is the case with Old Catalan—an extremely low-resource language with rich inflection and frequent homographs. In cases such as these, previous work on other languages such as Middle Welsh or Classical Tibetan can allow for a semi-supervised method of manual tagging and correction. More importantly, they allow one to build up a training set in an incremental way (Meelen and Pujol i Campeny 2021; Horvath et al., this volume).

These examples are just a few ways in which 'words' may combine linguistic variants from different varieties. This mixing can be at the level of the whole language, at different syntactic levels, or even at the root, infix, or suffix (as well as phonological and/or orthographical). Multilingualism in DH is often framed through a lens of diversity, or representativeness of the 'number' of standard languages used in a call-for-papers, during a conference, publications, etc. In its worst case, this diversity is dismissed as tokenistic and can be no more than the provision of translations of particular texts. But multilingualism is also about language itself, the variety of language(s), and the diverse ways such multilingualism can be represented digitally to mirror its human expression. In short, there is so much opportunity for heterogeneity in language, whether they be the Spanishes, the Arabics, the Chineses, the Dutches, and so on. The textual forms in which such multilingualism appears may take on a variety of forms. In short, 'multilingualism' can mean different things. In the same way, discerning what a 'text' is, is not a simple question. The exclusion of minority voices—or in the case discussed later, 'invisible' ones—inevitably perpetuates selection biases. Only recently have scholars begun to attempt digital text analysis using multilingual data. These include born-digital data as well as text in physical formats.

13.3. Multilingual 'Texts' in DH

How is it best to proceed when we try to analyse digitally varieties of languages that do not belong to any one linguistic taxonomy? One issue is that 'not all metadata standards are capable of encoding multilingual content in a sufficient way' (Arnold 2019). The need for computers to be able to assign binary values complicates a vast swathe of linguistic production which cannot easily be interpreted as Italian, Swahili, French, English, etc. This is particularly true for textual materials which ostensibly contain evidence of 'mixed' varieties of languages, such as lingua francas. Different issues arise when dealing with non-English language material in a digital space. To some degree, these challenges come about due to a lack of resources or available digital infrastructure to represent linguistic production in an adequate way. Wagner (2020), for example, identifies several challenges for the use of non-Latin scripts in the digital space. Challenges include limitations of script reproduction (directionality, image-text connection, Unicode as gatekeeper), mapping of

characters, OCR (Optical Character Recognition) with less ground truth to build on, and missing agreed standards for transcription of non-Latin scripts and metadata (see also Horvath et al., this volume, on challenges of non-Latin script tools).

Dombrowski (2020) explicitly addresses the question of preparing non-English texts for computational analysis. She notes right from the start that the methods developed for computational text analysis in English 'tend to be developed with certain assumptions about how "words" work'. While English fits some of these assumptions quite nicely, many languages do not. These include language-specific properties of English orthography (words are separated by a space when written), morphology (minimal inflection), and others. For languages like Arabic and Quechua, as well as historical languages such as Latin and Sanskrit, repetitions of a 'word' may be obscured to algorithms with no understanding of grammar due to variation in number, gender, or case in which that word occurs.[3] Consequently, text modification is necessary before an algorithm can count various forms as the same 'word'. In some cases, the goal 'is to arrive at a different kind of understanding of a text using some form of word frequency analysis'. The question of being able to distinguish what a 'word' is allows for much more interpretable computational results than if one gives the algorithm a form of the text intended for human readers.

One problem which researchers often come face to face with in textual analysis is stemming and lemmatisation. Vanetik and Litvak (2019, 7, 16) offer brief but useful comments regarding the issues involved. They provide an example of stemming performed by Python NLTK (Natural Language Toolkit) using the Porter stemming algorithm, noting that 'most existing and available algorithms for stemming and lemmatisation are strictly language-dependent tasks'. This goes for many projects in DH, which often rely on available models and projects that have been designed according to standard language criteria. While these authors point to online resources for particular stemming algorithms developed for various European languages (such as Snowball), no evaluation of particular models or their implementation is discussed. Researchers face the question of choice. Selecting a particular stemmer (Porter, Snowball, Lancaster, WordNetLemmatiser, etc.) will naturally lead to different results—and different types of results—according to the overall aim of analysis. In any case, the objective of stemming remains the same: the reduction of inflectional forms and derivational morphologies to a common base form. Languages which have a highly inflected morphology may appear in many different types of 'texts' from the past, including trade material exchanged via lingua francas, as discussed later. In this case, we are dealing with an 'invisible' language whose variety remains buried at the bottom of archives.

Each intervention in this form of knowledge recovery can be seen as an additional layer of transformation. These interpretations occur at multiple levels. First, mixed multilingual data is already a transformation as the data are imagined in creative ways. Cross-fertilising two (or more) separate linguistic varieties to produce new data is already one layer of extrapolation. The assumption is that the data will be interpretable by an imagined user-speaker. Second, the written configuration of these forms occurs via processes of abstraction that do not follow any norm,

other than what is imagined as the phone-grapheme mapping in the writer's mind. The question goes to the very heart of what a language model is but also what a language is. At each transformation, semantic and linguistic values may become lost or distorted. Similarly, Viola and Fiscarelli's (2021) project on a repository of twentieth-century newspapers shows how enriching a digital heritage collection ultimately means sacrificing content. Any form of historical linguistic (re)production will also require critical digital literacy to ensure the greatest degree of preservation and knowledge access.

