THE UNIVERSITY
of EDINBURGH

A Facebook Dialect: orthographical standardisation in Romanised Lebanese-Arabic

Omar Naboulsi

Doctor of Philosophy
University of Edinburgh
2020

Abstract

With the advent of the internet, the new communicative opportunities afforded to millions of its users across the globe have not always come without drawbacks– and in some cases, unexpected advantages. For speakers of colloquial Arabic dialects, such as that of Lebanese colloquial Arabic, the traditional Arabic script used for writing both Classical Arabic and its associated colloquial forms was not available for use in the first programs and applications that enabled digital communication. The resulting adoption of the Roman script has persisted well beyond the availability of the Arabic script for online communication, and is considered a non-standard orthography, used for the writing of a non-standard language, offering its users both constraint (for the representation of sounds for which the Roman script is not suited) and freedom (for the writing of certain colloquial Arabic features of that the Arabic script is not suited, as well as from the generalised constraint of *standard language culture*). This puts the Roman script orthography of Lebanese colloquial Arabic in a unique position, where users do not have a direct *standard reflex* to which to refer or recourse, meaning that unlike non-standard orthographies such as those used to write English dialects, or even creole languages such as Jamaican Creole with a standard lexifier (in this case also English), there is no means by which users can tend towards (or away from) a codified, standardised manner of writing. And yet what emerges is not unbound chaos, but an effective and in many cases expressive writing that generally serves the practical (if not ideological) needs of its users well. Though the QA dialects and in particular their online CMC manifestations have been studied extensively over the past two decades, the opportunity to understand how written conventions form on a grassroots level when there is no *standard reflex* from which users can draw has not yet been taken advantage of. This study adopts a 'mature' understanding of the sociolinguistics of writing and a modern understanding of standardisation as a cultured and imposed paradigm, with which we can consider the non-standard writing of Lebanese colloquial Arabic as it is used in the city of Tripoli in Lebanon not as an orthography that is simply awaiting standardisation (or which can be expected to inevitably standardise), but rather as flexible, dynamic writing well-suited to its use outside of the *standard language culture* paradigm, and yet within which written conventions nevertheless can be observed, and a process of *conventionalisation* and its effects can be detected and described. The city of Tripoli, due to its troubled history, has

a history of Facebook groups initially formed for the discussion of news not otherwise covered by mainstream media, but which have evolved over time to become discussion boards for members of the city, seeing regular Roman script writing and so serving as the first corpus for this study, alongside a series of experimental interviews conducted in Tripoli in 2016 that allow the novel comparison between spoken and written forms in a manner not yet exploited by studies of grassroots conventionalisation, allowing us to ultimately describe this process and produce novel conclusions about how conventionalisation works for non-standard orthographies untethered to a single standard form or the imposed constraints of standard ideology.

Lay Summary

People use all kinds of language to express themselves online, where concern over using 'correct' forms can be less important. For dialects without a *standard* way of writing, digital communication is an ideal space for written self-expression otherwise discouraged in other types of writing. Some languages have come to be written using a different *writing system*, like dialects of Arabic in the Roman script, due to the Arabic script not being available in the early years of the internet. While this also introduces complications, such as how people choose to express sounds that the Roman script has no letters for, it also leads to a new freedom to express native dialects. Within just such a situation, in the city of Tripoli in Lebanon, we explore the emergence of this new kind of writing, which occurs without the *guidance* of established conventions (such as those available in *standard writing*), leading to sometimes chaotic and difficult to read sentences, since each person can *write their speech* in the way they see fit (which we call *transcriptional writing*). We argue that this type of informal and non-standard writing is distinct from other informal writing, like for example the writing of English words as *cuz* and *wot*, as those are always *choices*, for which standard alternatives ("*because*" and "*what*") are available, but which is not the case for Lebanese Arabic in the Roman script. Building on the basis of this freeform writing, we firstly identify how these apparently random forms are actually based on traceable linguistic and social factors, and, most importantly, how it is possible to observe variation becoming gradually *limited* in an organic way, without the guidance of governments and language institutions and other means by which *standardisation* actually occurs. We call this organic development of writing conventions *grassroots conventionalisation,* and contrast it with what is usually labelled *standardisation,* which we argue is a complicated, long-term and ultimately ideological process, rooted in the specific cultural sphere of the modern west. Instead, we strive to understand through our example of the online writing of the Lebanese Arabic of Tripoli how linguistic variation can be resolved, to a certain extent, outside of what we call *standard language culture,* ultimately arguing that non-standard writing of this kind offers many the opportunity of writing within which variation need not be random and overwhelming, but instead can come to be organically arranged in a way that allows for both vernacular expression, the expression of social identity, and at the same time, effective digital communication, without the need to *standardise* the writing or language or to introduce the principles of *standard ideology*.

Table of Contents

Chapter 1: The Sociolinguistics of Arabic

1.1 Introduction

Our study is situated as a study of grassroots conventionalisation within a non-standard orthography as it is used in online communication, following in the paradigm of Hinrichs (2004), Deuber and Hinrichs (2007) and, to an extent, Rajah-Carrim (2008), though it is novel for a number of reasons, the first being the focus on non-standard writing with no *standard reflex,* which its users can neither *converge* upon nor *diverge* from, another being the use of a combination of spoken and written forms of the same words (produced by the same individuals) to enable a concrete discussion of phonetic-graphemic divergence and convergence. To address the question of our thesis, however, we are first required to develop a mature sociolinguistic approach to a number of core concepts. The first of these is the study of Arabic and its dialects, wherein we situate Lebanese colloquial Arabic as it is spoken (and written) in Tripoli within its sociolinguistic context, which work we undertake in this first chapter. The second is the sociolinguistics of writing, for which we adopt Blommaert's (2013) *mature sociolinguistics of writing,* understood not as an autonomous phenomenon but instead as a series of resources available (or otherwise) to any given group of potential users; this will be the focus of Chapter 2. Thereafter, we devote Chapter 3 to developing a modern understanding of the process of standardisation, and what the word *standard* means both academically and in popular perception*,* as well as understanding the historical (and specifically European) roots of standard language culture: a non-universal arrangement of variation, to which alternatives are possible (and exist). We expand on this in Chapter 4 by considering the notion of non-standard *orthographies* specifically and how they are characterised, where we also develop our ideas of *standard reflexes* and a novel approach to categorising non-standard writing. We introduce our final ingredient– the sociolinguistics of Computer-Mediated Communication– in Chapter 5, where we also return to reinterpreting *standardisation* as *conventionalisation* in light of the studies of non-standard CMC writing that we re-term *grassroots conventionalisation* studies, and where we finally reconsider the use of colloquial Arabic in a specifically digital CMC context and review the latest work in this field from the perspective of the deeper understanding of standard language and writing that we developed in previous chapters. We finalise the contextualisation of our study in Chapter 6 with a discussion of the recent history and

sociolinguistic situation of Tripoli, Lebanon, and in that same chapter conduct our preliminary analysis using a pilot dataset we term Dataset 0, before using our full corpus of Facebook group comments in Chapters 7 and 8 to analyse the conventions used in the online writing of Tripolitan Lebanese colloquial Arabic users, where we develop models for understanding how conventionalisation occurs. This is further developed in Chapter 9, where we utilise the written data of Dataset 2 collected by experimental interviews in Tripoli in 2016, and in Chapter 10 we use the spoken data of Dataset 2 to discuss the interplay between the writing and speech of Tripolitan Lebanese colloquial Arabic speakers, allowing us to bring together our theoretical and practical work in our concluding Chapter 11. We begin this process in this chapter, therefore, with an introduction to the most pertinent topics from within the sociolinguistics of Arabic.

1.2 A Brief Review of the Nature and History of Arabic

1.2.1 Origins & Spread

Arabic is classified as a language of the Semitic family, sharing roots with such languages as Hebrew, Phoenician and Akkadian, with a posited common ancestor termed proto-Semitic (Holes, 2004, 10). It is spoken in the Middle East and across the diaspora, with recent estimates putting the number of native speakers at around 250 million (ibid: 1). Though Arabic certainly existed before the revelation of the Qur'an to the Prophet Muhammad (peace be upon him) and the advent of Islam, its attestations before this are limited, and scholars have relied primarily on the body of pre-Islamic oral poetry that survives, and which, as a result of the strict poetic structure of these constructions, is itself problematic for the task of reconstruction a pre-Islamic spoken Arabic (ibid: 10-11). Though some have posited a possible high prestige status for Arabic even before the advent of Islam (Eid, 1990; Ferguson, 1996), what is certainly clear that under Islam, Arabic attained high status, esteem and longevity (Albirini, 2016, 10). The Arabic language swiftly spread alongside Islam in the second half of the seventh century, though it did not spread into a vacuum, but rather to regions with speakers of other, often related languages. Greater Syria, including the region that would become modern-day Lebanon, had been under Byzantine control before the Islamic conquest, and its population primarily spoke variants of Aramaic (a related Semitic language) in addition to Greek among the ruling and mercantile classes (Holes, 2004,

19). The arrival of the Arabic language into the region would therefore create a complicated linguistic landscape, particularly for the task of differentiating dialectical taxonomies for the various vernaculars of Arabic, many of which are posited to have inherited substratal elements (as defined by Thomason & Kaufman, 1988) from these pre-existing spoken languages (see for example Zu'bi, 2019 for Aramaic influences on colloquial Palestinian Arabic dialects).

1.2.2 First Distinction: The Codification of Classical Arabic

Holes (2004, 35) lists four primary factors that contributed to the spread of Arabic in the Arabian peninsula: preconquest contacts between Arabs and inhabitants of the regions later conquered, Islam, (initially the least influential of the factors but which would play a keener role in subsequent generations), urbanisation (with Arabic coming to be the language of the multilingual city), and finally migration and assimilation (though the initial conquest brought relatively few Arabic speakers to new lands, the migrations would later follow would be more influential in scale and effect). What is generally labelled Classical Arabic (CA) emerges from this historical landscape, the codification of its style, structure and form aided by works of great Arab linguists and grammarians such as Sibawayh, Al-Fareedhi and others, who set about codifying the language, using as reference the text of the holy Qur'an, the corpus of pre-Islamic poetry as well as the 'judgment of Bedouins', who were considered in their conservatism– linguistic and otherwise– to be the bearers of the 'authentic' language, a belief which remains prevalent to this day (Albirini, 2016, 11). Both Versteegh (2014, 60-61) and Albirini (2016, 10-12) describe this as a process of *standardisation*, though to better understand to what extent this label is applicable or indeed useful, requires a deeper understanding of the process of standardisation, which will be our primary focus from Chapter 3 onwards.

1.2.3 The Writing of Classical Arabic

Though there is evidence of the writing of pre-Islamic variants of Arabic that would later form Classical Arabic, it was with the emergence of Islam and the drive to eliminate ambiguous readings in light of the Qur'anic manuscript that the writing of Classical Arabic would come to be codified in the form of what we call the Classical Arabic orthography (Versteegh, 2014, 61). Part of this codification was the elimination of the two primary

sources of orthographic ambiguity, the first of which being phonemes that were not distinguished graphemically, such as /s/ and /ʃ/ both written with skeletal form <س>, /r/ and /z/ sharing grapheme <ر> and so on, which Versteegh traces to the roots of Arabic writing in the Nabataean script, which did not distinguish these phonemes as they were not distinct in the Nabataean language, and for which Versteegh suggests the possibility that these phonemes were distinguished even in a pre-Islamic context using diacritic dots, a convention that would be adopted with the codification of the Classical Arabic orthography, in which /s/ and /ʃ/ for example are distinguished as <س> and <ش> (ibid: 63). This is complicated, however, by the second ambiguity, that being the semi-Abjadic nature of Arabic writing, wherein, as with most other Semitic scripts, short vowels are mostly unwritten (Daniels, 2013, 415; Versteegh, 2014, 63). The resolution for this ambiguity came only after the emergence of Islam, and utilised a similar diacritic dot resolution as that adopted to distinguish skeletal letter-forms, and was only later replaced with a system whereby shorthand forms of the long vowels were used diacritically instead of dots for marking vowels, along with other conventions such as the shadda <ّ> as a diacritic marking of gemination, and the hamza <ء> as a marker for the glottal stop (ibid: 63-64). Even so, it took some time for the universal adoption of these conventions, even for their initial use in Qur'anic manuscript writing, and thereafter, as a generalised, codified orthography for the writing of Classical Arabic (ibid: 64). Though the use of diacritic dots for distinguishing letter forms has become standardised, the diacritic marking of vowels remains optional (except primarily within theological writing), and vowels are generally seldom marked except in rare cases where the writer deems the distinction necessary; outside of this, the marking of vowels is generally left to context and the reader (Daniels, 2013).

1.2.4 Second Distinction: Classical & Modern Standard Arabic

Classical Arabic is further distinguished– at least academically– from the modern formal Arabic used in television, newspapers, legislation and other formalised avenues, which is generally labelled Modern Standard Arabic (MSA). This distinction, however, is complex; there is no set point in time at which it can be agreed that Classical Arabic evolved into Modern Standard Arabic, and moreover, present-day speakers of Arabic themselves make no distinction between the two, calling both by the same name (in Arabic, *fus'ha*; Holes,

2004, 5). The distinction between CA and MSA is primarily related to their positions in time, specifically relative to the intrusion of the western world with its culture, philosophy and languages. Both Versteegh (2014, 221) and Albirini (2016, 11) trace this process back to Napoleon's conquest of Egypt in 1798, the traditional date for the encroachment of the western world into the Islamic world. This distinction is therefore not linguistic but concerned with categorisation and a sociolinguistic understanding of how the language is used. The linguistic consequences of this newly emergent power dynamic between west and east were felt in the case of CA first as a result of a heavy wave of translation from European languages (French foremost) into Arabic, leading to borrowings and calques of foreign words and expressions into the formal Arabic language (Albirini, 2016, 11). The process of *westernisation* of CA into MSA is defined by Abdulaziz (1986) in three parts: the modernisation and westernisation of major urban settlements such as Cairo or Baghdad where a western lifestyle was adopted; the work of western-educated figures and the literary and intellectual movements they led; and finally the Arabic language academies that were set up to formalise the language into a modern standard in the European model. In orthographical terms, this also manifested in attempts to replace the Arabic script with a Roman one, such as by the British in Egypt, concurrent with a general push to replace the use of CA with the local colloquial Arabic (QA) variants, such as by the French in Algeria (Versteegh, 2014, 174). These efforts were resisted on anti-colonial grounds informed by pan-Arabist and Islamic ideologies (Albirini, 2016, 14; see also Mejdell, 2006). While Abdulaziz takes a largely positive view of the modernising effect of a benevolent west on the Arab world, Albirini paints a more accurate picture in his description of the survival of the Arabic language in spite of (and not with thanks to) the interjection of European culture (Albirini, 2016, 12). The continuing homogeneity of what is called MSA is a direct result of this popular and intellectual resistance, along with a resistance to grammatical change within MSA (noted initially by Abdulaziz, 1986). In light of this homogeneity, and the lack of distinction between CA and MSA by native speakers, the term Standard Arabic (SA) is generally used to describe both CA and MSA (Albirini, 2016, 3; Holes, 2004, 47), a convention we too adopt henceforth, as well as distinguishing further between SA as a spoken language and SA orthography as the standard writing thereof in the Arabic script.

For Albirini, the very same factors that transformed CA into MSA are also responsible for a new wave of influence in the present internet age, going so far as to anticipate a new form of SA which he terms a post-MSA (Albirini, 2016, 12). In his view, the first encounter between what was Classical Arabic and the colonial west was a physical encounter, and its result, MSA, now itself is undergoing a second, virtual encounter with those same forces of socio-cultural hegemony. While MSA is certainly impacted by the global spread of CMC (Computer-Mediated Communication, including texting, messaging, social media, and so on), it is perhaps more useful to view European influence on Arabic as a singular, ongoing process, dating as far back as 1798 and consistently in effect since. This makes CMC but the latest iteration of the same asymmetrical one-way cultural 'exchange'. MSA was not fully formed after a definable and finite instant of contact with the west, but rather is the result of years of this relationship, which we can perceive as still ongoing, now taking on an added virtual capacity. The internet and CMC certainly do have a unique and unprecedented impact upon speakers of Arabic, though it is not MSA but the vernacular QA dialects that have been most impacted, and in many ways *empowered* by CMC, meaning that perhaps the biggest threat that CMC poses to MSA is its *disempowerment* in favour of the vernacular dialects of QA. In this sense, what colonial authorities at the dawn of the 20th century failed to do is being instigated instead by the development and spread of CMC, where QA– at least in a digital context– is no longer bound to oral communication but through CMC is used in written communication in direct contest with MSA, as seen in studies like Al-Tamimi and Gorgis (2007) and Mimouna (2012) who examine the variable use of MSA and QA online.