Another useful approach is to consider mixed language varieties in the past precisely as 'data'. In the case of the *Dictionnaire* discussed later, the very existence of the document implies that its author had conceptualised a separate linguistic variety. In other words, 'data need to be imagined as data to exist and function as such, and the imagination of data entails an interpretative base' (Gitelman and Jackson 2013, 3). In this case, the title of the object, its subtitle, and its textual layout, belie the author's assumptions in relaying (writing) a linguistic variety which they had envisioned as being such. Considering hybrid language from the past as data also requires us to look '*under* data to consider their root assumptions'. Put another way, 'data need to be understood as framed and framing' (Gitelman and Jackson 2013, 4). All linguistic production can be considered as data in this regard. In terms of the *Dictionnaire*, the title invented by the author describes the linguistic variety they are recording (*langue franque ou petit mauresque*). The subtitle indicates to future readers its teleological value and intended audience: to aid French persons in Africa (*à l'usage des Français en Afrique*). In the layout, the data are framed as lexical entries, with two columns on one folio representing a one-to-one mapping from French to the Mediterranean Lingua Franca (MLF). When we consider these aspects as 'data', the interpretation of hybrid language from the past is likely to take on new meanings created by their end users, as occurs with other historical matter.

The rest of this chapter turns to issues of stemming in data from one particular document from the past—the *Dictionnaire de la langue franque ou petit mauresque*. After brief background notes by way of introduction, I deal with some questions of morphology and stemmers in preparing the non-English elements for analysis.

13.4. The *Dictionnaire de la langue franque ou petit mauresque*[4]

The *Dictionnaire* was written in 1830 and published in Marseille, France. It has been described as recording the most comprehensive and complete lexical entries for a Mediterranean trade language used sometime around the Middle Ages and throughout the early modern period and known as MLF. Even though it is a 'dictionary', there are no definitions. Rather, it provides French terms in the left-hand column and corresponding terms in 'lingua franca' in the right-hand column. Images of the title page of the *Dictionnaire*, and the first page of recorded forms beginning with the letter A, are provided in Figures 13.1 and 13.2, respectively.

DICTIONNAIRE

DE LA

LANGUE FRANQUE

OU

PETIT MAURESQUE,

SUIVI

DE QUELQUES DIALOGUES FAMILIERS

ET

D'UN VOCABULAIRE DE MOTS ARABES LES PLUS USUELS;

A L'USAGE DES FRANÇAIS EN AFRIQUE.

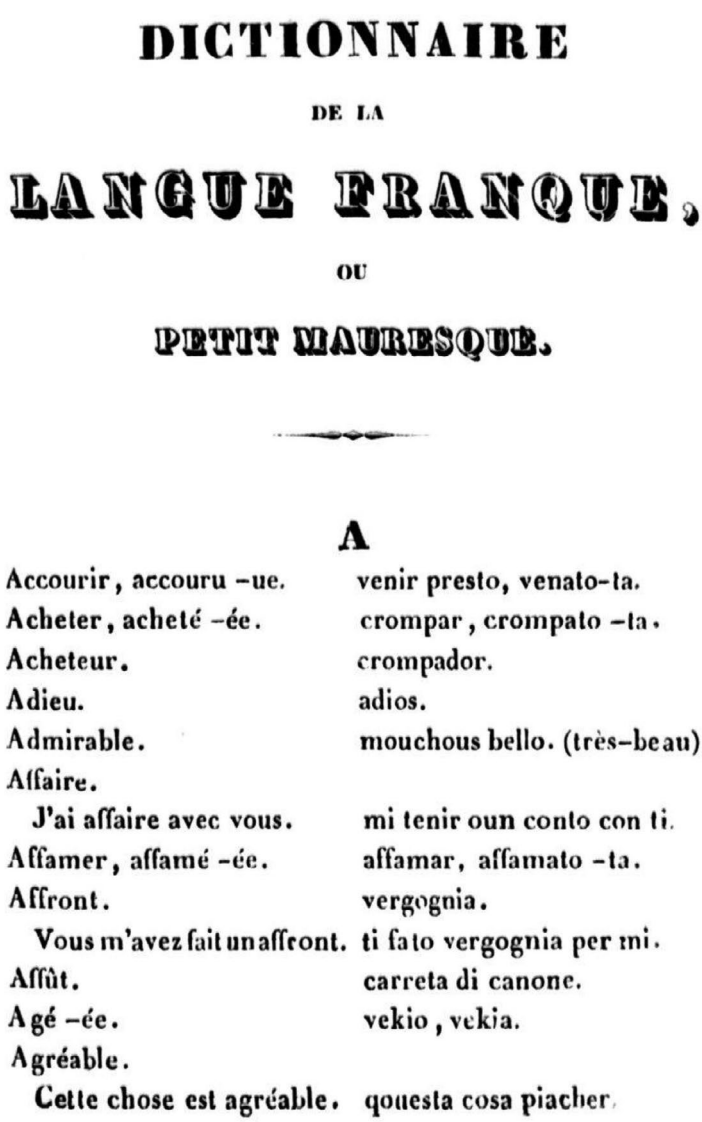
MARSEILLE,

TYPOGRAPHIE DE FEISSAT AINÉ ET DEMONCHY, IMPRIMEURS,
Rue Cannebière, n° 19.