1.3 A Third Distinction: Colloquial Dialects of Arabic

1.3.1 The Development and Classification of QA Dialects

Ferguson (1959a) famously argued for a koiné origin for all QA dialects as a result of the 14 features he demonstrates to be shared across all QA dialects, indicating a monogenetic origin for QA dialects. Blau (1981) takes the position that the pre-Islamic linguistic differences (discussed in 1.2.2) did not cease to exist following codification into what came to be CA, and argues that the medieval QA dialects (from which modern QA forms derive) themselves derive, in turn, from the pre-Islamic dialects that were close to CA but did not undergo the same preservational effect of codification and therefore continued to vary into

the medieval and then modern period. A similar position has been strongly adopted by Owens (2006), who takes a historical-comparative approach to reconstructing the common ancestor of the modern QA dialects (in effect, Ferguson's proposed koiné), an approach heavily criticised by Versteegh (2014) on the grounds that it ignores the 'sociocultural circumstances of the acquisition process' by not justifying the transmission of this prototypical form from its pre-Islamic origins to the current spread of QA (Versteegh, 2014, 140). Versteegh is instead a proponent of the polygenetic view for the origin of the QA dialects, characterised by convergences through time and contact with pre-existing languages in the areas into which Arabic was propagated. This is described in Versteegh (1984) as a cycle of pidginization, creolisation and decreolisation, whereby the indigenous people of the conquered lands pidginised the CA of the conquerors, which over generations came to be a creole language passed on from parents as a mother tongue, before eventually usurping the original CA as the primary spoken language in any given region. Holes (2004) concedes the likelihood of the linguistic accommodation that is part of such a process, but insists that such an intricate model as Versteegh's does not make sense for this period, citing the lack of written evidence in any of the literature of the period for the existence of such a pidgin as Versteegh proposes (Holes, 2004, 23). Holes also points out that such a drastic a series of events would have rendered the resulting language unrecognisable (grammatically, phonologically and lexically) from the CA that preceded it, and would have made for a vastly different series of modern QA dialects. For Holes, the written evidence points instead to continuity, citing Hopkins' (1984) analysis of informal written documents dating to 100-200 years after the first contact between Arabic and the peoples of the conquered regions. Hopkins (1984) examines artefacts (chancery documents, personal letters, inventories, bills and so on) and finds both cases of features that no longer exist in modern QA dialects (that were thus lost at a later date), alongside the absence of features also missing in modern QA (which had thus already been lost at the time). Most strikingly, he also finds evidence of the first emergences of features prevalent in and strongly associated with modern QA. In summary, Hopkins finds 'a very impressive continuity in colloquial Arabic usage', concluding that the modern vernacular's roots 'lie very deep' (Hopkins, 1984, xlvi). The evidence, therefore, does not support such a model as Versteegh's, but rather points in the direction of continuity from pre-Islamic to Colloquial Arabic.

Irrespective of precisely in what manner the modern QA dialects came to be, they are defined as consisting of any number of regional, vernacular dialects and sub-dialects, used by speakers in a wide range of primarily informal contexts. These are divergent primarily lexically and phonologically, but also share a wide base of common features (Mitchell and El-Hassan, 1994), including a similar grammatical structure (Soltan, 2007, Aoun et al, 2010). Being uncodified, spoken QA is flexible and dynamic, wherein 'new concepts, expressions and styles can be easily introduced' (Albirini, 2016, 15). In the following chapters, we develop an understanding of this in the specific context of standardisation, and thus understand QA dialects to function as non-standard languages (and their written forms as non-standard orthographies). Speakers of QA are not inducted into the languages via formal education (nor is their native understanding of their QA speech thus modified by formal education) but they are instead learned at home, from family and friends, used in everyday speech and enjoy no official status (ibid: 14). In attempting to arrange the many, often fluid variants of QA, linguists have primarily resorted to non-linguistic taxonomic criteria, based on factors such as geography and ethnicity, and ultimately QA dialects are most commonly associated with the country within which they are spoken, hence Lebanese QA, Syrian QA, Iraqi QA, and so on (ibid: 30-31). This is, however, often problematic, particularly in border regions, where Albirini gives the example of the Syrian QA spoken in Deir-Az-Zour being more closely related to the Iraqi QA dialects just across the border than it is to the rest of the dialects classified as Syrian QA (ibid: 30). This is related to a more general problem of language classification and the way in which we think about language, particularly standard languages, and wherein non-standard language (such as the QA dialects) are still more difficult to delineate outside of a process of standardisation (see 3.3.1).

Given the non-uniformity of QA dialects even within each national classification, there has recently been much discussion with regards to the variation of QA dialects within each of these classifications. Versteegh classifies Lebanese QA within a broader group consisting of both Lebanese and Central Syrian dialects, which includes the QA dialects of Beirut, Tripoli and Damascus (Versteegh, 2014, 198). Within Lebanese QA (henceforth LQA), the prestige urban dialects of the capital cities are spreading at the expense of rural dialects (ibid.), and, to some degree also at the expense of non-capital urban dialects, such as Tripolitan LQA

(see 1.3.3). Moreover, these prestige dialects are not necessarily the traditional dialects spoken in the respective capital, but often take the form of distinct prestige forms. Srage (1997) describes the emergence of a 'constituted urban dialect' that is unlike the traditional sub-dialects of Beirut, just as it is from the dialects of the surrounding region (Srage, 1997, 30). Germanos (2007) understands this in the same terms of *koineisation* that has been proposed to be central to emergence of the QA dialects from CA in the first place (Germanos, 2007, 161). Thus, in addition to the complex arrangement of the typical classification of QA dialects by country, we recognise the still more complex interplay between QA variants within each country, with the case of the prestige form of Beirut of particular pertinence to our study. We discuss further the role of prestige within the sociolinguistics of Arabic in 1.3.3, and continue to contextualise our study in its Lebanese (and Tripolitan) locus in Chapter 6 (6.1). Additionally, following this initial discussion of QA dialects as oral languages, we discuss the writing of QA in a CMC context in Chapter 5 (5.3).

Finally, we consider the attitudes of Arabic speakers towards the use of QA, where, generally speaking, and particularly in an interview or survey-feedback context, attitudes to QA are negative relative to an idealised and potentially ideological preference for the superior form of SA (for example, Al-Muhannadi, 1991, Ennaji, 2007). Ennaji (2007), in a survey of Moroccan QA speakers, receives the familiar feedback that Moroccan QA is seen as a 'corrupt form of [Standard] Arabic" (Ennaji, 2007, Albirini, 2016, 84-85), and the notion of *corruption* is a widely prevalent attitude among speakers of almost all Arabic QA dialects, a view extending even to urban vernaculars in opposition to 'purer' Bedouin forms (Miller, 2007, 4). Albirini highlights some common perceptions of QA as juvenile, 'a language without a grammar', and not even suited to linguistic inquiry, with academic interest in the vernaculars viewed 'with suspicion' (Albirini, 2016, 82). Such attitudes are not surprising considering the prestige associated with SA, owing to its theological, literary, legal governmental and educational superiority for over a millennium, as well as the prestige afforded to written languages that for example QA have not accrued (Albirini, 2017, 36, citing Haugen, 1966 on the prestige of written languages). It does not necessarily follow, however, that these self-declared attitudes reflect the full picture, and we return to this discussion in 1.3.3.

1.3.2 Models of Disglossia

Before we are able to discuss perceptions of prestige or indeed the complex sociolinguistic spread of vernacular variants within the space of any single national QA label (such as Lebanese QA), we first discuss *disglossia,* how it functions in an Arabic sociolinguistic context, and how linguistic attitudes towards this concept have shifted over time. At its core, the discussion of disglossia is a discussion of the relationship between the QA dialects and SA in contemporary Arab societies. Historically, this relationship has been defined as *disglossic* (first applied to Arabic in the landmark paper of Ferguson, 1959b), though the present landscape, while still building upon the initial notion of disglossia, is more complex. Ferguson, writing in 1959, borrowed the term *disglossie* from the work of Marçais (1930, who himself adapted it from Krumbacher, 1902), and applied it to several languages, including Arabic. The essence of Ferguson's argument was that there is a division in society between different forms of language, based primarily on prestige, which results in the co-existence of a duality of variants, each with a definable use, function, context or purpose. Disglossia, for Ferguson, is the division between divergent 'high' (H) and 'low' (L) functions of language-use within a society, where a high form is used in the fields of literature, law, government, media and education, but not in everyday conversation, for which the low form is instead employed (Ferguson, 1959b, 336). In the case of Arabic, Ferguson designates SA as H and QA as L within any given Arabic-speaking speech community. Owens (2011) takes issue not only with Ferguson's assertion that these two forms (H and L) socially distinct, but also that they can carry any inherent linguistic properties that align with their social classification. Ferguson expects the H-form to be more morphologically complex, for example, but for Owens descriptions of such a linguistic nature cannot be conflated with social distinctions: prestige is not inherent in any given system, but is conceptually associated with certain forms and in certain ways by the *users* of these systems, with no bearing on the linguistic structure of any variant (Owens, 2011).

Since Ferguson's landmark paper, a vast body of literature has been produced, much of it dedicated to modifying or further clarifying Ferguson's original model. Blanc (1960) began an enduring trend of expanding Ferguson's H-L duality by introducing a scale of intermediate varieties, ranging from Pure Classical (H), through Modified Classical, Semi-

Literary Colloquial and finally Koineized Colloquial before arriving at L: 'Pure Dialect'. Blanc was followed in this scalar approach by the likes of Bishai (1966), Badawi (1973), Diem (1974), Elgibali (1993), Bassiouney (2006), with Badawi's own breakdown of levels of Arabic into five distinct categories perhaps being the most widely cited of these. While this approach recognises that the linguistic ecosystem of Arabic-speaking societies cannot be defined with a simplified duality of forms (H and L), nevertheless the compartmentalisation of the intermediate forms is problematic (and was initially identified as such by Hawkins, 1983). Owens (2011) summarises succinctly the problematic nature of such approaches: none of these authors provide an empirical framework for defining the categories they propose, with the implication strongly being that such a framework of categorisations is all but impossible to actually achieve. For Owens, beyond SA as H and QA as L, there exist no truly definable categories of speech. Mitchell (1986) is commonly credited with being the first to move away from the sub-categories model, proposing instead a series of *styles,* in which even the shortest utterances can contain a variety of forms, such as in inter-dialectical communication whereby strongly local dialectal features tend to be elided (something also observed by Blanc, 1960 within his sub-categories model), or indeed certain contexts in which features of SA are intermixed with colloquial QA features. Instead of negating the notion of sub-categories, Mitchell instead defines them non-rigidly as interchangeable modes within a body or utterance. Nevertheless, the question of defining the space between the H-form of SA and the L-form of QA remains central to the question of disglossia, and much of the most significant literature concerns itself with addressing this question, in many cases acknowledging the dynamic nature of the relationship between the two forms without attempting to delineate set compartments in an ascending scale. Caton (1991) considers Ferguson's work to be idealised and prescriptive and advocating instead for a descriptive approach to disglossia, such as in the example of distinct speech communities in Yemen, one urban and one rural, which display markedly different ideological approaches to SA, for whom its use alongside QA therefore differs in function. For Albrini (2011) too, though speakers frequently switch between SA and QA, there remain nevertheless distinct functions for each form. According to Albirini, this code-switching allows speakers to encode their utterances in terms of seriousness, complexity, importance and intention. Unlike Fishman (1967), who considered SA and QA as two distinct languages which variously inhabit compartmentalised domains on the basis of contextual social factors, Albirini's

proposal, at its core, defines disglossic variance between H and L forms in terms of *function* (i.e. speaker intention) rather than purely context, which is not precluded, but rather seen to be part of the functional *intention* of a speaker, alongside other factors. In this way, Albirini brings Disglossia more fully in line with present-day sociolinguistic approaches to speakers and speech communities.

To a large extent, what we observe across the span of this literature is a series of definitions and re-definitions of the ways in which SA and QA interface, based on the earliest notions of H and L forms as not segregated and distinct contextual registers, but rather between which there is complex, dynamic interaction that is primarily socially motivated and driven by speaker intentions. To a large extent, all work on this subject– to some extent– consist of modifications to Ferguson's original theory of disglossia, which despite its shortcomings, nevertheless remains at the core for our understanding of how members of a speech community utilise different forms and registers of language. The *in between forms* that are neither H nor L have been approached variously, and alternative terms such as *Polyglossia* and *Contiglossia* have been proposed (Albirini, 2016, 20). Since the work of Mitchell (1986), these approaches have focused defining a *spectrum* rather than insisting on discrete and measurable compartments such as the ones Hawkins (1983) and Owens (2011) take issue with. Among the more recent contributions to the study of disglossia, Auer (2005) posits several unique forms of disglossia using a diachronic approach that considers a changing historical and linguistic landscape. In the case of Europe, Auer demonstrates at various points of time the existence of various different disglossias (Type 0, A, B or C) characterised by what kind of languages occupy the H, L and intermediary roles, and what their relationship to one another is. Bidaoui (2017) applies Auer's framework to Arabic, determining it to be Type C (which Auer calls *diaglossia*). Bidaoui shows how a QA (L) dialect can influence and modify the SA form (H), giving the case of LQA <ya3ni> (/jaɛni/, literally "*it means"*, frequently used as a filler word) as used by guests and commentators appearing on Al-Jazeera television in highly formal contexts where they are otherwise using SA, thus demonstrating an adoption into SA of an initially QA feature. Bidaoui also applies Auer's notion of an intermediary register (resulting in the forms H – M – L) which he contends is a discrete form of speech, and which he defines by demonstrating a single variant form within a single language (Algerian QA) that shows divergence between two variants that Bidaoui

labels Standard (H) and Intermediate (M; Bidaoui, 2017, 70). This singular instance, however, is not nearly evidence enough to posit a rigid, universal M form that can apply across all the manifold forms of QA, and indeed is precisely what Owens (2011) takes issue with when considering the compartmentalisation approach to defining the scale between H and L, and where a singular example from a single QA dialect is unlikely to satisfy Owen's requirements for an empirically-demonstrable M-form. In effect, Bidaoui's three-point classification is essentially a simplification of Badawi's (1973) five-point breakdown of the levels of Arabic, bringing the discussion back full circle to its problematic academic history. More useful, however, is the discrete introduction of non-related languages to the same spread of registers, wherein Bidaoui produces a spread of H – M – L – *English* – *French*. In this, we have an interesting new model that considers languages like English and French as part of a disglossic continuum, a notion (first suggested also by Fishman, 1971) that despite the problematic role of a distinct M form, can still be of particular relevance to our own study of LQA.

1.3.3 Language Prestige & Sociolinguistic Paradigm

Against the backdrop of a deeper understanding of disglossia, we can now further discuss the function of prestige and speech variation within the QA forms. Sallam (1980) introduces the concept of variation across social groups in his study of variance between the utterances of male and female speakers, thus bringing Arabic sociolinguistics in line with mainstream sociolinguistic study (Owens, 2011, 972). This approach is taken up by Abdel-Jawad's (1981) study of linguistic variables across Amman's diverse composition of ethnic-social groups (generalised as Jordanian, Palestinian and Bedouin). The importance of this study lies in its introduction of sociolinguistic variation within communities, in contrast to previous studies which only described variations *between* communities, and has heralded a wealth of literature on intra-communal variation within the Arabic-speaking world (Owens, 2011, 972). Ibrahim (1986) further refines the work of Abdel-Jawad (1981) within the paradigm of the quantitative sociolinguistic approach, defining variable phonemes with complementary roles in terms of prestige, for example, a female, urban prestige marker in /ʔ/ and a complementary /kˁ/ pronunciation more favoured by male speakers and which, in being the SA form, has a prestige of its own. Ibrahim considers these tendencies to be in line with sex

differentiation through linguistic usage across other language communities, where gendered preferences are not only questions of prestige, but also performance of gendered identities of masculinity and femininity. The use of urban features by women are socially-charged acts of feminine identity performance, for which the masculine equivalent is usually the use of either Bedouin or else rural or local QA features (Albirini, 2016, 197-198). Al-Wer (2014) has further suggested that the performance a gender role can be equally viable for both sexes, where in certain contexts women can draw on the prestige associated with masculine forms for assertive or even identity-performing purposes. On the basis of feminine urban prestige form /ʔ/ and masculine SA prestige form /kˁ/, Ibrahim proposes a co-existence of prestige forms that derive their prestige from different sources, both of which mark different *prestiges*. Abdel-Jawad (1987) similarly identifies a case where prestige derives from *locality,* demonstrating a local variant in Nablus, Palestine to be more prestigious than SA equivalents, and where individuals tend towards those local features even when they are not part of their native repertoires. In this way, SA does not hold hegemony over prestige, but rather its prestige is limited (and its hegemony diluted) by sociopragmatic considerations, allowing for a wider spread of prestige forms. For this notion, Ibrahim proposes a continuum of prestige:

- **Capital:** highest-prestige QA variant per nation; Beiruti LQA, Damascene Syrian QA.
- **Urban:** below *capital,* but above other variants.
- **Rural:** less prestigious, speakers assimilate to urban variants when moving to cities.
- **Bedouin:** despite low apparent prestige, a notion persists in the Arab world that Bedouin preserves archaic and thus *true* features of Arabic tongue, thus granting it its own of prestige.