1830.

Figure 13.1 Title page of the 1830 *Dictionnaire*, printed in Marseille. Bibliothèque nationale de France.

Source: https://gallica.bnf.fr/ark:/12148/bpt6k6290361w.texteImage



DICTIONNAIRE

DE LA

LANGUE FRANQUE,

OU

PETIT MAURESQUE.

A

Accourir, accouru -ue.	venir presto, venato-ta.
Acheter, acheté -ée.	crompar, crompato -ta.
Acheteur.	crompador.
Adieu.	adios.
Admirable.	mouchous bello. (très-beau)
Affaire.	
J'ai affaire avec vous.	mi tenir oun conto con ti.
Affamer, affamé -ée.	affamar, affamato -ta.
Affront.	vergognia.
Vous m'avez fait un affront.	ti fato vergognia per mi.
Affût.	carreta di canone.
Agé -ée.	vekio, vekia.
Agréable.	
Cette chose est agréable.	questa cosa piacher.

Figure 13.2 The first page of the *Dictionnaire* showing French and MLF. Bibliothèque nationale de France.

Source: https://gallica.bnf.fr/ark:/12148/bpt6k6290361w.texteImage

Previous analyses have attempted to describe the linguistic make-up of MLF data contained in the *Dictionnaire*, and in other texts taken to contain MLF (Baglioni 2018; Nolan 2020a, 2020b; Operstein 2017, 2018, 2021). Nevertheless, much debate remains about the precise nature of its origins, whether it ever creolised

(Operstein 1998)—therefore providing evidence for a separate linguistic variety—and whether it is in fact a language at all (Brown 2022, 2023). These questions become acutely relevant when attempting to (re)present digital forms of hybrid language in the past. MLF is an extinct trade language, with a Romance lexical base, used around the Mediterranean basin. The *Dictionnaire* has been termed 'the only comprehensive source' [*die einzige umfassende Quelle*] by Schuchardt (quoted in Coates 1971, 25).

Research on MLF is characterised by a general lack of agreement on precisely which data constitute this language variety (Selbach 2017; Brown 2017, 2022). Regardless of whether any particular variant can be said to belong to MLF—or belong to French, Italian, Spanish, etc.—MLF demonstrates a highly inflectional verb and noun morphology (Operstein 2018). In linguistic morphology, stemming is the process of reducing inflected lexemes to their word stem, base, or root. Oftentimes, researchers are interested in the morphological root of the word, in order to see the various types of inflections that a lexeme may display. Stemming is helpful for tagging or parsing purposes, or in order to group particular verb or noun morphologies together when they appear in unstructured texts. The question of stemming, therefore, is particularly useful for being able to identify all word forms which may belong to MLF, not just in the *Dictionnaire* but in other historical corpora too.

13.5. Error Measurement in Stemming Algorithms

Two error measurements exist in stemming algorithms. The first is overstemming; the second is understemming. In overstemming, an error occurs where two separate inflected words are stemmed to the same root but should not have been. This is similar to receiving a 'false positive' in statistical analyses when performing particular types of tests. On the other hand, understemming is a type of error where two separate inflected words should be stemmed to the same root, but they are not—a 'false negative'. The Porter stemmer, for example, stems 'universal', 'university', and 'universe' to 'univers'. This can be seen as a case of overstemming. Although these three words are etymologically related, their modern meanings are so different that treating them as synonyms in a search engine will reduce the relevance of the search results. On the other hand, an example of understemming is 'alumnus' > 'alumnu', 'alumni', 'alumna' > 'alumna'. This word, which has entered into standard English, has kept its Latin morphology, and so these near-synonyms are not conflated. Stemming algorithms attempt to minimise each type of error, often with varying results.

The semi-structured nature of the data in the *Dictionnaire*, and the numerous duplicates it contains, also means that stemming is a particularly useful activity to carry out on such data. 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*, and *écouler*. 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, 224).[5] 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 creolised or not. Identifying a list of stemmed entries allows for easier identification of the lexical variation inherent in MLF data. It also means that such data can be used as a diagnostic to search for evidence of MLF in other data sources. If one eliminates the duplicates in the MLF column from the 2,120 total number of entries, 1,887 unique lexemes remain. The question of deciding which stemmer to apply to such data is not obvious. This is particularly the case for mixed-language data, since models are designed with specific languages in mind. These models are usually based on standard languages. Different stemmers can produce varying results, and their outputs can vary according to the algorithm used, the lexical stock included in their databases, as well as how they deal with lower-case versus upper-case letters, apostrophes, diacritics, and accents, just to name a few.

The stemmers discussed in the next section include two widely used algorithms that are freely available online, Porter Stemmer and Snowball.[6] Before entering into the following discussion, a brief word is needed to introduce each stemmer and contextualise the development of how they were produced. The Porter Stemming Algorithm (or more simply, 'Porter Stemmer') is a process for removing the 'commoner morphological and inflexional endings from words'.[7] The Snowball algorithm is similar to the Porter Stemmer. It is also known as Porter 2. Generally speaking, its algorithm is a better version of Porter Stemmer and is more aggressive. It allows for 'a small string processing language for creating stemming algorithms'.[8] It is one of a handful of stemmers which contains a model for Italian—one of the main donor varieties to MLF (Brown 2022). In the discussion below, focus is placed on the latter of these two stemmers.

13.6. Discussion

Choosing the right model for stemming is crucial in order to achieve meaningful results. The Romance morphologies listed on the Snowball homepage describe the various forms used in creating the model, for both derivational and inflectional morphology. This model was one of the first which attempted to analyse multilingual data. Samples of vocabulary for specific languages are also available. By way of example, the first ten words listed and stemmed for Italian are provided in Table 13.1.

As can be seen, the presence of word-final accented vowels has a drastic impact on the returned stem, leading to cases of understemming. The last two items in the table show the difficulties involved. The penultimate item, *abbandono*, is a first-person present singular verb (whose infinitive is *abbandonare*), while *abbandonò* is a third-person singular 'past historic' (*passato remoto)* verb with the same infinitive—but the stemmer returns different results. Nevertheless, the first four entries in the table, showing reflexes of past participles, which differ according to gender and number, return the correct morphological stem *abbandon-*. Similarly,

Table 13.1 First ten words of Italian vocabulary with stemmed forms generated by Snowball

Word	*Stem*
abbandonata	abbandon
abbandonate	abbandon
abbandonati	abbandon
abbandonato	abbandon
abbandonava	abbandon
abbandonerà	abbandon
abbandoneranno	abbandon
abbandonerò	abbandon
abbandono	abband
abbandonò	abbandon

the underlying model for Porter Stemmer provides output data based on the particular model used to create the different stemmed equivalents. For example, importing NLTK, creating a variable, and running the model on the terms *bat* and *batting* will produce *bat, bat*.

```
import nltkfrom nltk.stem import PorterStemmer
ps = PorterStemmer()
print(ps.stem('bat'))
print(ps.stem('batting'))
```

Running this script on MLF data does not return meaningful results, since these data do not form part of the library. Running the same model on *contentar* and *contento*, or *commandar, commando* does not stem the entries at all—another case of understemming. The same type of result is returned even when the input data is identical except for the presence of a diacritic. This occurs with the corresponding items provided for the French entries *entrepôt* and *dépot*, translated in the *Dictionnaire* with the terms *déposito* and *deposito*, respectively. Running the same script on these entries:

```
print(ps.stem('deposito'))
print(ps.stem('déposito'))
```

similarly provides no stemming to the data. This goes for all lexical roots which may contain stem extenders or present both verb and noun pairings for particular items. A case in point is *escambiar* and *escambio*, for which the term *échange* is provided in the 'French' column in the *Dictionnaire*. The implications of this algorithm for mixed data in Romance, whether historical or not, raises further questions about the typology used for the underlying model for derivational suffixes (called *d-suffix*) in the model.[9] One advantage of Snowball is that it contains a list of suffixes used in the algorithm across different Romance languages. A summary of *d-suffix* typologies for French, Spanish, Portuguese, and Italian is provided in

the documentation for Snowball on its homepage, and they are reproduced in Table 13.2. The first column lists the linguistic category of the particular suffix (one for adverbs, six for adjectives, 13 for nouns, one for verbs), while the second column lists equivalent forms in English.[10]

What is striking about the underlying pattern in this table with respect to MLF data is that the endings only find a tenuous similarity between standard language forms and the actual forms listed in the *Dictionnaire*. In some cases, the model may provide spurious results. This is the case for *-amento*, listed as a Portuguese ending with the corresponding Italian ending, given as *-mente* for nouns—but an entry such as *testamento* is also standard Italian. In some cases, the suffixes in the model are present in only one of the four languages given (e.g., adjective *ique* is only French). In other cases, some items are common to several languages (*ador*, common to both Spanish and Portuguese), while others are common to the last three (*ico*, *oso*, *ente*, and so on). How useful is this stemmer when applied to MLF data in the *Dictionnaire*, and what can we say about the validity of the model underlying the Snowball stemmer in cases of hybrid data?