It is clear from both this work as well as our prior discussion of disglossia that we cannot understand prestige as a dichotomy between SA and QA, but as a sophisticated and dynamic contextual scale that functions variously in different social, local and national contexts. Not only are urban QA dialects perceived to be more prestigious than rural ones, but the prestige of the capital dialect is also often seen to be more prestigious than competing urban dialects of other cities, as a result of sociopolitical and socioeconomic factors (Albirini,

2016, 37). Haeri (1991) goes as far as suggesting that capital urban dialects function as *national standards*, and while we will see in our discussions of Chapter 3 and onwards that questions of standard languages and standard functions are complex, we nevertheless can understand the capital dialect to fulfil (along with SA and other forms) some of– though not all– the prestige functions that are served by the *standard* in countries which where a single, standardised national language exists outside of a heavily disglossic context. Bedouin dialects, meanwhile, play dual a function, whereby their low socioeconomic prestige co-exists with the prevalent perception of them being closely connected to the original tribal Classical Arabic (Ferguson, 1968; Hussein & El-Ali, 1989).

Though we see in these discussions a shift towards a wider application of mainstream sociolinguistic theory to the study of Arabic, Owens (2011) identifies various approaches he regards to still be missing from Arabic sociolinguistics. For Owens, the western-centric sociolinguistic approach is based largely on English and western societies in general, and yet there exists a realm of possibilities for the introduction of paradigms exclusive to (or at least vastly more prevalent in) the study of other languages such as Arabic. Owen also raises questions about standards, standardisation and prestige for which Arabic sociolinguistics can assist the integration into a 'cross-cultural' model of study, for example in the unconventional position of an 'unspoken' language like SA as the *standard*, while also sharing standard functions with non-standard colloquial prestige varieties such as urban capital QA. This is in addition to the lack of dialectological studies pointed out by Horesh and Cotter (2016) for various QA dialects, which they see as being gradually addressed in part by collaborative dialectological and sociolinguistic work on heretofore unstudied dialects, whereby descriptive methodologies combine with sociolinguistic analysis. Though our thesis is primarily predicated on the emergence of conventions, nevertheless the combination of dialectological work with sociolinguistic analysis fits into the paradigm proposed by Horesh and Cotter, while our work also addresses Owens' call for an understanding of questions of standard language and standardisation from a uniquely Arabic perspective.

1.3.4 Summary

We have traced the emergence of Arabic from a local language spoken in the heart of the Arabian peninsula to its present global and disglossic status, as well as formulating an understanding of the development of the academic perspectives pertaining to its study. This work provides us with a sociolinguistic basis for understanding the use of language within an Arabic-speaking country such a Lebanon, in particular with regards to the use of a non-standard colloquial variant such as LQA and more particularly still, the urban but non-capital LQA of Tripoli. We will build upon this understanding in the specific context of the use of written LQA in CMC in Chapter 5 (5.3), by which time we will have developed understandings of standard and non-standard language and writing, as well as of the sociolinguistics of writing and CMC writing in particular. We then finalise our contextualisation of our study in the specific grounding of Tripoli, Lebanon in Chapter 6 (6.1), where we apply the paradigms developed in this chapter to our own localised context. We have also developed a basic understanding in this chapter of the Arabic script writing of SA, which will be of relevance throughout the rest of our work, and we now move on to adding a sociolinguistic understanding of literacy more generally, and of orthographies specifically.

Chapter 2: Writing, Literacy, & Orthography

2.1. Sociolinguistic Perspectives of Literacy

2.1.1 Historical Roots of the Study of Literacy

The distinction between spoken and written language has been central within the field of sociolinguistics, even if it took until the second half of the previous century for this distinction to take shape, and therefore for the study of writing and writing systems itself to become a relevant topic of inquiry (Sebba 2009, Coulmas 2013). This has been traced back (for example by Coulmas, 2013, 2-3) to Saussure's emphasis on speech as the natural, organic and innate manifestation of language in opposition to a perceived artificiality of writing which is necessarily learned, within which paradigm Saussure saw writing as having a disruptive effect on its users' perception of language, considering how writing makes language 'visible' to its speakers (ibid: 3; De Saussure, 1916/1978). Joseph (1987, 34-35) calls this 'the awareness of discrete units' of language that is introduced through the visual medium of (specifically alphabetic) writing. The early dismissal of writing as a subject of study, however, belied the fact that writing systems do have a profound effect on members of literate communities, the effect of which could be ignored, but not negated. In the words of Coulmas, in the early years of sociolinguistics 'the baby of writing [was] thrown out with the bath water of its messy effects on linguistic analysis' (Coulmas, 2013, 4). Leonard Bloomfield (1933) added to the Saussurean view the idea that the study of writing inhibits the study of change, as writing is more stable than spoken language, though we note that this is less true in the case of non-standard writing, and further complicated when we consider that even within standardised orthographies, writing itself can even be a vehicle for language change, as discussed by Joseph (1987, 66-67). Some of Saussure's objections to the study of written language followed from his view of writing as a *system*, in opposition to his emphasis on the importance of the 'social fact' underpinning language– an attitude in line with modern sociolinguistic motivations, if not conclusions (Coulmas, 2013, 3). Saussure's misgivings were vindicated in the early sociolinguistic study of writing, within which literacy was indeed perceived to be a system rather than a social practice, with work such as that of Ong (1982) and Olson (1988) focusing on the differentiation of spoken and written language as two discrete functions.

2.1.2 New Literacy Studies: A New Approach

The problem of language and writing and the misgivings of Saussure and Bloomfield would only be addressed with the emergence of the body of work known as New Literacy Studies, wherein writing is seen to interface with both the structure of society and the cultural practice of its members (Coulmas, 2013, 17). This approach began with works such as Scribner and Cole (1981), Heath (1983), Street (1984) and Graff (1987). In particular, Besnier's (1988) study of spoken and written forms of Nukulaelae Tuvaluan extensively demonstrates how written and spoken language cannot be taken as discrete categorical systems but must be approached in terms of social context, specifically through the social contexts of writing, literacy, speech and orality specific to the culture or community in question. The question of speech and writing is thus no longer about inherent differences between the two principal modes of communication, but rather a question of meaning-construction in different contexts, not only in terms of which system is utilised, but in light too of the wider social system within which these communication channels reside, function and are given meaning (Roberts and Street, 1997, 169). Besnier's work dismisses entirely the notional divide between speech and writing and even the possibility of a more refined notional spectrum between the two, pointing out that even the properties most emphatically associated with one could (and do) appear in the other (Besnier, 1988, 731). Thus we come full circle: far from Bloomfield and Saussure's desire to distinguish written and spoken forms of language, preferring the latter for its perceived naturalism and social immediacy, modern sociolinguistic study approaches both writing and oral language as part of an intricate social and communal fabric. The work of Street (1984) demonstrates another aspect of New Literacy by challenging the notion of writing as *autonomous*: a singular, individual and indivisible set of skills that are distinct from socio-cultural factors usually studied by sociolinguists (Sebba, 2007, 13). Street demonstrates through the study of the literacy and illiteracy divide in Iran that literacy is not merely the ability to decode written language into meaning (which many individuals considered to be functionally illiterate were proven capable of doing), but rather it holds a distinct social meaning that supersedes the basic notion of literacy as simply the ability to read, where a division of literate and illiterate is an oversimplification of the effective existence of different *literacies* premised on different social, cultural and ideological backgrounds (Street, 1984, 129-130).

2.1.3 A Mature Sociolinguistics of Writing

Blommaert (2013) calls for a 'mature sociolinguistics of writing', within which it is necessary 'unthink the unproductive distinction between "language and writing"', and instead to distinguish the sub-molecular resources that structure writing, which vary in availability and importance from community to community (Blommaert, 2013, 440-445). Blommaert at once refuses the conflation of speech and writing (as Besnier, 1988) but also adapts the ideas of Street (1984), taking literacy not only to be non-autonomous, but further breaking it up into component parts as *resources* that are inherently social in nature and subject to availability within any given community. Building on a previous landmark study of two instances of what he calls *Grassroots Literacies* (Blommaert 2008; see 5.2.4), Blommaert (2013) identifies what he calls *sub-particles* of writing, beginning with technological and infrastructural resources as the basic materials that allow writing, and graphic resources, which form the visual, 'design' aspects of writing, ranging from the basic ability to form shapes all the way to the fact that the shape of writing can essentially indicate its very genre (indentations indicate poetry, colourful writing indicating publicity, and so on). These sub-particles also include non-literate linguistic resources such as those concerning the language variety used, and in particular the tension between formal and informal language. Notions of standardised correctness are apparent in the pressure to use normative forms that generally accompanies the act of writing, where even instances that call for the use of vernacular forms still come with certain social expectations of correctness. Blommaert's interrelated resources that together form the 'infrastructure' of writing (Blommaert, 2013, 442) are far from being either monolithic or autonomous, but instead represent the development of a 'mature' New Literacy, and a sophisticated sociolinguistic approach towards the once-neglected sociolinguistic study of writing.

2.2 From Literacy to Orthography

The study of *literacy*, being a generalised study of writing, is distinct from that of *orthography,* as the study of discrete writing systems. Thereafter follows our second distinction: Baker (1997, 93) considers a *writing* system to be any graphic representation of language, and an *orthography* to be the writing system as it is used by a specific language. A

writing system, such as the Roman script, when assigned specific graphemic correspondences to the phonemes of a particular language (such as English) is then an orthography, in this case the standard English orthography. The study of both writing systems and orthographies, has until recently, been paid relatively little attention even within the sociolinguistic study of literacy (Sebba, 2007, 12), much in the same way that the study of literacy itself had previously been disregarded, though as with literacy, so too has the study of orthography garnered much more attention in the past decade or so.

2.2.1 Orthography and Cognitive Autonomy

Orthography too was initially considered an *autonomous* system*,* in the same manner that literacy itself was been until this model was challenged by Street (1984), and so too does Sebba challenge the *autonomous* model of orthography from within the same New Literacy paradigm, suggesting that orthography too must be considered within a socio-cultural framework that takes into account the social, cultural and ideological factors affecting orthography (Sebba, 2007, 14-18). The autonomous model for both literacy and orthography has carried colonial, western ethno-centric associations, one example being the work of Goody and Watt (1968), who held the position that a 'Great Divide' exists between literate and non-literate societies, as well as between users of different writing systems (Goody and Watt, 1968, 15-16). Goody and Watt's initial cognitive division between literate and non-literate societies is addressed by Joseph (1987), who notes that while there are certain cognitive changes that take place upon the introduction of alphabetic writing to a community, chief among them the development of an awareness of the 'discrete units' of language and thus the introduction of *language* as a concept into human awareness (Joseph, 1987, 34-35), to call the notion a 'Great Divide' in which non-literate societies are inherently inferior, is in Joseph's words a *cognitive fallacy* (ibid: 40). Joseph points out the circular reasoning that grants inherent value specifically to the modes that developed from within western (or European) culture, including writing, standardised language and the western conception of an education that takes place within such a context. In this way a culturocentric conflation takes place between notions of intellectualisation, modernisation and westernisation that forms a feedback loop, from which such fallacies as the 'Great Divide' emerge (ibid: 34-41).

Goody and Watt also propose a second divide, between phonetic writing systems (most commonly used in the western world), which they take to be superior to and more sophisticated than other systems, such as logographic ones (Goody and Watt, 1968, 37-38). Logographic systems (which map logograms rather than phonemes onto graphemes) are in widespread use, with the Chinese writing system being the most widely-used example thereof. Thus, a major implication of Goody and Watt's conclusion is that a huge number of people in the non-western world are inherently disadvantaged by their 'underdeveloped' writing system in comparison to the phonetic (and overwhelmingly, alphabetic) writing systems used in the west and which Goody and Watt find superior. Sebba (2009, 37) traces Goody and Watt's autonomous approach back to the work of Gelb (1963) and Diringer (1968), who both proposed scales of 'development' of writing systems ranging from pictographic, through logographic, ultimately to phonetic. Sebba also traces a line forwards from Goody and Watt to the more recent work of Hannas (2003), who compares western and East Asian cultural and individual 'characteristics', which he suggests can be traced to their respective writing systems (Hannas, 2003, 6-7; Sebba, 2009, 37). This leads Hannas to posit an inhibition in abstract thought and creativity in East Asian people, as well as tendencies towards political conservatism and what he calls 'group-based behaviour' (though, luckily, manages to stop just short of calling these attributes a built-in herd mentality and abiding love for autocracy). These attitudes are not only problematic in a socio-cultural and political sense, but linguistically, too. Scribner and Cole (1981) dismantle both the cognitive argument that underpins such perspectives as well as offering a sociolinguistic rebuttal in their study of Vai speakers in Liberia, where they find no tangible cognitive effect related to which of the various available writing systems an individual is best acquainted. Instead, all apparent differences are instead explained by the simple factor of education, given that those with good command of the Roman writing system have become acquainted with it through formal education, and thus any cognitive differences derive from the schooling itself, and not the type of script that it equipped them with. Scribner and Cole's work has been significant in dispelling the link between writing systems and cognition, and moreover, concludes that a sociolinguistic approach is most useful, in which the interaction between a writing system, its users and the social and cultural context is the best way to understand the writing systems and how they work. Literacy is not merely

'knowing how to read and write a specific script', so much as it is about 'applying this knowledge for specific purposes, in specific contexts of use' (Scribner and Cole, 1981, 236), recalling both Street's (1984) sophisticated sociolinguistic view of the effects of literacy in his study in Iran and Blommaert's (2013) view of writing as a series of social and technological resources.

2.2.2 Orthography and Linguistic Autonomy

In addition to the view of orthographies as autonomous with regards to their cognitive effects, and like literacy itself prior to thew New Literacy paradigm, orthographies too have been considered purely *linguistic systems*. This approach is reasonably less colonial in its implications, and can be seen in such works as Seifart (2006), who maintains a strictly pragmatic view of orthography, arguing that purely linguistic factors take precedence in the development of orthographies, which should be based on the structure of the language they are writing, rather than conventions of other orthographies, even if those have come to hold social meaning to the speakers of the language in question (Seifart, 2006, 288). Much of the literature of the past decade or so, however, has demonstrated the importance of social and cultural meaning underpinning orthographic choice, following the example of Sebba, who has been producing such literature as far back as the turn of the millennium (for example, Sebba 2000). These include Clifton (2013), Donaldson (2015), Kelly (2018), Lüpke (2018), all of which we discuss in the sections to follow. The concept of autonomy (both cognitive and linguistic) underpins missionary and colonial introductions of orthographies to imperial colonies, with assumptions that certain (phonemic and alphabetic) writing systems are inherently beneficial, and in which some manner of linguistic purity is prioritised over social meanings and realities pertaining to orthography. We highlight therefore the importance (to our own study as well as to the general field) of social meaning in studying orthography, which, like literacy, cannot simply be taken to be an *autonomous* vehicle with cognitive or linguistic value, without careful consideration of the social and political-ideological context of the formation of writing systems.

2.3 The Social Indexing of Orthography

Orthography is socially indexed by choice, whether on an individual, small-group or institutional basis, primarily the choice of which letters (graphemes) are used to signify sounds (phonemes). Almost every orthographical choice can be socially charged, whether the choice is made by an individual, community or institution. A useful way to frame the social pressures on orthographic choice is through the notions of *distance* and *closeness* (Sebba, 2007, 109), where orthographic *distance* is born of the need to distinguish the orthography in question from its neighbours or from a former colonial language, whereas orthographic *closeness* is a means of signalling community and inclusivity through the mirroring of a particular orthography. The interplay between these notions is central to understanding the role of orthographies in social indexing.

2.3.1 Decisions in the Establishment of Orthographies

We can observe the effect of *distance* historically, such as in the case of Polish orthography which was distinguished from Czech conventions by adopting digraphs in place of diacritics (Rothstein, 1977, 225), just as Lithuanian would then adopt Czech conventions to distinguish itself, in turn, from Polish. Another example given by Sebba (2009, 42) is Faroese, which was consciously modelled to be as distinct from Danish orthography as possible (Lindqvist, 2003). Though these are top-down, institutional decisions, they still inform our understanding of socio-politically (rather than linguistically) motivated orthographical decisions. The same duality of *distance* and *closeness* distinction also exists in more recent examples of orthographical development, such as in Clifton (2013) who relates how speakers of Vanimo (a community of 2,700 people spread across three villages on the northern coast of Papua New Guinea) requested the advice of Clifton and his wife in order to develop an orthography. The two local dialects, Vanimo and Waromo, are distinguished primarily by the former using /h/ and /g/, which the latter language conflates to a single sound /ʔ/. The desire for a shared orthography with the least amount of division led to the committee's choice of <g> and <h> for Vanimo, and a digraph <gh> for the Waromo phoneme /ʔ/, leading to orthographies with high graphic similarity, thus allowing orthography to index similarity, or *closeness* (Clifton, 2013, 2). Clifton also gives examples of where distance is preferred, such as the Tanchangya and Chakma languages, both spoken in

Bangladesh by 150,000 and 21,600 people respectively, and considered closely related to the degree that a common literature was proposed for both by Maggard et al (2007). In spite of this, what Clifton calls an 'ambivalent relationship' between the two communities is also reflected in their orthographical choices: despite the prior adoption by Chakma speakers of a Burmese-based orthography, which speakers of Tanchangya speakers could have easily adopted themselves, they instead pointedly developed their own script, heavily based on the Chakma one but with characters differing in many cases only in orientation (Clifton, 2013, 6)- distinguished in most cases for the sake of distinction, a heavily social and ideological rather than linguistic decision.