When it comes to nouns, the model accounts for six suffixes present in MLF. Of these, four are French: *ateur* (1 occurrence, *sacrificateur*); *ousif* (1, but not an adj., French translation is *nègre, esclave*); *ation* (6, *sitouation, sommation, condanation, dispération, fortification, ration*), and *ant* (1, *diamant*). There is only one for Spanish *ancia* (1, *bilancia* 'scale'), and one for Portuguese: *amento* (6, *testamento* 'witness', *armamento* 'armery', *campamento* 'encampment', *désarmamento*

Table 13.2 d-suffixes of Romance language stemmers

		French	*Spanish*	*Portug.*	*Italian*
adverb	LY	(***e***)***ment***	(***a***)***mente***	(***a***)***mente***	(***a***)***mente***
adjective	IC	***ique***	***ico***	***ico***	***ico***
adjective	ABLE	***able***	***able***	***ável***	***abile***
adjective	IBLE	-	***ible***	***ível***	***ibile***
adjective	OUS	***eux***	***oso***	***oso***	***oso***
adjective	ENT	***ent***	***ente***	***ente***	***ente***
adjective	IVE	***if***	***ive***	***ivo***	***ivo***
noun	ANCE	***ance***	***anza***	***eza***	***anza***
noun	ISM	***isme***	***ismo***	***ismo***	***ismo***
noun	IST	***iste***	***ista***	***ista***	***ista***
noun	MENT	***ment***	***amiento***	***amento***	***mente***
noun	ATOR	***ateur***	***ador***	***ador***	***attore***
noun	ATRESS	***atrice***	-	-	***atrice***
noun	ATION	***ation***	***ación***	***ação***	***azione***
noun	LOGY	***logie***	***logía***	***logía***	***logia***
noun	USION	***usion***	***ución***	***ución***	***uzione***
noun	ENCE	***ence***	***encia***	***ência***	***enza***
noun	ANCE	***ance***	***ancia***	***ância***	***anza***
noun	ANT	***ant***	***ante***	***ante***	***ante***
noun	ITY	***ité***	***idad***	***idade***	***ità***
verb	ATE	***at***	***at***	***at***	***at***

'disarmament', *dgiouramento* 'pledge', *ornamento* 'decoration'). There are none for Italian. As mentioned earlier, it is not the case that certain suffixes are mutually exclusive to the language under which they are listed. In other words, the model provides for a framework for mixed language data if and only if the suffixes contained in the data uniquely match those listed under the standard languages listed in Table 13.2. In other cases still, overstemming is present. This goes in particular for the one occurrence of the adverb listed in the model—*(a)mente* (1, *altramente* 'otherwise'). It is also present in the following morphologies:

adjectives

ico (1, *antico, rico*) but also produces nouns *arsénico* 'arsenic', adj/noun *poublico* 'public' and *risico* 'risk'
oso (9 occurrences: *curioso* 'curious', *fourioso* 'furious', *piloso* 'hairy', *prétzioso* 'precious', *sérioso* 'serious', *spacioso* 'spacious', *superstitzioso* 'superstitious', *vénimoso* 'poisonous', *vergognioso* 'shameful') and 1 noun *riposo* 'rest'.
ente (1, *proudente* 'prudent')

nouns

ista (2, *vista* 'vista', *provista* 'supply')
ante (1, *levante* 'levant, east')

In the case of **ante**, the model also produces *brillante* 'bright' which, while listed in the noun class, is an adjective in standard Italian—a clear case of misidentification and overstemming. Similarly, *poublico* can be both adjective and noun. The same goes for *riposo*. What the model fails to distinguish in these cases is not just the linguistic category of the terms being reported, but a one-to-one mapping of the suffixes against these categories leads to spurious results also for standard language data, let alone mixed data. What is needed is a model to distinguish different forms according to different languages with a much broader-based morphology and that is capable of reporting when particular items can be distinguished as actual suffixes in one language or not.

Occurrences such as these arise in other language pairings, especially **anza**, common to both Spanish and Italian (2 occurrences, *spéranza* 'hope', *miscolanza* 'mix'), as well as Portuguese and Italian **ivo** (2, *cativo* 'bad', *vivo* 'alive') but also **ador**, which appears in the class of adjectives for both Spanish and Portuguese, with ten occurrences (10, *balador* 'dancer', *biancador* 'laundryman', *cantador* 'singer', *caschador* 'shoe', *conspirador* 'conspirateur', *comprador* 'buyer', *lavorador* 'worker', *peskador* 'fisherman', *salvador* 'saviour', *segador* 'sawyer'). As I have argued elsewhere, part of the issue in analysing mixed language data using existing infrastructure can be explained by the fact that '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' (Brown 2023). The language

of the MLF presents mixed phenomena that do not easily fall into a pre-established taxonomy. The consequence is that the modelling of particular forms of hybridity, whether contemporary or historical, requires particular tools and further development in a digital space. Table 13.2 contains 48 unique forms. Running the model on the MLF entries from the *Dictionnaire* shows there to be 15 unique suffixes (14 noun; 1 adverb). In other words, even though the model may appear comprehensive when it comes to standard language data and includes a substantial number of suffixes across four of the major standard Romance languages, only 31.25% of these forms are identified in MLF data.