2.3.2 Variation in Established Orthographies

In addition to orthographic decisions at what Sebba calls the *development* stage of an orthography (Sebba, 2007, 41), where a standardised orthography is institutionally developed, we also observe the same socially-motivated pressures as an ongoing, variable process both in standard as well as non-standard orthographies. Separatist ideologies, for example, can be expressed through orthographical choice, even if these choices do not result in a fully standardised form of writing. Speakers of Galician, for which there is no standardised set of writing conventions, choose between Spanish or Portuguese orthographical features, and in doing so, mark both their identity as well as their perception of what Galician *is:* Portuguese writing conventions (such as <ç>, <õ> or <ã>) reflect a belief that Galician is a variant of Portuguese, whereas use of alternatives derived from Spanish writing conventions marks a belief in Galician as an independent language separate to Portuguese (Álvarez-Cáccamo and Herrero Valeiro, 1996, 148–149; Sebba, 2007, 40). Here it is not a single, synchronic moment of choice but a range of socially-indexed variants which co-exist, and within which variation even the simplest (diachronous) orthographical choice can in fact be a socially meaningful act of allegiance. Another example is Basque, whose speakers generally desire to distance Basque from Spanish, and do so through the use of distinctly non-Spanish features such as <k>, <ts> and <tx> (Álvarez-Cáccamo & Herrero Valeiro, 1996, 149; Sebba, 2007, 40). Unlike Galician, whose speakers use Spanish features to mark distance from Portuguese, speakers of Basque, being under Spanish dominance, use distinctly non-Spanish features to mark distance from Spanish, demonstrating the

contextual complexity of social and political identity. Orthographical choices on both an institutional and individual follow largely the same pattern with regards to distance and closeness, the primary difference being whether the decision is one-time and synchronous (associated with codification and standardisation), or an ongoing, diachronous decision (associated with non-standard and uncodified orthographies). It is also true, however, that some standardised orthographies allow for variation, such as the Dutch orthography that allows for what Sebba calls *licensed variation,* where variant spellings of certain words are considered correct within the standard, such as <kommunikasie>, <kommunikatie> and <communcatie> (Sebba, 2007, 38). Seuren (1982) considers his own preferred rendition, <kommunikatie>, to mark a mid-point between archaism and modernity, and uses it for this purpose specifically (Seuren, 1982, 77-78). Diachronous orthographic choices are therefore also viable to a certain extent within standard orthographies that allow for licensed variation, allowing for a means of marking social meaning even within the standard.

2.3.3 Post-Colonial Orthographies

The question of orthographic distance and closeness, along with other socially-motivated considerations, is particularly pertinent in the post-colonial context and the establishment of post-colonial orthographies. There is often a motivation to distance the language of a newly-independent country from that of former occupiers, which is performed in part by orthographical means. The Dutch representation of /u/ as <oe> was, for example, rejected for the writing of both Indonesian (Vikør, 1988) and Sranan (Sebba, 2000) following their respective independence from the Netherlands, replaced instead with the form <u> (Sebba, 2009, 40). In Haiti, however, we find a more complex interplay of orthography and identity. For speakers of Haitian creole, French orthographical conventions have become integrated and signal a Haitian identity that, later, was threatened by the attempted introduction of a new orthography that used English conventions instead, such as <w>, <k> and <y> (Schieffelin and Doucet, 1994, 191). For Haitians, the perceived 'Anglo-Saxon letters' formed part of a colonial proposal rooted in US imperialism, and portended both an attempted shift towards the English language itself, and even a shift from Catholicism to Protestantism (ibid: 191; Sebba, 2007, 40). That Catholicism and the French writing system were artefacts of a previous colonial ruler did not preclude them from becoming symbolic of a native Haitian

identity, especially in the face of a new colonial threat, and thus *distancing* from the new imperialist threat of English means, in turn, closeness to the former French colonial writing. Ultimately, the attempted institutional introduction of a new orthography was rejected by the community on a highly socially and politically meaningful basis. We also observe here several analytical paradigms intersecting: there is not a single Haitian orthography, but a selection of established but non-standardised orthographies based on French writing, which are in competition and yet all of which carry a shared social meaning indexed by their use of French features, which come into conflict with the attempted establishment of a new, standard convention based on English writing. Socially meaningful choices therefore take place both where there is a discrete or synchronous point of adoption for a standard orthography on an institutional level, but can also be observed occurring diachronously along the same lines in communities of users of established orthographies, whether or not those orthographies are standardised. Linguistic analysis alone, as in the autonomous model, cannot account for such orthographical decisions.

2.4 The Ideology of Orthography

The discussion of distance and closeness revolves primarily around the view of what a particular orthography *should be*. The ideological question here is: *what is our perceived or desired identity*? We consider this an *internal ideology*, made apparent through the choices of the community (or national institutions) in question. On the other hand, ideological debates that take place outside of the exclusive decision-making sphere of the community are questions of *external ideology*, not concerned with the shape of any specific orthography, but rather what *orthography itself* is*.* We have already touched upon such external ideological debates, such as whether the *type* of writing system an orthography is based on provides cognitive advantages, such as in the works of Hannas (2003) and Goody and Watt (1968) before him, and their phonemicist or structuralist approach to the preference of phonetic over logographic writing. West-centric views and attitudes, however, are not only limited to a discussion of writing systems. Languages in the west, more generally, have been judged by two primary criteria: whether they are written (Goody and Watt's 'Great Divide'), but also whether or not they are *standardised*. Thus, we now consider a more general view of the ideology of writing, particularly with regards to

institutional attempts to formulate or designate standard orthographies in the context of the perceived prestige of writing, and, particularly, *writing in a standardised way*. We thus further extend our discussion of colonialism in light of ideological considerations (2.4.1), before developing this discourse towards the question of standardisation (2.4.2), leading us into our expanded discussion of standardisation in Chapter 3 to follow.

2.4.1 The Tyranny of Colonialism

According to Donaldson (2015), the colonial western attitude towards an inherent superiority of written languages becomes intertwined in the post-colonial era with a western tendency towards developmentalism, culminating in a desire to equip the unwritten languages of Africa and Asia with the perceived boons of writing. This initiative, however, often fails completely, such as in the example of the Manding languages that Donaldson discusses, or indeed in case of the multitude of failed attempts to introduce orthographical standardisation to the entire region of West Africa discussed by Lüpke (2018), where the proposed standard orthographies are ultimately 'barely used' because of the purely linguistic basis of their construction that ignored the social realities of the region (Lüpke, 2018, 129). The Manding languages, being the languages of Guinea (Maninka) Mali (Bamanan) Burkina Faso (Jula) are mutually intelligible when spoken but have multiple orthographies so that written communication between these groups is not possible when different scripts are utilised. Even in Mali alone there are multiple writing systems in place for the writing of Bamanan (Roman script, an IPA-based script and the French writing system, with the additional option in the latter of using or not using diacritics). This poses a great challenge to any organisation seeking to provide the language with a standardised orthography (Donaldson, 2015, 1-2).

The very question of choosing an orthography, according to Eira (1998, 174), can be broken up into different 'discourses': scientific, political, religious, technological, historical or pedagogical. Eira demonstrates how conflicting approaches that depend on different discourses can clash, given that each approach makes sense from the perspective of a single discourse, but much less from another (Eira, 1998, 171). Pedagogy in particular has been a central discourse for the adoption of a standardised orthography for languages without one,

with the argument being that learnability of the orthography is crucial, not only for the sake of learning to write the language itself, but also because education has long been considered more effective when it takes place in one's mother tongue (Thomas and Collier 2002). This recalls our discussion of the autonomous perspectives of orthography, whereby writing is considered a systemic vehicle for cognitive benefits, belying a culturocentric approach that considers *education* to be synonymous with the western conception of an education system. A new perspective prioritising the appreciation for the social and cultural benefits of preserving the same West African linguistic and orthographic diversity that Donaldson discusses (rather than stifling it with western concepts of language and education) is more fiercely championed by Lüpke (2018), whose work we discuss in 2.4.2 to follow. Donaldson, though he does not go so far as Lüpke, nevertheless does not accept either pedagogy nor Eira's other discourses as sufficient alone for the development or selection of an orthography in the case of the Manding languages. For him, the social practice already in place must, above all, take precedence. In this we see again echoes of the conflict between the structuralist perspective of orthographical autonomy, versus the sociolinguistic approach that considers the socio-cultural reality to be of the greatest significance. Donaldson himself frames his discussion in the context of an ideological tension: a stand-off between what he calls the desire for a 'continent-wide empire of letters' (Donaldson, 2015, 6), in which uniformity across as many languages as possible is desirable in order to connect, through orthography, vast swathes of the African geography (an approach historically desirable especially for colonial officers), a vision opposed to the retention of locally-intimate systems and local social idiosyncrasies tied up with orthography even if it leads to disconnected orthographies across a wider geographical and social area. In short, we understand this as the tension between an imposed (orthographical) standard in opposition to indigenous, sovereign desires and social requirements. Donaldson, however, does not do away with the other discourses proposed by Eira, but rather places the social discourse above all others, giving it precedence over autonomous orthographical perspectives such as how learnable or efficient a proposed orthography might be. This is not dissimilar the proposal of Stebbins (2001), who resolves this debate by simply adding another discourse category to Eira's list, which he calls 'community'. In effect, Donaldson's conclusion is the same: implementing a standardised orthography for languages that lack it must consider, alongside the traditional factors, the socio-ideological factors particular to

the community in question. Lüpke, without moving away from the same West African landscape, goes further.

2.4.2 The Tyranny of Standard Language Culture

Unlike Donaldson, who introduces social indexing to the broader debate of orthographic development without challenging its standard (and standardised) context, Lüpke (2018) makes a compelling argument for the abandonment of the entirety of what he calls *standard language culture* for West African writing (Lüpke, 2018, 1). Lüpke demonstrates, as Donaldson did, how the communities of the West African region employ a rich variety of overlapping languages and dialects, which Lüpke calls language *repertoires,* demonstrating how these are typically written in a number of different scripts, derived from French, Arabic and even a previous failed attempt at creating standardised orthographies for the more widely-used languages, but in spite of these attempts and influences, this orthographical landscape ultimately reverts to mirroring in writing the same diversity present in indigenous oral practices. Lüpke concludes that standardisation and the adoption of standard language culture will be detrimental to the linguistic diversity of the region, preferring instead a symbolic written standard for the sake of representation, but the actual use of the orthographies that developed indigenously (ibid: 28). Such orthographies, though they derive from writing systems introduced to the region externally, have become naturalised for the writing of the spoken variants of the region through not an institutional but communal and organic process. Here there is less significance given to *distance* and *closeness*, but rather, what Lüpke calls *lead-languages* are used as an orthographical basis (for example standard French sound-symbol correspondences), out of which emerge what he calls *language-independent* orthographies, the users of which do not concern themselves with the linguistic or social meaning of the language that provides the orthographical basis, and where non-standard language-independent orthographies are not tied to any single local language, but instead are variously used for different *dialects,* or *repertoires*, as Lüpke labels them. We expect our own work to align more closely with this approach, whereby standard English and French act as lead-languages for the writing of Lebanese QA as written in the Roman script, but seldom indexing social meaning related to the languages of English or French. Both the West African context and our expectations for our own LQA context are

unlike the case of Haitian creole whereby (as discussed in 2.3.3) French orthographical features have come to be indexed within Haitian identity and thus the introduction of English features is perceived as a new colonial threat to identity. In our case, we will find the mapping of sound-system correspondences of English and French to be more pragmatic in nature, and thus more akin to Lüpke's description of lead-language repertoires in the context of West African writing.

We see in Lüpke therefore a new way in which writing systems are adopted and used to great effect to mark social and ethno-linguistic identity in a way that mirrors our examination of social indexing in the previous section, but which also is less intimately tied up to the specific social meaning of colonial and post-colonial ideologies; instead orthography is repurposed to serve the complex fabric of local oral variation through conventionalised but not standardised orthographies. Crucially, Lüpke does not see standardised writing as the inevitable final state of an orthography (Lüpke, 2018, 25), an assumption that is frequently held (sometimes only by implication) in much of the work in the field. Both Donaldson and indeed Clifton (2013) whose work on Southeast Asian languages we reviewed in 2.3.1, though they are both highly aware of the social realities underpinning writing and orthography, and argue for the need to give indigenous social identity priority in arranging an orthography, nevertheless concern themselves with providing a *standardised* orthography for the communities in question. This implementation of a standardised orthography, even if it is based on the non-standard writing already utilised, nevertheless seeks to codify and freeze it in place, placing its speakers firmly within standard language culture. This forms an external ideological debate over the purpose of an orthography and the role of standard language culture, complementary to the internal ideological decisions of distance and closeness discussed in 2.3. In Chapter 3 to follow, we address more fully question of *the standard* and the deep ideological issues that underlie it.

2.5 Conclusions: The Sociolinguistics of Orthography

Following our examination of the sociolinguistics of Arabic in our first chapter, we have now developed an understanding of the sociolinguistic study of writing and orthographies, tracing briefly the history of the development of this study and addressing the most relevant

strands of sociolinguistic debate regarding autonomy, ideology and social indexing within it, in particular in a post-colonial context and in light of the nascent debate about standardisation and how it fits into questions of ideology and colonisation. In keeping with the broad New Literacy paradigm, we view writing and orthographies as social constructs that exist within communities whose change and variation is driven by social factors as well as political and sometimes also ideological ones. Literacy exists not as a monolithic entity but a series of interconnected socially-explicit resources. The historical, culturocentric attitudes assuming the superiority of certain writing systems or indeed standardised (versus non-standard) orthographies, premised on qualitative judgements within western academic literature historically favouring whichever system is used in the west have finally been challenged by recent scholarship with growing frequency, as we see most clearly in the recent collection by Weth and Juffermans (2018) entitled *The Tyranny of Writing*, which contains Lüpke (2018) whose work we have examined, though this thinking has roots not only in the relatively recent work of Milroy (2001) but in fact goes as far back as Joseph (1987), who challenges the very notion of the *standard* as a western and European construct. Within this context, we begin to understand the standardisation of writing in the form of tyranny that can be stifling to areas with rich linguistic and orthographic variation.

In the case of Tripolitan LQA and its written CMC form that will be the focus of our study, we must therefore consider the complex collection of external ideological debates revolving around orthography, the majority of which are tied into westernisation, just as we have to consider the internal ideology of the orthographical decisions of individuals. That the Roman script writing of LQA derives primarily from two standard orthographies (those of English and French) will recall the example in this chapter of the adoption of French orthographical features for users of Haitian, though we expect these lead-languages to result in a situation where orthographical choices are not indexed on the basis of identity-performance rooted in the source of origin of each orthography, but where the various orthographical features map onto the linguistic diversity of the spoken language of the region. This is not to say that the choice of English or French features cannot be indicative of social meaning, but that we expect this to be the exception rather than the rule, at least within our Tripolitan context (and where the Beiruti context is likely to differ significantly in this regard). The question of standardisation within the study of orthographies is of the greatest importance to our work,

as the focus of our study will be to determine whether *standardisation* (or *conventionalisation*) takes place within the CMC writing of the language community of Tripoli. For this reason, we devote the next chapter to developing an understanding of the very notion of *standard* and what it means, before returning to apply this understanding of standardisation specifically to the question of writing and orthographies in Chapter 4 thereafter. We will add to this an understanding of the sociolinguistics of CMC in Chapter 5, and in that context develop our understanding of standardisation into one of *conventionalisation,* allowing us to understand in that chapter the work done so far on grassroots standardisation and to re-contextualise it within a new concept we will term *grassroots conventionalisation*, and which we finally then are able to apply to the Roman script CMC writing of LQA within our own analysis.

Chapter 3: Standard, Non-Standard & Standardisation

3.1 Language Standards and Standard Languages

3.1.1 The Process of Standardisation

Joseph (1987) remains one of the most influential works on standard languages and standardisation, presenting a comprehensive overview of the standardisation of both language and writing and breaking down the factors involved in the standardisation process as well as examining the consequences of standardisation at length. Joseph demonstrates in detail how the notion of the standard is rooted in the specific cultural and civilisational milieu of the modern West (an idea we encountered in the previous chapter in the work of Lüpke, 2018), which Joseph defines not geographically but as a cultural and political entity which exists not only in Europe where it is rooted, nor the primarily Anglophone countries that it has been exported to, but also the multitude of places and cultures upon which it has been imposed (Joseph, 1987, 22), a definition which we too adopt. We also clarify a distinction between a standard language and a standard orthography, which though they are often intimately intertwined, understanding the complex ways in which they interact necessitates understanding them as interrelated but distinct.