13.7. Conclusion

This chapter has considered two separate but closely related notions in multilingual digital humanities. The first concerns the question of what 'multilingual' and 'text' might mean in situations of language contact that can potentially render both terms ambiguous. These terms are used in a variety of ways to refer to different forms of languages in contact across diverse media. Languages come into contact through a variety of means. The discipline of language contact itself is immediately concerned with the effects of that contact in its own right and as its own field, into circumscribed linguistic categories, into multiple categories, or into none. The task for the researcher is made more difficult when the linguistic data do not fall neatly. In most cases for Western European languages, we are dealing with data that must fit anachronistically into particular glottonyms (proper names of an individual language or language family), since what these 'languages' refer to is often defined at a date after the time when such data are produced.[11] The issue becomes complicated, given that the 'mental habit' of working only with English data and literature leads to a situation of 'monolingual-Anglophone obliviousness with regard to language' (Nilsson-Fernàndez and Dombrowski 2022, 83).

This chapter has also attempted to reveal the problematic nature of using models based on standard language datasets when applied to historically mixed language. Using a historical dictionary as a test-case, the so-called *Dictionnaire de la langue franque*, error measurement can be discerned when searching for the presence of derivational suffixes using one particular stemmer. The question of stemmer selection is in itself problematic for mixed languages. One advantage of the model discussed in this chapter is the seemingly large variety of derivational suffixes it contains for a variety of Romance languages, including French, Spanish, Portuguese, and Italian. Nevertheless, applying that model to MLF data reveals the paucity of suffixes which it is actually able to stem. The standard languages used to create the model also lead to cases of error measurement, particularly overstemming. Errors may occur when suffixes are assigned to multiple languages but also because the model will identify certain reflexes as belonging to a particular linguistic class when they belong to another entirely. Issues arising when working on multilingual data are becoming more and more pressing, as more data become available.

Overall, what is needed is further development of the currently available tools in order to work with hybrid data in the past. A greater rapprochement between the communities of NLP and DH could indeed 'bring methodological advances to NLP, while at the same time confronting DH datasets with powerful state-of-the-art techniques' (McGillivray et al. 2020; see also Gil and Ortega 2016; Wagner 2020). Part of the aim of this chapter has been to answer the question: How is it best to proceed when we try to analyse digitally varieties of languages that do not belong to any one linguistic taxonomy? In short, there is no *best* way. The methodological decisions and justifications that are needed will likely vary according to the precise research question being posed and desideratum of the research being conducted. What one can do is pay particular attention to the issue of multilingual variation, ensuring that parameters are set as openly as possible to take account of such variation. Even with the specific case presented here, we have seen that multiple avenues are possible for applying stemmers to historical data. There is not just one 'word' which pre-established suffixes might appear in; nor is there just one language they can be said to belong to.

The field is slowing developing, as other questions of standardisation, authority files, and character recognition are being brought more out into the open (Arnold 2019). Possible future prospects of research for multilingual DH are wide-ranging. Further avenues are already being explored, beyond simply providing resources in languages other than English. This includes work on postcolonial digital humanities, sentiment analysis, as well as recent initiatives such as *Saving Ukrainian Cultural Heritage Online*, and many, many others. All this work has a focus which is not (just) on English. Some of it is not even in so-called standard language(s), or contemporary languages. More openness on a technical and/or conceptual level is also needed. Knowledge infrastructures housed in libraries, for example, could be enhanced if a greater openness to linguistic variation is incorporated into technical solutions. Another important corollary in answering the question of how best to proceed regards precisely (re)interrogating what the data actually are and how they are digitally (re)presented. In other words, it is important to define precisely what the model is doing and is searching for. Applying one of the available stemmers can be useful in many circumstances. Oftentimes models can be refined even further or developed more in order to provide a greater sense of definition for what the researcher is attempting to discriminate in hybrid data. In large measure, these developments are the result of researchers engaging with more culturally and linguistically diverse data. Such advances promise well for the future of multilingual digital humanities.

Notes

1 The author is grateful to the editors and the anonymous reviewers for generous feedback on this chapter.

2 Matthews (1997, 58) notes that '"code" itself is used by some sociolinguists effectively of any distinct variety of language'.

3 The strong focus on English source material, academic environments, and tools has had far-reaching consequences for DH more generally. I cannot address this issue here

for reasons of space. However, as Cro and Kearns remark (2020, §1): 'It is problematic that DH in the US has focused primarily and nearly exclusively on implementation in English-only environments given both the diversity of languages and experiences in the country itself, of the student populations targeted by DH in American higher ed institutions, and the purported global, public, open perspective indicative of work in DH (Gil and Ortega 2016)'. On the pedagogical issues and question of language faced by these researchers in the implementation of advanced French seminars, see Cro and Kearns (2020, §13).