The process of standardisation begins with a dialect community in which geo-political factors eventually cause one community or locality to come into a position of power over others (even if it is initially only as a 'first among equals'; ibid: 2). This dialect gradually attains further status and may come to be seen as the definitive dialect for the entirety of its localised region in a process that Joseph terms the *synecdoche*, at which point the dominant dialect is the *dialect proper,* existing in tandem with alternative sub-dialects (ibid.). But, crucially, it is only within a western cultural context and through the distinct process of standardisation that this local nexus of languages develops further into a distinction between a *language* and its *dialects*. While the synecdoche can be expected to occur universally because 'hierarchisation characterises all linguistic behaviour' (ibid: 60), it is only the specific cultural pressure of the process of standardisation that further produces the division of language variants into the framework of (standard) *language* and (non-standard)

dialects (ibid: 2). Where this pressure exists (whether in western cultures or in other cultures influenced by them), the development of the synecdochal variant into a standardised language requires the fulfilment of a series of involved and complex requisites. A body of *language standards*, which emerges once the synecdoche is established (ibid: 7) is one of these, even if these language standards by themselves do not constitute standardisation, and are instead closer to informal *conventions*; Joseph himself further defines *relative standards* (ibid: 163) as the conventions which give a language its form, contrasted with the prescriptive *absolute standard* found in the cultural manifestation of standard language. The process of standardisation transforms the *conventions* or *relative standards* of the synecdochal variant into the prescriptive *absolute standards* of a standardised language. Conversely, this means that it is possible for a non-standard language to have conventions and rules (Joseph's *language standards*) without having to undergo standardisation.

Beyond the synecdoche and the emergence of language standards, the existence and use of a writing system is among the most important pre-requisites for standardisation. Though we define standard language and standard writing independently, the link between the two is nevertheless complex and in some cases recursive. For Joseph, writing, and more specifically *alphabetic* writing first allows for the development of an awareness of language as an independent cultural object (ibid: 35), itself also a prerequisite for standardisation. Joseph also points out that alphabetic writing and standardisation need not necessarily be *synchronic*, but rather it is sufficient for language-consciousness to have emerged in the past through alphabetic writing, without requiring this writing to be persevered and in-use at the point of standardisation, nor is alphabetic writing necessary for the transmission of language-consciousness once it develops (ibid: 65). Beyond its initial role in language-consciousness, writing is also the vehicle that drives crucial elements for the transformation of a synecdoche into a standard, including some of the language functions a standard must fulfil, in particular *codification* and *legislation*, both of which must additionally be accessible to users of the language. The use of a language variant in legislation provides it with a perception of prestige over time (ibid: 61), and functional legislation, moreover, necessitates a writing system to be available to a literate public. Codification, perhaps even more so, is bound to writing, comprising as it does the codification of phonological and

morphophonological structure through orthography, and of syntactic and morphosyntactic structure through an established and prescriptive grammar (ibid: 65). Significantly, this means that the process of phonological codification is generally inextricable from the process of developing an orthographical standard (for Joseph, 'the creation of a uniform orthography', ibid: 65) for the standardising language. Syntactical codification, through grammar (or a number of grammars) and particularly the transmission of grammar, is also bound to the act of writing, and both grammars and dictionaries play a role in the codification process, both of which encode lexical-semantic structure, even if dictionaries alone are not sufficient for the standardisation of spelling, and thus it follows that the creation of a dictionary is not, in and of itself, a mark of a standardised language (ibid: 71-72).

The relationship between writing and standardisation is intimate and multi-layered, though it is quite clear that standardisation cannot not truly take place without writing. For Joseph, an unwritten language can only be considered standardised 'metaphorically' (ibid: 6). It is also clear that the standardisation of writing, to a certain extent at least, takes place concurrently with the standardisation of the language itself. As such it is difficult to imagine a standard language without an accompanying standard orthography, except perhaps non-synchronously, in cases where a standardised language is written with an alternative, non-standard orthography despite the availability of a standard one (see 4.1 for an exploration of these possibilities). However it is also crucial to understand that, just as the synecdoche does not necessarily lead to standardisation, writing itself (even in the most *ideal* case of a conventionalised synecdochal dialect written with an alphabetic writing system) does not inherently lead to standardisation either. Joseph gives historical examples of Ancient China, Egypt and Sumer, all of which had writing systems, hierarchised dialects and other prerequisites of the standard but did not develop into something we can equate with standard languages (ibid: 20). The first key point is that the use of writing within a community does not itself lead to the formation of cultural resources like grammars and dictionaries that, through writing, codify the language and thus develop a standard orthography as well as a standard language. Rather, these are a part of the *ideology* of standardisation, which does not inherently emerge from the act of writing alone. Secondly, there are also further factors that do not pertain directly to writing, and yet which are

significant prerequisites for the standardisation process. One such 'precondition for standardisation' is superpositioning (ibid: 45), defined as the existence of 'two or more languages of significantly different prestige within a single-speech community' (ibid: 48). This usually foreign language initially functions in the role of H (a *high prestige* form), in advance of the not-yet-standard language (which becomes contrasted as *low prestige* L in the meantime) itself becoming standardised. The modelling of the new L language on the H form takes place in the domain of language rather than writing, though as we explore in further detail (in 4.2), the way that new orthographies form operates on a similar basis whereby the new orthography is moulded on the basis of a pre-existing standard orthography, often one used for the writing of a foreign standard language (Sebba, 2007, 58-59), and a similar modelling occurs in the case of non-standard orthographic developments too. In the domain of language, a variety of cultural functions that ascribe prestige to whichever language fulfils them are initially fulfilled by the foreign H-form standard language before they are later transferred to the standardising L-form as it takes on the role of H. These prestigious cultural functions are primarily functions of spoken language, though almost all of them have a written reflex too. While the notion of standardisation exclusively occurring on the basis of a previous standard form may seem counterintuitive, Joseph explains that this is the case because 'the numerous cultural modules which constitute [standardisation] have grown and changed' over time, meaning that the full repertoire of standardisation is now more complex than at previous points in history (Joseph, 1987, 49). This means that, short of any new standard undergoing this centuries-long process individually and in a fraction of the time, the only way to achieve standardisation in its current form is to model it on a language that has already gone through this long-term historical process (ibid: 49-50). This ultimately goes back to Latin as the first standard, which Joseph says, for over a millennium 'was the only language employed in what we would identify a standard-language functions' (ibid: 50). Such a situation still prevailed in 15th and 16th centuries France, with Latin as the H form and French as the L form prior to its standardisation (ibid: 49), a more recent example being Russian in the 18th and 19th centuries, where Russian aristocratic elites were bilingual in French (as the H form) and Russian (as the L), resulting in French serving as the primary model for the standardisation not only of Russian itself but also other Slavic languages (ibid.). This, in turn, is what allows new standard languages to be inter-translatable with

other standard languages, which forms another of Joseph's requirements for standardisation (ibid: 6). Here we recall from our discussion in 1.2.4 the case of Classical Arabic, which had undergone a series of localised codifications akin to that of standardisation and came to serve a majority of the standard language functions in the communities that used it, but only became *standard* in the European model after its transformation into Modern Standard Arabic in the wake of contact with the culture of the European west and concerted effort in the shape of Arabic language academies (Abdulaziz, 1986), specifically taking the form of translations from European languages into SA (Albirini, 2016, 12).

It is not always the case, however, that any given non-standard form is a clear contender for the fulfilment of the requisite functions and therefore of standardisation. Joseph describes the two primary means by which emerges the *primary dialect* for standardisation, the first being *circumstantial emergence* where this primary dialect is essentially the same as the synecdochal variant, emerging as a 'by-product of non-linguistic prestige factors' as the only real viable variant for standardisation, as in the case of English (ibid: 60). On the other hand, *engineered emergence* occurs when there is no one obvious choice of 'prime' variant, often leading to much debate by proponents of particular dialects, including attempts at forcing the use of a favoured variant in some of the prestige functions in order to encourage its adoption as standard (ibid: 61). In this case, a standard is *engineered* rather than emerging *circumstantially*. These two modes of emergence need not be mutually exclusive, and Joseph gives examples of circumstantially emergent variants that nevertheless required engineering to secure their role (such as in standard French and standard Spanish; ibid: 61). *Circumstantial emergence* also occurs in cases where missionaries (or more recently, language planning committees and other institutions) identify a synecdochally emergent prestige-variant within the language community they are invested in, making the process simpler as there is no debate or disagreement about which form to use (ibid: 61). In cases where the synecdoche has not occurred, the missionary (or equivalent) is forced instead to recourse to *engineered emergence,* and must either assemble a standard or choose a language variant to act as the basis of the standardisation process (ibid: 61). Such a choice, in turn, is premised on a tension between distinction and inclusivity, which we discuss in an

orthographical context in 4.3.2, borrowing Joseph's same paired contrasts adapted for our discussion of emergent and engineered *orthographies* which we first delineate in 4.2.1.

3.1.2 Writing Standardisation

Standardisation is a complex, iterated process that has evolved over time and which comes with specific requirements for its fulfilment, not only linguistic but also cultural and ideological, where an emerging standard comes to be associated with qualities such as clarity, richness, order and intellectuality, thus boosting its perceptions of prestige (ibid: 75). The role of writing within this greater standard language culture, as we have seen, is of great importance at several stages of the process, ranging from the introduction of *linguistic awareness* to functions of formality through written communication, newspapers, public documents and legislation. The standard also plays the function of lingua franca, used across potentially mutually unintelligible dialects within a single polity (ibid.), meaning that in the function of education, the standardising language becomes a means of 'retraining' children from the near-fluency they possess of their native dialect to instead speak and write the standard, a significant part of which occurs in tandem with the training of children in standardised writing and the utilisation of standard orthographies in the teaching process itself. Beyond grammatical and orthographical codification, there are also certain functions that occur entirely through writing, such as the function of literature, which for Joseph (ibid: 76) is a 'cultural manifestation by which language ceases to be an impartial means for conveying messages, and becomes a message itself' (and is itself a reflex of the spoken language function of broadcasting; ibid: 78). Scientific and technological writing was historically the last function ceded by the 'archetypal standard language' of Latin (ibid: 79), where Latin even rewrites the emerging standard, reforming it in order for it to be functional within this field, and thereby also adds not only lexical items but affects even its phonology, morphology and syntax (ibid.). The majority of the prestigious cultural functions required of the standard have some role for writing, meaning that writing plays a role in every stage of the process, from the creation of language-awareness, to codification, to fulfilment of a large number of prestige functions. However, while writing is *utilised* for these purposes, it is not the act of writing itself that leads to these developments, but rather they are developments for which writing is largely prerequisite, but which writing itself does

not cause. We further examine the standardisation of writing and the concept of standard orthographies specifically and in greater detail in Chapter 4 to follow; meanwhile we continue in this chapter to examine the general concept of standardisation and its features common to both spoken and written language, building on the conceptual framework of Joseph (1987) that we have developed in this section.

3.2 Unravelling the Standard

3.2.1 Ideological Perspectives: The Power of Prestige

Milroy (2001) explores in depth the biases inherent in linguistic inquiry that are rooted in what he calls *standard language culture* (a term widely adopted since, including by Lüpke, 2018; henceforth referred to as SLC). Milroy makes a distinction between the associations of prestige inherent in this *standard ideology* and what he calls the *linguistic* process of standardisation, which he defines as 'the imposition of uniformity upon a class of objects'– a definition offered as non-ideological (Milroy, 2001, 531). For Milroy, cultural associations like prestige, formality, and carefulness emerge from the standard *ideology* of SLC, whereas a linguistically-standard language is simply a form of language in which variation has been eradicated (or at least greatly limited) and in which uniformity reigns (ibid: 530-531), distinguishing uniformity as a linguistic property of the language, apart from the social qualities such as prestige are attributed to it by its users (ibid: 532). A consequence of Milroy's division is his objection to the conflation of *standard language* and *prestigious language forms* (such as in Labov's, 1990 view of *changes from above*, which treats prestige and standardness largely interchangeably; Milroy, 2001, 533). For Milroy, the linguistic uniformity of a language is not inherently prestigious, but rather, only the standard languages which overlap with, but do not entirely encompass *languages with high uniformity*- are ascribed prestige by their users within SLC. For Milroy, this means neither that uniformity always evokes prestige, nor that the perception of prestige is exclusively derived from *uniform* languages. We have seen examples of this ourselves, such as the case of the non-standard, non-uniform yet still high-prestige urban capital QA variants discussed in 1.3.3, supporting Milroy's proposal that prestige does not systematically correlate with linguistic uniformity, in so far as prestige can also derive from sources other than uniformity. The converse, however– Milroy's claim that linguistic uniformity does not itself invoke

perceptions of prestige– is more problematic. At the heart of this is the question of why we should expect even *linguistic* standardisation (as merely a linguistic and non-ideological *imposition of uniformity*) to take place outside of the cultural context of SLC. As we saw in Joseph (1987), the very process of standardisation that leads to linguistic uniformity (or even leads to the *desirability* for linguistic uniformity) is itself a cultural process, not only in the prestige attached to it, but even in the very impulse to codify language beyond the synecdoche (which is already associated with non-linguistic prestige factors; Joseph, 1987, 60). Given the non-universality and specifically west-centric nature of the standardisation process (a view advocated by Milroy himself; Milroy, 2001, 530-531), a purely linguistically standardised language cannot be expected to occur outside of it in a culture-free linguistic vacuum. This challenges Milroy's proposition that a standard language should not be measured by how much prestige it has (which he takes to be an ideological definition) but instead by how much uniformity it possess (the linguistic definition), as this claim essentially redefines what we understand a 'standard language' to mean. Prestige alone cannot be the measure of whether a language is standard, for reasons we have discussed, but neither is uniformity sufficient by itself as an alternate measure of the *standardness* of a language. Rather, both are part of an over-arching and ideological process, from which the concept of prestige cannot be stripped. While the perception of a language as prestigious is not necessarily an indication of it being standard, a language cannot be considered standard without the perception of prestige. To reverse this definition, as Milroy does, is to redefine standard language entirely outside of both the academic context it is discussed within, as well as the real cultural and social context within which it exists in the real world.

Milroy also distinguishes between an inherent notion of prestige that a language itself is seen to autonomously *possess*, versus a prestige accorded to it by its users (ibid: 532), an important distinction also discussed in the previous chapter (2.2.1 and 2.2.2). Milroy however uses this paradigm to claim that a standardised language 'may, or may not acquire [prestige]' (ibid: 533), based largely on whether a standard language is adopted by the community in which it emerges. In the sociolinguistic approach, it is certainly true that a standardised language which is not in use by any community cannot autonomously hold any kind of prestige, and equally it is entirely possible to imagine cases where a standard language is institutionally *engineered* for a community and then entirely rejected by it.

However, even in such extreme cases, it is difficult to escape the associations of prestige. We have in fact already encountered cases of the non-adoption of formulated standard orthographies in Lüpke's work (2.4.2), and yet in spite of the fact that members of these communities did not adopt the use of these standardised orthographies devised for their languages, they nevertheless see them as 'a source of great pride', even if they themselves do not write at all, or else do so using the flexible non-standard orthographies at their disposal (Lüpke, 2008, 16). This is the case for example for Wolof in Senegal (McLaughlin, 2008) and Bambara in Mali (Canut & Dumestre, 1993). For Lüpke, these standard orthographies are not used not for *communication*, but instead 'identity creation and self-representation' (Lüpke, 2008, 17): such is the ideological power of SLC, and the perception of standard languages and orthographies as prestigious, that even when a uniform writing system is not seen to be useful enough for a language community to adopt, they nevertheless continue to exist as purely ideological constructs, as sources of pride for the communities for which they were devised. This challenges Milroy's idea of prestige being easily separable from standard languages: even when they are rejected and unused, they do not exist as purely *linguistically standardised* forms but even then continue to signal prestige. Prestige cannot be taken to be a variable, but in fact is a constant of SLC, where the very process itself of standardisation produces not only a language which is linguistically standard, but one which members of the culture in which it emerged deem prestigious. This is a view espoused in recent scholarship, such as Gal (2018, whose work we examine in 3.2.2 to follow), who is clear in her view that standardisation is 'best approached as an ideological phenomenon' (Gal, 2018, 222). While Milroy himself argues for a similar view, his separation of linguistic and ideological uniformity remains problematic. To accommodate a redefinition of *standardisation* as a linguistic phenomenon separate to cultural and ideological considerations would be to upend a great deal of how we understand standard language, standardisation and SLC. Instead, we formulate an alternative solution: it is *conventionalisation* instead which imposes some degree of uniformity on a language without necessarily leading to its users attributing prestige to it. Though the label of *imposition* is no longer useful here, conventionalisation remains the best way to imagine the emergence of a degree of uniformity within an orthography without invoking the ideological preconceptions linked to standardisation, and concurrently, without also requiring an ideological justification for why uniformity is being *imposed* in the first place (the answer to

which will always be the role of SLC). Indeed, the fact that some degree of order is likely to arise inherently (and thus without imposition) links directly to Joseph's notion of the synecdoche and of *relative language standards* that we can expect to arise universally (Joseph, 1987, 7). We will build in the next chapters (and throughout the thesis) on this view of conventionalisation as a means of understanding the emergence of *language standards*, as *conventions,* through which (a degree of) *uniformity* can emerge outside of the ideological context of the prestige of SLC. Milroy's (2001) work is nevertheless indispensable to understanding the ways in which both linguists and non-linguists have understood language inherently through the prism of SLC, and we return in 3.2.3 to discussing his demonstration of how the very process of standardisation is itself tied up in a broader ideological background, situated in Europe and closely connected to western modernity.

3.2.2 Unravelling the Standard: A System Among Many

For Joseph, both the synecdoche and linguistic conventions (as *relative language standards*) can be expected to emerge universally (Joseph, 1987, 7) and without standardisation, and so language communities can (and do) develop and function without the existence of or indeed need for SLC. Standard language forms but 'one pattern of development among many possible effective ones' (ibid: 51). Susan Gal (2018) builds on this concept, considering the duality of standard and non-standard language to be but one possible *axis of differentiation* for organising the variation that exists within any given language community (Gal, 2018, 229). Gal demonstrates the ideological nature of the divide between standard and non-standard by giving a number of examples of alternative arrangements of linguistic variation, demonstrating how, despite its global prevalence through classical colonialism and neo-colonial dynamics of prestige, any *universality* of SLC is merely illusory, being but a single, heavily cultured means of arranging variation for which alternatives exist. Moreover, Gal questions the notion of *correctness* in language, which for her is entangled within SLC, outside of which there are very few prescriptivist judgements, giving examples such as Worora of north-western Australia and the Tolowa spoken in California within which correctness does not exist as a linguistic concept (Gal, 2018, 227). Gal thus sees prescriptivism itself as a product of SLC, within which correctness is decreed by specialists such as linguists and teachers, rather than following 'the conventions of use by communities

of speakers' (ibid: 227). This approach, in turn, has its roots in the 'elite European approach', which Gal calls 'ideologically grounded [and] not a natural fact or objective "view from above"' (ibid: 227).

The first of Gal's examples is Bóly, a small rural town in Hungary in which she conducted field study in the 1980s and 1990s, at which time about half the town was still bilingual, speaking both Hungarian and German (ibid: 230). As part of her work there, Gal reconstructed the linguistic reality of the town in the interwar period, in which period only German was spoken– or rather, two forms of German: *Bäuerisch* and *Handwerkisch*: farmer's-language and artisan's-language. These forms did not fit into a model of standard or non-standard, and neither was considered *correct* or *incorrect*, though there were phonological and grammatical differences between them. Instead, these two forms were closely connected to two primary identities: the farmer's-language indexing the *authentic*, *traditional*, *restrained* and *monotonous* attributes associated with farmers, and the artisan's-language, contrastingly perceived as *fancy, innovative, ornamented* and *various*, indexing the artisanal identity (ibid: 230). The split between these two languages was part of a larger set of culture binaries through which these two sub-communities were divided, within which any kind of activity could be performed 'in an artisan way or a farmer way' (ibid.), including eating habits, entertainment, clothing, architecture and so on. Though members of each community saw their own way of life and language as superior, there was no *absolute* value judgement between a high prestige or *correct* code and a lower prestige or *incorrect* code as would exist in a split between standard or non-standard. Each form espoused a different *type* of value, and Gal observed cases where an individual might switch between codes, if it was necessitated (ibid: 231). In this way Gal provides an emphatic example of how variation in a localised community can be organised in a radically different manner to the received notions of standard and non-standard. Another example is offered by Gal through her retrospective analysis of the work of Kuipers (1998) on the community of the Weyewa highlands of Sumba, Indonesia. There, the split is not between two profession-based sub-communal identities, but between two different *functions*: everyday language and ritual language, the latter possessing rich syntactical and phonological features and being *dense* with meaning (Gal, 2018, 231). And yet it is not *standardised* in the manner that we would expect, having no state or institutional support, nor being taught through

schooling; just as importantly, it was not perceived as *more correct* than the language of everyday speech, even if it was perceived to be richer with meaning. As we saw with the German variants of Bóly, here too the dual language forms are closely associated with identity: ritual language is perceived as *angry,* versus a more emotionally neutral everyday language, and the anger of the ritual language is further understood to be a personal quality suited to leadership (ibid.). Again we see another arrangement of language that does not follow the standardisational model of the west: for Gal, it is better understood as another *axis of differentiation*, this time contrasting *activities* rather than *identities,* and these activities also have associations with identity, but not with correctness or prestige in an SLC understanding. Given that the Weyewa registers are arranged by function, use of the ritual function outside of its correct context is unlikely to garner higher prestige but the opposite, echoing to some degree the role of SA within Arabic disglossia discussed in 1.3.2, whereby the use of highly-formal SA instead of colloquial QA in an informal context leads to ridicule, not prestige.

Gal's alternative *axes of differentiation,* being linguistic systems defined by mutually-contrastive sets of registers within a population, can be *normalised* and *conventionalised* and yet still do not create regimes of standardisation (ibid: 229-230), again recalling the distinction between *standards* and *conventions* (Joseph's *relative language standards*), which exist outside of the context of standardisation. For Gal, these normalised conventions allow the formation of alternative axes for communities to arrange the variation that exists among their speakers. Nor should we expect that, given enough time, these *conventions* might somehow come to 'evolve' into *standards*, as had been the prevalent historical view of language, but rather, we actually see that these cultural *arrangements* of the linguistic space within each community (at least at a certain point in history) are alternative developments which run in parallel to the standard versus non-standard dichotomy of the west, and are only replaced by SLC through external imposition. While they exist, they form a cultural and linguistic organisation specific to the cultural reality of their communities, in the same way that the arrangement of standard languages emerged within the cultural sphere of the west. For Gal, standardisation is a conceptualisation closely related to the conceptualisation of modernity itself (a view also held by Joseph, 1987 and Lüpke, 2018): the standard/non-standard distinction is itself modernity enacted within the communicative

domain, and 'standard languages are thus one of the practices that constitute the axis of modernity' (Gal, 2018, 233), just as the binary of farmer's-language versus artisan's-language was one of the practises that defined the cultural axis between farmers and artisans.

Gal applies the same analysis to the western cultural arrangement of language variation as she did for the alternative cultural arrangements discussed above, based on a list of contrastive values that, in their internal opposition, constitute *modernity*. In this way, standard languages can only be understood by considering their reflex: non-standard languages, against which their identity is indexed and their values defined (ibid: 233). For Gal, therefore, standard language indexes ***anonymity*** (versus *authenticity*)*,* ***universality*** (versus *particularity/emplacement*)*,* ***reason*** (versus *emotion*)*,* ***progress*** (versus *tradition*)*,* ***literacy/education*** (versus *orality*)*,* ***centrality*** (versus *periphery*), and finally, ***homogeneity*** (versus *variety*; ibid: 233). Standard languages are anonymous by being generalised constructs, where non-standard languages are rooted in a particular locality; reason and education are values based on the perception of *correctness* that is a major pillar of standard languages, and contrasts an emotional orality that cannot *be* correct (or incorrect). This is also seen in Gal's contrast between *homogeneity* of the standard (Milroy's *imposed uniformity* reimagined as a factor within an ideological binary, a far cry from a non-ideological linguistic process) versus the variety permitted by the non-standard. The role of writing, for Gal, is tied up with *education*: even if a non-standard language is written, it cannot be *correctly written* unless it is standardised (along with its orthography). The very concept of education is rooted in not only writing but standard writing (even if an *oral education* is possible, it is seldom acknowledged in the modernist axis as *true* education). The most striking difference for Gal between the western modernist axis and the other cultural axes she examines prior is that the modernist axis is the only one that exceptionally claims *value* for one side (the standard) that makes it superior to its antithesis (loosely, the non-standard) in claiming both *quality* and *correctness* exceptionally (ibid: 234). Gal's application of this same dual dissection to western SLC serves as a final demonstration of how it is merely another axis of differentiation, a means of structuring variation not superior but analogous to the other examples she presents: even if the end result is a structure of correctness versus incorrectness, this can only be understood within the polar cultural

features and *social meanings* that are indexed by the two types of language that have emerged from within western (and originally European) culture.

3.2.3 Unravelling the Standard: A European Cultural System

Gal redefines *intellectualisation* in the context of standard languages as *modernisation,* both of which to Joseph are essentially synonymous, both being equivalent to westernisation (Joseph, 1987, 39-40). Gal's demonstrations of what non-western arrangements of variation look like reinforces Joseph's view of standard language as 'a specifically Western concept that has been spread by cultural tradition' (ibid: 7) as well as the rejection (such as by Milroy 2001, 539 and Joseph, 1987, 86-87) of the standard language arrangement as a universal inevitability. For Gal, the emergence of nation-states is a primary driving force behind the association of language with identity, and she notes that even European ideologies themselves 'were generally more diverse before the rise of nation-states and standardised national languages' (Gal, 2018, 228), with the post-First World War transition from multi-cultural, multi-national and multi-lingual empires (Habsburg, German and Ottoman) to monolingual nation-states being central to the imposition of the 'ideology of state-centred monolingualism on the region's ethnolinguistic mosaic' (ibid: 225). This is later echoed in Lüpke's concern that the imposition of standard orthographies on the diverse writing systems of West Africa will too force the same state-centred monolingualism, which Lüpke fears will have the same fatal effect on that region's own 'ethnolinguistic mosaic' (Lüpke, 2018, 16; see 2.4.2). As in West Africa, the initial imposition of standard ideology even in parts of Europe was met with resistance in some areas, particularly in secondary cases where it was 'imposed from above' rather than in the communities where it first developed (Milroy, 2001, 542). We can envisage this same process at work repeatedly: first emerging in western Europe, then being imposed on the emerging nation-states of central and eastern Europe, and thereafter upon the rest of the world through colonial endeavours and post-colonial factors.

Milroy's view of standardisation is as a gradual and progressive process, by which a language develops 'over time higher and higher levels of standardisation', equated with 'greater and greater acceptance of the ideology of standardisation' within a culture (ibid:

542), echoing Joseph's description of how 'the numerous cultural modules which constitute [standardisation] have grown and changed' over time (Joseph, 1987, 49). For Joseph too, standardisation is linked to (though not directly caused by) the political state of Europe in the late 18th century, where the distribution, functional range and even the structural basis of standard languages changed significantly (ibid: 43-44), ultimately leading to the European nations 'heading steadily in the direction of internal linguistic uniformity' (ibid: 48). In this sense *national status* (the function of the standard as a marker for national identity) can be understood as another step towards a still higher level of standardisation, affording the standard a novel ideological status (ibid: 72). The initial demarcation of a language through the process of standardisation (and the perception of clear linguistic boundaries this allows for, discussed further in 3.3.1) is followed by the emergence of a distinct standard language associated with the nation-state (and thus elevated still further above other languages, dialects and registers). In this way, the standard becomes *national property*, belonging to a nation in the same way that a flag or a national anthem does, indexing modernity in the same way these other national properties do (ibid.).

For Milroy, the largely European inclination towards standardisation is not unique to language but part of a broader impulse towards the standardisation of multiple heretofore variable cultural objects. Here Milroy's definition of *imposing uniformity on a class of objects* is most useful, in the context of the wider European push for standardisation of cultural artefacts such as currency or infrastructure, concurrent with the emergence of European nations and nationalism (Milroy, 2001, 541). The notion of standard language, to Milroy, is inseparable from the emergence of other standards in that specific geo-temporal European moment, citing Heilbroner's (1999) report of a businessman travelling through Germany in the year 1550, who comes across a vast array of communities with their own standards and regulations (Milroy, 2001, 541). Even just in the Baden area, the businessman comes across 112 measures for length, 92 square measures, 65 dry measures, 163 cereal measures, 123 liquid measures, 63 liquor measures and 80 pound weights (Heilbroner, 1999, 22). Milroy thus links the development of standard language with these non-linguistic forms of emerging standards such as monetary systems, weights, measures and factory weight goods, and sees the initial impulse for the standardisation of language as one that occurs in tandem with these other standardisations, themselves led by the development of

international trade and the global capitalist system (Milroy, 2001, 541). In this view it is not linguistic factors but in fact economic, commercial and political ones that drive the standardisation of language, and Milroy concludes that standard language cannot be considered as simply another dialect, but rather must be seen as a construction of its own (ibid: 543). This is echoed by Gal, for whom standard language lies on one extreme of the axis of differentiation in the western notion of modernity, while all other forms of language lie to the other end, defined by and indexed against the standard (Gal, 2018, 224). Jaffe (2000) too, in her seminal work on non-standard orthography, emphasises this same point: for her, 'every use of a non-standard form silently invokes the prescriptive power of the standard language myth' (Jaffe, 2000, 511). Milroy's proposal of a *non-ideological* definition of linguistic uniformity is above all refuted by how powerfully he himself demonstrates how the very impulse to apply uniformity to language is born of a geographically and culturally emplaced ideology based on the European development of nationhood and the paradigm of modernity, directly correlating with analogously emergent cultural standards.

3.3 What does *Standard* mean, and to whom?

3.3.1 Standard Language & Academic Attitudes

We have developed a nuanced understanding of the *standard,* at least from an academic perspective. The standard is not autonomously and inherently prestigious, nor does it derive perceptions of prestige and legitimacy from linguistic properties but instead from society and socio-politically accorded prestige. SLC imposes a hegemony of purportedly fixed and absolute linguistic norms, the value and correctness of which is widely recognised, even by those who do not fully use or even know how to use them (Gal, 2018, 222). Normalised, conventionalised and culturally-meaningful arrangements of language are entirely possible in many other ways that do not follow the standard versus non-standard paradigm so linked to the western paradigm of modernity. These perspectives, building upon Joseph (1987), remain recent admissions into the sociolinguistic world, where the clustering of all manner of *non-standard language* on one end of Gal's axis of differentiation in opposition to *standard* has not yet been extensively challenged in academic discourse, and in much linguistic enquiry the standard remains held up as the primary way of understanding of what language is (Milroy, 2001, 530). The implicit view of *language* as *standard language,*

complete with boundaries, structure and unity, an implied monolingualism and clearly defined norms of correctness (oral and orthographical) allows for the study of languages as discrete and clearly-bounded entities (Gal, 2018, 226). For Milroy, 'where there is no centralisation and standardisation, languages are much more fluid and unstable entities than linguists would have believed', and so 'do not easily fit into the structuralist account of languages as coherent systems of interdependent parts' (Milroy, 2001, 540).
Standardisation brings into focus language as a singular and bounded object both in the perception of the language communities that adopt SLC but also in the western field of linguistics, which, being itself built upon a foundation of SLC, has historically consisted of a mode of enquiry steeped in the standard ideology that defines what language *is*. Milroy makes it clear that it is only through the predication of linguistic study on standard languages that we attain an image of language as a discrete or fixed object, with finite and definable boundaries (ibid.). Heryanto (1990) points out that 'language is not a universal category or cultural activity', meaning that 'not all people have a language in a sense of which this term is currently used' (Heryanto, 1990, 41). Thus, SLC-rooted linguistic inquiry is ill-suited for the study of language in cultures completely disconnected from standard language ideology. Milroy offers the studies of George Grace (1990, 1992, 1993) on Austronesian languages as demonstrations of the difficulty of defining what constitutes a language at all, who finds in some instances speakers with no conception of either possessing a language or belonging to a language community (Grace, 1991, 15). In non-standard cultures, languages can be both indeterminate and even undeterminable (Milroy, 2001, 540). Neither the farmer's-language nor the artisan's-language described by Gal (2018) can entirely be called *languages* in a conventional linguistic understanding, despite both the social and linguistic differences between them. Grace, as a result of his work on Austronesian languages, gives an enticing indication of how a new an approach might be formed on the basis of what he calls 'pools of linguistic resources' (Grace, 1981, 263-264). This is a term we have encountered in our previous chapter's discussion of Blommaert's *mature sociolinguistics of writing* (2.1.3), in which Blommaert proposes the study of writing should be based on pools of sub-molecular *orthographical resources*, indicating that perhaps the field may indeed be moving in the direction suggested by Grace and, by extension, Milroy, who finds Grace's proposition of a new system of linguistic analysis highly appealing (ibid: 540). Blommaert's *mature sociolinguistics of writing* is thus part of a broader, mature

sociolinguistic approach to how we conceptualise what language *is*, both of which see a shift in linguistic thought away from firm, fixed boundaries and instead towards variable *resources.*

The usefulness of the notion of resources in both written and spoken contexts is not entirely coincidental, given that orthography is one of the core ways in which standard languages attain the boundary-demarcation. Jaffe (2000) describes orthography as a 'linguistic boundary-marking device' which 'both differentiates a code from other codes, and displays the internal coherence and unity (sameness) of that code'; thus for Jaffe orthography is 'one of the key symbols of language unity and status itself' (Jaffe, 2000, 505). We established (in 3.1.1 and 3.1.2) the key role of writing and orthography in the standardisation process, to which we now add a new role in *demarcation*. It is not through orthography alone that boundary-marking takes place, particularly in cultures not accultured in some way to SLC: Grace's Austronesian languages cannot be expected to develop an awareness of distinct and discrete *language* as in the western paradigm simply by writing their language. In such cases alphabetic writing might lead to internal language-awareness of the internal units of language (as per Joseph), but even this would not be sufficient for the demarcation of different languages as distinct entities with clear limits regarding where one ends and the other begins. It is *standardisation* that leads to this manner of demarcation, and only in that context does the orthography- a distinctly important part of the standard language- play a role in establishing this demarcation.