4 The full title as shown on the first page of this text 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*.

5 Britain (2012, 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'.

6 Some work is already being done to improve stemming for non-English languages, such as the GitHub repository of lemmatization lists in languages other than English, available at: https://github.com/michmech/lemmatization-lists

7 I quote directly from the main homepage of the Porter Stemming Algorithm, available at: https://tartarus.org/martin/PorterStemmer/.

8 https://snowballstem.org/.

9 The term *i-suffix* is used to denote an *inflectional* suffix (e.g. the addition of *-ed* to verbs in English). Since most Romance languages tend to have a highly inflected verb morphology, the verb 'tends to dominate initial thinking about stemming in these languages' (http://snowball.tartarus.org/texts/romance.html).

10 This table has 84 cells with data from four Romance varieties, with three cells left blank (presumably since no corresponding form exists in that variety). Removing the three blank cells from this count leaves 81 forms. Removing duplicates (e.g. *ismo,* which appears as a noun suffix in Spanish, Portuguese, and Italian) across two or more languages, leaves 48 unique forms.

11 These issues are not restricted to these languages, of course. For a recent paper on word segmentation, unknown-word resolution and a morphological parsing system in Hebrew, see Goldberg and Elhadad (2013); also Sarma et al. (2022).

References

Arnold, Matthias. 2019. "Multilingual Research Projects: Challenges for Making Use of Standards, Authority Files, and Character Recognition". *ADHO DH 2019*, Utrecht, The Netherlands. Session Workshop: "Towards Multilingualism in Digital Humanities: Achievements, Failures and Good practices in DH Projects with non-Latin Scripts". Available at: https://zenodo.org/record/5911047#.Ym8o29NBxpQ.

Baglioni, Daniele. 2018. "The Vocabulary of the Algerian Lingua Franca". *Lexicographica* 33, 1, 185–205.

Britain, David. 2012. "Koineization and Cake Baking: Reflections on Methods in Dialect Contact Research". In *Methods in Contemporary Linguistics*, edited by Bernhard Wälchli, Adrian Leemann and Andrea Ender, 219–238. Berlin: De Gruyter Mouton.

Brown, Joshua. 2015. "Testimonianze di una precoce toscanizzazione nelle lettere commerciali del mercante milanese Francesco Tanso (?-1398), Archivio Datini, Prato". *Forum Italicum* 49, 3, 683–714.

Brown, Joshua. 2017. "Multilingual Merchants: The Trade Network of the 14th Century Tuscan Merchant Francesco di Marco Datini?". In *Merchants of Innovation. The*

Languages of Traders, edited by Esther-Miriam Wagner, Bettina Beinhoff and Ben Outhwaite, 235–251. Berlin: De Gruyter Mouton.

Brown, Joshua. 2022. "On the Existence of a Mediterranean Lingua Franca and the Persistence of Language Myths". In *Language Dynamics in the Early Modern Period. Volume 1*, edited by Karen Bennett and Angelo Cattaneo, 169–189. London: Routledge.

Brown, Joshua. 2023 "Whose Language? Whose DH? Towards a Taxonomy of Definitional Elusiveness in the Digital Humanities". *Digital Scholarship in the Humanities* 38, 2, 501–514 Available at: https://academic.oup.com/dsh/article/38/2/501/6827888.

Coates, William A. 1971. "The Lingua Franca". In *Proceedings of the Fifth Annual Kansas Linguistics Conference*, edited by Frances Ingemann, 25–34. Lawrence: University of Kansas.

Cro, Melinda A. and Sara K. Kearns. 2020. "Developing a Process-Oriented, Inclusive Pedagogy: At the Intersection of Digital Humanities, Second Language Acquisition, and New Literacies". *Digital Humanities Quarterly* 14, 1.

Dombrowski, Quinn. 2020. "Preparing Non-English Texts for Computational Analysis". *Modern Languages Open* 45, 1, 1–9.

Dombrowski, Quinn. 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. Available at: www.ucl.ac.uk/digital-humanities/events/2021/apr/ucldh-online-histories-and-futures-linguistic-diversity-dh-rescheduled.

Dombrowski, Quinn. 2022. "Non-English DH Is Not a Thing". Talk presented at online conference *Digital Approaches to Multilingual Text Analysis*. Recording. Available at: https://metodhology.anu.edu.au/index.php/2022/01/20/digital-approaches-to-multilingual-text-analysis/.

Galina Russell, Isabel. 2014. "Geographical and Linguistic Diversity in the Digital Humanities". *Literary and Linguistic Computing* 29, 3, 307–316.

Gil, Alex and Élika Ortega. 2016. "Global Outlooks in Digital Humanities: Multilingual Practices and Minimal Computing". In *Doing Digital Humanities: Practice, Training, Research*, edited by Constance Crompton, Richard J. Lane and Ray Siemens, 22–34. New York: Routledge.

Gitelman, Lisa and Virginia Jackson. 2013. "Introduction". In *"Raw Data" Is an Oxymoron*, edited by Lisa Gitelman, 1–14. Cambridge, MA: The MIT Press.

Goldberg, Yoav and Michael Elhadad. 2013. "Word Segmentation, Unknown-Word Resolution, and Morphological Agreement in a Hebrew Parsing System". *Computational Linguistics* 39, 1, 121–160.

Horvath, Aliz. 2021. "Enhancing Language Inclusivity in Digital Humanities: Towards Sensitivity and Multilingualism". *Modern Languages Open* 1, 26, 1–21.

Mahony, Simon. 2018. "Cultural Diversity and the Digital Humanities". *Fudan Journal of Humanities and Social Sciences* 11, 371–388.

Matthews, Peter. 1997. *Concise Dictionary of Linguistics (Oxford)*. Oxford and New York: Oxford University Press.

McGillivray, Barbara, Thierry Poibeau and Pablo Ruiz Fabo. 2020. "Digital Humanities and Natural Language Processing: 'Je t'aime . . . Moi non plus'". *Digital Humanities Quarterly* 14, 2.

Meelen, Marieke and Afra Pujol i Campeny. 2021. "Old Catalan Morphosyntax: Developing an Annotated Corpus". *Journal of Open Humanities Data* 7, 30.

Nilsson-Fernàndez, Pedro and Quinn Dombrowski. 2022. "Multilingual Digital Humanities". In *The Bloomsbury Handbook to the Digital Humanities*, edited by James O'Sullivan, 83–92. London: Bloomsbury.

Nolan, Joanna. 2020a. *The Elusive Case of Lingua Franca. Fact and Fiction*. London: Palgrave Macmillan.

Nolan, Joanna. 2020b. "Mediterranean Lingua Franca". In *Arabic and Contact-Induced Change*, edited by Christopher Lucas and Stefano Manfredi, 533–550. Berlin: Language Science Press.

Operstein, Natalie. 1998. "Was Lingua Franca Ever Creolized?". *Journal of Pidgin and Creole Languages* 13, 2, 377–380.

Operstein, Natalie. 2017. "The Spanish Component in Lingua Franca". *Language Ecology* 1, 2, 105–136.

Operstein, Natalie. 2018. "Inflection in Lingua Franca: From Haedo's *Topographia* to the *Dictionnaire de la langue franque*". *Morphology* 28, 2, 145–185.

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

Sarma, Neelakshi, Ranbir Sanasam Singh and Diganta Goswami. 2022. "SwitchNet: Learning to Switch for Word-Level Language Identification in Code-Mixed Social Media Text". *Natural Language Engineering* 28, 3, 337–359.

Selbach, Rachel. 2017. "On a Famous Lacuna: Lingua Franca the Mediterranean Trade Pidgin?". In *Merchants of Innovation. The Languages of Traders*, edited by Esther-Miriam Wagner, Bettina Beinhoff and Ben Outhwaite, 252–271. Berlin: De Gruyter Mouton.

Spence, Paul. 2014. "Centros y fronteras: el panorama internacional de las humanidades digitales". In *Humanidades Digitales: desafíos, logros y perspectivas de futuro. Janus,* Anexo 1, edited by Sagrario López Poza and Nieves Pena Sueiro, 37–61. SIELAE: Universidade da Coruña.

Sula, Chris Alen and Heather V. Hill. 2019. "The Early History of Digital Humanities: An Analysis of *Computers and the Humanities (1966–2004)* and *Literary and Linguistic Computing (1986–2004)*". *Digital Scholarship in the Humanities* 34, Supplement 1, 190–206.

Terras, Melissa and Julianne Nyhan. 2016. "Father Busa's Female Punch Card Operatives". In *Debates in DH 2015*, edited by Matthew K. Gold and Lauren F. Klein, 60–65. Minneapolis, MN: University of Minnesota Press.

Vanetik, Natalia and Marina Litvak. 2019. "Multilingual Text Analysis: History, Tasks, and Challenges". In *Multilingual Text Analysis. Challenges, Models, and Approaches*, edited by Marina Litvak and Natalia Vanetik, 1–29. Hackensack, NJ: World Scientific.

Viola, Lorella and Antonio Maria Fiscarelli. 2021. "From Digitised Sources to Digital Data: Behind the Scenes of (Critically) Enriching a Digital Heritage Collection". In *Proceedings of the International Conference Collect and Connect: Archives and Collections in a Digital Age*, edited by Alison Weber, Maarten Heerlien, Eulàlia Gassó Miracle and Katherine Wolstencroft, 51–64. CEUR—Workshop Proceedings, vol. 2810.

Wagner, Cosima. 2020. "Challenging Research Infrastructures from a Multilingual DH Point of View—Impulses from Two Workshops on Non-Latin Scripts". *Disrupting Digital Monolingualism Workshop*. Available at: https://languageacts.org/digital-mediations/event/disrupting-digital-monolingualism/ddm-workshop-videos/.

Watts, Richard. 2015. "Setting the Scene: Letters, Standards and Historical Sociolinguistics". In *Letter Writing and Language Change*, edited by Anita Auer, Daniel Schreier and Richard Watts, 1–13. Cambridge: Cambridge University Press.