3.3.2 Standard Language & Popular Perceptions

The attitudes not only of users of standard languages, but any society that has been subjected to the prestige associations of the western cultural domain of language will to some extent be coloured by SLC, given the global prevalence of western modernity. Where variation exists, there is within SLC a perception of there being but one possible resolution: some forms are right, and some are wrong, and even when there is uncertainty about which of several forms is correct, the assumption very much remains that only one form may *actually* be correct (Milroy, 2001, 535-36), a view echoed also by Gal (Gal, 2018, 227; see 3.2.2). Subscription to the ideology of SLC (almost never a voluntary or certainly not

conscious decision) strips the possession of language from native speakers, and instead puts it in the hands of 'nameless institutions' which Milroy likens to 'high priests', charged with maintaining 'canonical correctness' (Milroy, 2001, 537). These Joseph describes as the 'persons who act as forces of linguistic stability', who are 'established in cultural roles' within the community in question and charged with giving the standard 'stability across time' (Joseph, 1987, 6). We must therefore understand that members of standard language cultures believe their judgements on incorrect language use are 'purely linguistic judgements sanctioned by authorities on language', believing there to be no ideological (or prejudiced) element behind such judgements, and often making judgements on their own erroneous use of language, as perceived by themselves, admitting these to be lapses or examples of incompetence, rather than challenging the standard itself (Milroy, 2001, 536). Any sociolinguistic study therefore, even while adopting the views of Joseph, Gal, Milroy, Jaffe and others on the cultural reality underlying *the standard*, must nevertheless retain an awareness of the linguistic and cultural attitudes of members of societies party or otherwise subjected to SLC. We find striking examples of how important a role standard languages have attained on a global level in Sebba (2009), who cites several examples where users insist on prescriptivist spelling systems for their orthographies, such as the Polish orthographic reforms of the 1930s, the Dutch spelling reforms of the 1950s and the Portuguese-Brazilian spelling accord of 1990, within which examples all attempts to allow flexibility through optional orthographical variation were met with resistance (Sebba, 2009, 44). Jaffe describes an equally telling incident through a column written by two Corsican language activists objecting to the variation that existed in the writing of the language from author to author, concluding with the remark that 'every language has its rules. Corsican is a language. Those who oblige themselves to write correctly in French should apply the same rigor to Corsican' (Jaffe, 2000, 506; Perfettini and Agostini, 1994) aptly demonstrating the desire among users of an orthography for it to possess what Jaffe calls significant 'prescriptive power' (Jaffe, 2000, 506). We cannot therefore discount the role of the standard and its prestige in our sociolinguistic understanding of individuals and communities in both their own usage as well as their demands and expectations, which we expect to contain the same echoes of a desire for consistency, correctness and rigour, not because it is practically beneficial, but because the languages with the most prestige adhere to these

characteristics, and thus to not do so would be to demean or in some way lessen the value of their own language.

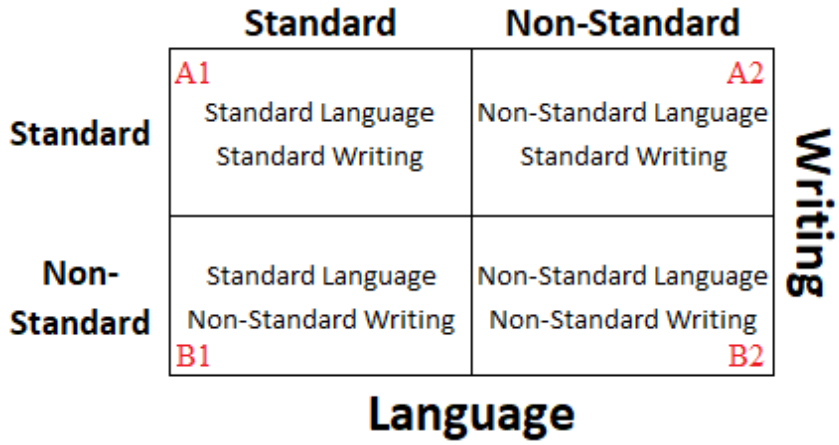
Chapter 4: Non-Standard Orthographies

We have discussed in Chapter 2 the sociolinguistic background of writing, literacy and orthographies, and in the prior Chapter 3 explored the ideological and cultural process of standardisation. We now focus in specifically on the question of standard and non-standard writing and orthographies that lies at the heart of our thesis. We have seen how writing plays a key role in the various stages of the standardisation process, whether in language-awareness, codification, prestige function or demarcation. The complex and multi-layered relationship between writing and standardisation is further explored in the first part of this chapter (4.1), which is dedicated to developing a model for disentangling and understanding the complex interplay between four concepts: standard, non-standard, spoken language and written language. This is followed in 4.2 by discussion of how orthographies (and specifically non-standard orthographies) emerge, and in 4.3 by a discussion of the features, uses and social meanings attributed to non-standard writing, before closing in 4.4 with a final discussion of the model we develop throughout the chapter for understanding the distinction between different types of non-standard writing in the form of two primary categories.

4.1 The Interfaces of Language, Writing & Standardisation

A language need not be standard to be written, and thus need not develop standard writing (Sebba, 2007, 102). We imagine non-standard writing to be used for the representation of non-standard language, and conversely standard writing to represent standard language. But could we imagine a standard language written with a non-standard orthography, or even, a non-standard language written using a standardised orthography? To address these questions, we create a visual representation of this four-way nexus (that is, to the best of my knowledge, novel) in Figure 4.1 below:

Figure 4.1



Such a conceptualisation provides a useful way of visualising the complex relationships between language and writing in standard and non-standard forms. Square A1 represents the primary relationship we have discussed thus far: standard language and standard writing. The non-standard orthographies used by speakers of non-standard languages are represented in B2, such as the Roman script writing of the Lebanese QA of Tripoli, where neither the spoken nor written form of the language is standardised. The case of B2 will naturally be central to most of our discussion, though we first discuss the less familiar combinations found in squares B1 and A2.

4.1.1 Non-Standard Writing of Standard Language (B1)

The idea of a standard language being written using non-standard writing (square B1) is complicated by the necessity of standard writing for the development of standard language. We can reconceptualise this by imagining a standard language that is *re-written* in a non-standard manner, even if a standard orthography has been historically available. A good example of this is the famous study by Androutsopoulos (2000) on German fan-zines that frequently contain non-standard writing within a counter-cultural sub-culture of punk zine writers and participants. Here it is important to determine whether the *language* being used itself can still be considered standard, since if a non-standard dialect of German is being written, this would merely indicate another example of B2. While some of the non-standard writing in Androutsopoulos's study does indeed reflect non-standard language use, there are also instances where *standard German* forms are used but deliberately misspelled in what Androutsopoulos calls cases of 'language-external symbolism', such as <zwex> instead

of standard <zwecks> and <Abwexlung> instead of standard <Abwechslung>, indicating no phonetic or dialectal variation at all (Androutsopoulos, 2000, 524). Such instances, and any case where standard written forms are rewritten for purely performative reasons outside of the representation of non-standard spoken forms are instances of a B1 relationship (other examples being <skool> for standard English "*school",* or <woz> for *"was"*). Such tendencies recall our discussion in 2.3 of *orthographic distancing*, such as Indonesian or Sranan respellings of words with <u> instead of markedly Dutch <oe> in a post-colonial context (2.3.3), though a key difference is that in the case of the German fan-zines it is not another language (colonial or otherwise) that the orthography is being distanced from*,* but in fact the very standard itself is being rejected through socially-meaningful orthographical decisions. The changes in Sranan or Indonesian are synchronous codifications adopted for official, standardised use, but in the case of the fan-zines, the changes are diachronous and socially meaningful acts of non-standard rewriting, thus resembling the diachronous, choices for example of Galician speakers between Spanish and Portuguese orthographical features that are not codified in a single synchronic moment (2.3.2). We find another example of socially meaningful respelling in Jaffe's (2000) discussion of respelled first names, which are pronounced the same way as 'standard names' are, but are given an orthographical flourish as a marker of identity and even deviance, which Jaffe considers to be 'a powerful act of self-representation' (Jaffe, 2000, 508). The B1 case of the rewriting of standard language using non-standard orthography is almost always an act of *expressivity*, whereby the use of the non-standard is a socially-meaningful choice rather than a case of necessity as is sometimes the case for the writing of non-standard languages without a standard representation, such as might occur in a B2 relationship.

The unusual occurrence of B1 for purposes other than expressivity can be imagined in cases where a language such as SA is transcribed using Roman script, such as when an Arabic script is not available. This is different to a modern, colloquial and non-standard Arabic dialect such as LQA being written in a non-standard script (such as the B2 CMC writing of our own study), but rather might occur where *standard Arabic* is transliterated using the Roman script instead of the SA orthography. An example of this is what Palfreyman and Khalil (2007) call *Common Latinized Arabic* (see 5.3.2)*,* which is not a standardised orthography but a conventional way of rendering SA street names and other street signs in

the Roman script, and as such forms a B1 relationship, where a standard language (SA) is written in a non-standard manner. Though this is less common than the examples predicated on the diachronous basis of individual expressive choices, it is nevertheless an example of B1 outside of the sole motivation of the creation of social meaning.

4.1.2 Standard Writing of Non-Standard Language (A2)

The second unconventional combination in Figure 4.1 is A2, where non-standard language is written using a standard orthography. This is particularly difficult to conceptualise given that a standard orthography is generally premised on the standard language it represents. We find, however, a potential answer in the case of non-standard Lebanese QA and its online writing using not the Roman but the Arabic script– a discussion particularly pertinent to our study, thus allowing us to further contextualise our work this section concurrently with our exploration of the A2 relationship. We have understood A1 to be SA written using the SA orthography, B1 to be the writing (or transliteration) of SA using a non-standard Roman script and B2 to be the non-standard Roman script writing of non-standard LQA (which writing we henceforth label *CMCR*, thus using the shorthand term LQA CMCR to describe the Roman-script CMC writing of LQA). Consequently, we understand A2 to be the writing of LQA using the standard orthography of SA. The Arabic script writing of LQA (which we label LQA CMC**A**), occurs frequently online but functions as non-standard orthography, belonging to the same B2 category as LQA CMCR, given that modifications to the standard Arabic orthography are necessary for the writing of the colloquial dialect. However it is in precisely in the instances where changes are *not* necessary that it is possible to glimpse instances of an A2 relationship, where the standard Arabic script of SA, as used for writing a non-standard dialect such has LQA, remains unchanged in its orthographical form, while contextually indicating a non-standard form and pronunciation. This is possible in part because of the optionality afforded by the Arabic script (see 1.2.3), particularly in the diacritics that mark short vowels, which are only seldom used and in specific contexts. In cases where the difference between an SA and QA word is only apparent in the short vowel sounds, the writing of both standard and non-standard forms does not change: the same standard Arabic writing represents both. In this way, we interpret SA writing as partly logographic, in the same manner that even standard English, according to Joseph, has

reverted ‘to a partially logographic state’ (Joseph, 1987, 66), which Sebba defines as a script that is ‘able to represent varieties which, in extreme cases, are also mutually unintelligible’, and where orthographic forms ‘do not necessarily tell the reader how that word is pronounced’ (Sebba, 2007, 110). The phrase <الحمد لله> (*“thank God, praise be to God”*) is realised as /al ħamdu lil.la:h/ in SA, and an identical orthographical form in dialect writing is realised in the case of LQA as /əlħamdəl.la/. Thus LQA can be at least partly written using segments of the same standard orthography of SA, despite being pronounced vastly differently. Some examples of this follow:

Figure 4.2

	SA Script	SA IPA	LQA IPA	*Translation*
1	أكلنا	/ʔakalna/	/ʔiʔkalna/	*“We ate”*
2	كلهم	/kul.luhum/	/kəl.lon/	*“All of them”*
3	انه	/an.nahu/	/ən.nu:/	*“(That) he is”*
4	مدارس	/mada:ris/	/made:rəs/	*“Schools”*
5	حساب	/ħisa:b/	/ħse:b/	*“Account”*
6	المعركة	/ʔal maʕraka/	/ʔəlmaʕərke/	*“The battle”*
7	قال	/kˁa:l/	/ʔa:l/	*“(He) said”*
8	وقت	/wakˁt/	/waʔət/	*“Time”*

In addition to vowel changes, in the case of forms #2 and #3 the <هـ> (/h/) is retained even when written in an LQA context where the sound is no longer produced, and in #2 the <م> (/m/) is frequently written even though it is pronounced /n/ in its LQA form. In these cases, there also exist alternative orthographical realisations that are more expressive, using the Arabic script to phonetically indicate the LQA phonetic forms (thus producing non-standard writing and a B2 relationship). Form #2 can be written as <كلن>, dropping the <هـ> that represents /h/ and using <ن> (to mark /n/) instead of the <م> that marks SA /m/. Form #3 can be written as <إنو>, indicating the changed initial vowel and representing the longer final vowel /u:/ using a written long vowel <و>. To which degree standard orthographic forms are retained– and to which degree LQA forms are indicated in writing– is generally at the discretion of each writer, and forms one of the primary tensions within certain types of non-standard writing, which we discuss further throughout this chapter (and in particular in 4.3.2). It is not the case, however, that all the forms in Figure 4.2 *can* be rewritten to indicate LQA forms, even if a writer desires to do so. Where the SA long vowel /a:/ is

realised in LQA as /e:/ (as is the case in forms #4 and #5), there is no way of representing this vowel using the Arabic script, meaning that <ا> (indicating /a:/) must be retained. This leads to ambiguity as to which vowel is indicated, and pronunciations such as /ħsa:b/ (form #5), more typical of Syrian QA cannot be distinguished from a LQA realisation of /ħse:b/. This is further compounded in form #7, where the <ق> (/kˤ/) is almost always retained in writing, even in cases where it is phonetically dropped in favour of /ʔ/, as in LQA, or even pronounced as /g/ in Gulf QA dialects. A word like <قال> (*"he said"*) is written identically for almost all Arabic dialects, making it impossible to discern from its orthographical form alone whether Lebanese QA (/ʔa:l/) or Qatari QA /ga:l/ is indicated without further context. This phenomenon, in turn, forms the basis for our understanding of the use of CMCR as allowing for written dialectal expression in a way that the use of the Arabic script (even in CMCA) does not (see 6.2.4). Arabic script writing can often be read in a multitude of QA dialects without it always being clear which dialectal form is marked (and even within a single strand of QA such as LQA, which specific local pronunciation is indicated). Words like <المعركة> and <أكلنا> (#1 and #6) have no real way of being rewritten phonetically to match their LQA vocalisations and are thus orthographically indistinguishable from their SA forms. It is in such examples that we see how the writing of SA is, in some instances, logographic. This also means that certain LQA CMCA sentences, by chance, can be written in a manner indistinguishable from SA as written with the SA orthography. We can use the words from Figure 4.2 to produce the following sentence:

أكلناهم كلهم قبل المعركة

"We ate them all before the battle"

This sentence reads as /ʔakalˈna:hum ˈkul.luhum kˤabl al ˈmaʕraka/ in SA, but in LQA produces /ʔəʔkalˈne:hon ˈkil.lon ʔabl əl maˈʕerke/. This sentence is, naturally, the contrived result of a deliberate attempt at creating such a form, and though it may be the case that in non-standard writing, such standard-legible forms appear on occasion by chance, the reality is that there is also a wealth of novel vocabulary present in QA dialects but not SA, in addition to questions of grammatical, syntactical and morphological difference that mean any full text will not truly resemble SA writing. As such we cannot conclude that the entirety of the standard orthography of SA can be used to fully express non-standard LQA, however

in even the limited possibility of rereading the same extract of standard writing either as standard language (SA) or non-standard language (LQA) we can glimpse how an A2 relationship might work, predicated generally on the logographic nature of the writing system. Other examples of logographic writing are more emphatic still in this regard, such as Chinese writing which can be used to write a multitude of dialects, even mutually unintelligible ones, because its characters can be independently phonetically interpreted by each dialect (Joseph, 1987, 36), but Joseph also cites the very logographic nature of this script as the primary reason that standardisation has been delayed in China, given that it allows the delay of the choice or emergence of a prime dialect by allowing for the writing of multiple dialects using the same orthography. As such, while we find in Chinese writing an example still clearer than that of Arabic of a single script representing a multitude of dialects, it is conversely this very fact that has prevented its standardisation even under the cultural pressure of the west and SLC, and thus Chinese writing, too, does not fully function as an example of A2. Nevertheless, logographic writing, in particular where standardised orthographies tend towards logographic features (such as Joseph claims for standard English, and which we further discuss in 4.3.3), provide the primary means of visualising the A2 relationship.

4.1.3 Further Distinctions: B1 and B2

The rewriting of a standard orthography without changing the phonetic or dialectal form indicated, primarily for the purpose of social indexing (such as German <zwecks> or English <woz>) leads to B2: the non-standard writing of a standard language. On the other hand, B1 is simply the non-standard writing of non-standard language, and while there is much scope within this for socially meaningful decision-making, it does not form the sole motivation but co-exists alongside other motivations, including dialectal expressivity, whether in the phonetic rewriting of words or indeed the writing of uncodified colloquial forms that do not feature in the standard at all. The B2 relationship, however, can be further divided into distinct cases, the first being non-standard languages for which there exists a closely-related standard orthography that can be used as a basis for the non-standard writing of the dialect in question, such as English dialects that can use standard English writing as a basis. This is distinct from the second case, which are non-standard languages whose users have no

immediate *standard reflex* to recourse to– no standard written form that can be used as an underlying basis for their non-standard writing. This is a distinction that is almost never made in the field of non-standard writing (see 4.4.2), and yet will be central to a great deal of our discussion. We create the following conceptualisation for this division:

Figure 4.3

[A] Type 1/SR	[B] Type 2/NSR
Non-Standard Orthography as a Reflex of a Standard Orthography	Non-Standard Orthography with no Standard Form in Same Language

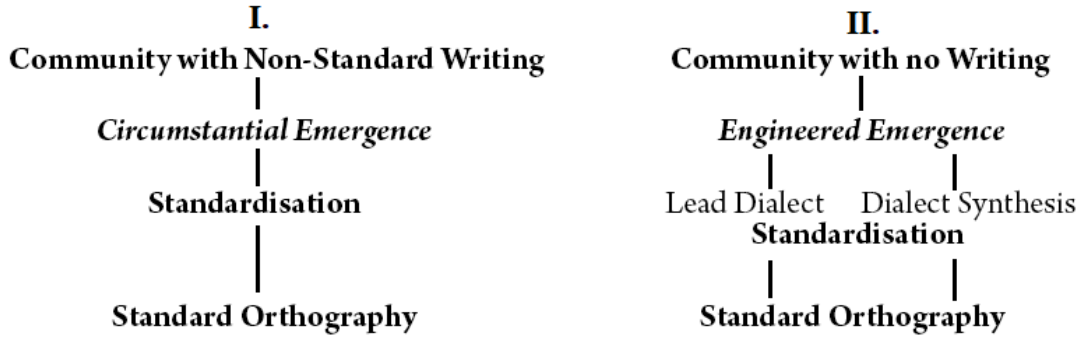
The non-standard writing of non-standard languages can develop in two primary ways: either through a divergence from a written standard that exists within the same community and uses the same writing system (Type 1/SR), or else by adopting a writing system (and some of its orthographical sound-symbol correspondences) from an external source for the expression of a local form (Type 2/NSR), either because there is no standard writing at all in the language community, or for other reasons such as the adoption of the Roman script to form LQA CMCR as a result of the initial unavailability of the Arabic script in early CMC applications (see 5.3.1). In either case, the non-standard orthography of Type 2/NSR has no standard orthographic mapping for writing the language for which it has been adopted, which results in greater creative freedom of expression but also greater variation due to the lack of established orthographic conventions, and unlike Type 1/SR, whose users have some degree of control over how much their orthographic productions diverge from the closely-related standard orthography, users of Type 2/NSR orthographies do not possess a *standard reflex* from which to optionally diverge. The writing of LQA belongs to both categories: it is Type 1/SR when written using Arabic script CMCA (diverging from the available resources of the SA orthography), and Type 2/NSR when it is written using Roman script CMCR, for which there are no standard orthographical resources available in the same script, and so for which there exists no immediate *standard reflex*. We will make this distinction throughout the rest of our work, and in 4.4.2 we summarise what it means in the context of the field as well as our own study specifically, holding to the view that the adoption of such a distinction would serve to delineate important distinctions in the field of non-standard writing.

4.2 How do Orthographies Develop?

4.2.1 Standard Orthographies

There are two primary ways that standard writing is usually introduced (Sebba, 2009, 41), which we can aptly describe by adapting Joseph's terms of *engineered* and *circumstantial* emergence (Joseph, 1987, 61; see 3.1.1). The introduction of an entirely new standard orthography for a language, usually in a community without any previous writing, can be described as *engineered orthographical emergence*, whereas communities in which a non-standard orthography is already in use, standardisation generally occurs through *circumstantial orthographical emergence*. In both cases, the emergence is institutionally led (or at least overseen by trained linguists and other professionals, such as the case of Clifton, 2013; see 2.3.1), and the orthography that emerges is a standard, even if it is not always adopted (such as in Lüpke, 2008, 16; see 3.2.1). The same difficulties that Joseph demonstrates for the choice of a spoken form suitable for standardisation are echoed in both cases: in circumstantial emergence, the multiplicity of non-standard writing systems requires a choice between variants, while in engineered emergence, where no writing previously exists, the choice is instead between which spoken variants are to be represented in the new system. This can be resolved either by the selection of a single spoken dialect upon which the new standard orthography is based, or else by producing a standardised orthography suitable for as many of the dialects within the community as possible, at the cost of a loss of phonetic detail (Sebba, 2007, 110), which we call *dialect synthesis,* and we further elaborate in 4.3.2. In circumstantial emergence, where a non-standard writing system exists in the community, standardisation can be modelled on the basis of the standard writing of a geographically or culturally proximate language community (such as the modelling of Estonian orthography on that of standard German; ibid: 58), or else the orthography of the colonial language with which SLC itself was introduced (such as the modelling of Haitian creole on the standard French orthography; ibid: 84). We summarise these processes in the following figure:

Figure 4.4A



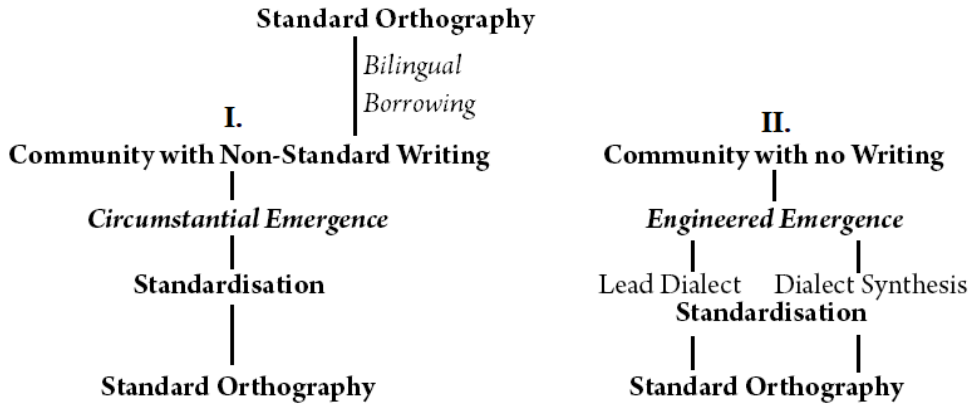
4.2.2 Non-Standard Orthographies

In our discussion thus far, we have somewhat subscribed to the ideology of SLC that speaks to the necessity of the standardisation of a non-standard orthography. While this is a suitable framework in cases where standardisation is actively desired by members of the community in question, we cannot take this as the default expectation for all (or even most) non-standard orthographies. While engineered emergence (process **II.** of Figure 4.4A) is entirely predicated on the process of standardisation, the circumstantial emergence of non-standard writing (process **I.**) describes a process by which non-standard writing organically emerges, which can then only optionally be standardised, in cases where this is desired. Our LQA CMCR writing belongs to this category, being circumstantially emergent rather than formed within a standardisational context. For Sebba, this emergence process is driven by bilingual speakers (Sebba, 2007, 58), and new orthographies, for the most part, and in the modern world especially, are modelled on pre-existing writing systems, while novel writing systems are only seldom devised (Sebba, 2009, 41). Orthographic development takes place within a broader process of interaction with another culture that possess a written tradition (Fishman, 1977, xiv), and bilingual speakers play a crucial role where they speak both the unwritten language as well as another written language, leading to the orthography of the written language being adopted and adapted to be used for writing the unwritten language (Sebba, 2007, 58-59). This is especially the case if there is a cultural or political motivation for a significant number of individuals within an unwritten language community to speak (and write) the same non-native language, such as in colonial and post-colonial circumstances (and as in the influence of standard French and Arabic writing on the non-standard writing of West African languages; Donaldson, 2015, 1-2), but also in cases where

there is cultural and political influence from bordering communities (such as the influence of Danish writing on the Norwegian Bokmål orthography; Sebba, 2007, 108).

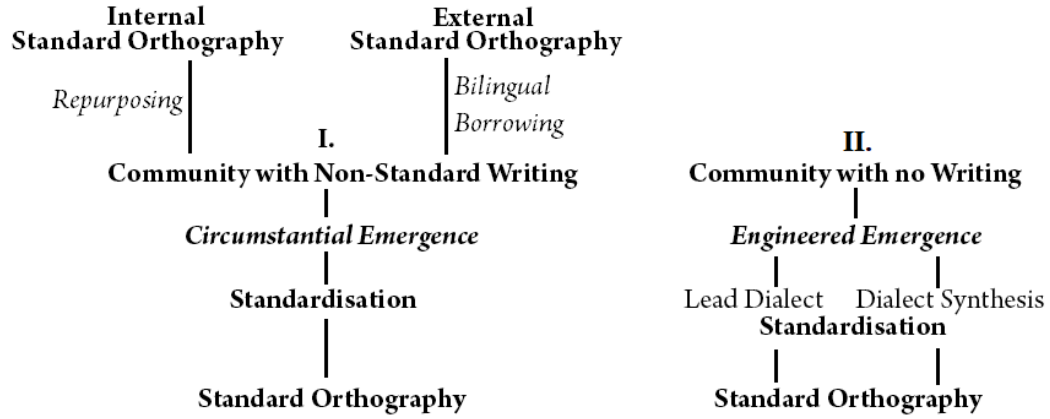
Sebba makes an important comparison between the standardisation of a language (modelled on an already-existing standard language, as described in 3.1.1; Joseph, 1987, 93) and the adoption of an orthography, modelled on an already-existing standard orthography (Sebba, 2007, 59). Thus Sebba's description of how 'bilingual elites […] transfer the conventions of the old standard to the new one' (ibid.) recalls Joseph's example of the bilingual Russian aristocracy of the 18th and 19th century speaking both French and Russian, thereby leading to the standardisation of Russian modelled on standard French (Joseph, 1987, 49). The strong historical German influence in Estonia, particularly on the education system and literacy of the country (Sebba, 2007, 59), led to Kurman's (1968) description of the Estonian writing of 1690 as a 'language poured into the mold of German' (Kurman, 1968, 9). Unlike the institutionally-driven adoption of standard orthographical conventions with the explicit aim of codifying a new standard orthography, the bilingual speakers involved in the borrowing of orthographical features need not be elites, and the writing that emerges out of such grassroots bilingual borrowing is not standardised. A *writing system* (such as the Roman script or the Classical Arabic script) only becomes an *orthography* when it is arranged for the expression of a particular language (see 2.2), irrespective of whether this arrangement is standard or non-standard. Thus, even when a standard orthography of a given language is borrowed for use in another, the spoken conventions of the new language will not correspond perfectly to the conventions of the language for which this writing system originally performed the role of standard orthography, to say nothing of the role this new orthography (along with the language itself) must play in formal functions and codification in order to attain the H-position of a standard. Thus, the new orthography that is adopted by bilingual speakers for the writing of a (usually previously unwritten) language either remains non-standard, or is later standardised as part of circumstantially-emergent standardisation. We therefore modify our original Figure 4.4A to represent this as follows:

Figure 4.4B



This does not, however, describe all types of non-standard orthographies. In the case of dialect writing within a language community already in possession of a standard, we must consider a different approach. The phonetic spelling of English dialects by divergence from standard English is one such example, such as AAVE (African American Vernacular English), a non-standard dialect (or language) with its own non-standard orthography that is in many cases conventionalised, though, given the strict rules of standardisation, cannot be considered standard (Pullum, 1999). AAVE is used within a specific but far-ranging sub-cultural community, using the same script as standard English and much of its orthographical rules and conventions. In this case, this non-standard writing system is not introduced via bilingual borrowing from the orthography of another language, but is actually a case of a community writing its own spoken code using the orthography of the standard, and as such is another example of *circumstantial emergence* of non-standard writing, though it remains equally viable for standardisation given the right circumstances– just as it is viable for it to remain conventionalised, variable and non-standard, with standardisation serving primarily to grant it the ideologically important perception of prestige among its users and their role in the wider national community. We make a final adjustment to represent this in Figure 4.4C below:

Figure 4.4C



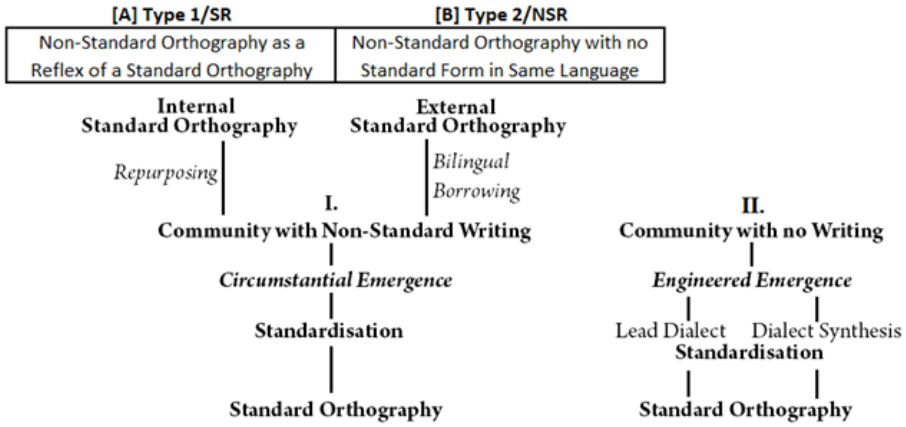
The *internal* standard orthography of standard English is repurposed for the writing of AAVE (representing not only socially-meaningful orthographical changes, as in a B2 situation, but also dialectally and phonetically meaningful ones), while bilingual borrowing conversely takes place through the adaption of an *external* orthography used for the writing of a different, often unrelated language, both of which processes lead to the use of non-standard writing within a community, which can then become standardised through *circumstantial emergence* (or, not at all). The writing of Lebanese QA is a result of process **I.**: there is a non-standard orthography in use within the community, whether using the Arabic (CMCA) or Roman (CMCR) scripts. The non-standard writing of CMCR of our study is a case of a *bilingual borrowing from an external orthography*- in fact, two external orthographies, those of standard French (owing to the former French colonial hold over the country and the use of the language in education) and standard English (owing to the more recent international neo-colonial prestige of the language and its widespread use especially online; see 6.1.2). The non-standard writing of LQA CMCA, on the other hand, using a modified Arabic script, is a case of *internal repurposing* of the standard Arabic writing for the writing of the LQA dialect. These two means of orthographical development, in turn, map onto the two types of non-standard orthographies we described in 4.1.3 and illustrated in Figure 4.3, reproduced below:

Figure 4.3

[A] Type 1/SR	[B] Type 2/NSR
Non-Standard Orthography as a Reflex of a Standard Orthography	Non-Standard Orthography with no Standard Form in Same Language

LQA CMCR belongs in column B, as no Roman-script standard orthography for expressing either SA or LQA is familiar to speakers of LQA, resulting in a non-standard writing based on the writing of external standard languages English and French, and with no standard reflex in the Arabic context of LQA. The writing of LQA CMCA, on the other hand, belongs in column A, where the standard Arabic writing of SA forms a standard reflex for the non-standard writing of LQA CMCA, allowing *optional* representation of non-standard forms. The writing of AAVE is in this context analogous to LQA CMCA: both variants use modified versions of the standard orthographies that serves the languages they both primarily derive from, and for which both standard and non-standard forms are available.

Figure 4.4D



Bringing all this together in Figure 4.4D above gives us a broad perspective of the notions we have discussed in this section. Type 1/SR (*standard reflex*) non-standard orthographies (as AAVE or LQA CMCA) are also cases of repurposed *internal* standard orthographies, while Type 2/NSR (*no standard reflex*) non-standard orthographies (as LQA CMCR) are a case of bilingual borrowing from an *external* standard orthography (or multiple such orthographies). Both ultimately lead to a community with non-standard orthographical resources, which can then optionally be standardised. This is in contrast to process **II.**, where a community with no prior non-standard writing, through engineered emergence, has a standard orthography

introduced, either on the basis of a single lead dialect or else a synthesis of a range of dialects, within which there is no room for non-standard writing.

4.3 What is Non-Standard Writing?

We now turn to the linguistic and social features and characteristics that set non-standard apart from standard writing. Jaffe's (2000) seminal introduction to non-standard orthographies centres around the social meaning of writing, where writing and spelling are not merely 'convenient' and 'arbitrary' codes- that is, not merely (autonomous) systems that make reading and writing possible (see 2.2.1 and 2.2.2), but rather, like all communication, are socially and ideologically meaningful. For Jaffe, 'orthographic choices and their interpretation are read as meta-linguistic, socially conditioned phenomena which shed light on people's attitudes towards both specific language varieties and social identities' (Jaffe, 2000, 498-499), in addition to the meta-linguistic meaning conveyed about subscription or non-subscription to the ideology of SLC at the heart of our discussion in the previous chapter. Adopting Jaffe's approach and making use too of our distinction between Type 1/SR and Type 2/NSR orthographies, we now discuss the nature of non-standard writing, understanding it to not only be a contrastive form to standard writing, but also in some cases as writing that is not in contact with SLC, and therefore, neither standardised nor defined by its non-standardisation.

4.3.1 Prestigious vs. Non-Prestigious Representation

For Jaffe, non-standard orthographies 'graphically capture some of the immediacy, the authenticity and the flavour of the spoken word' (Jaffe, 2000, 498). This is, however, not without its problems. Sebba elaborates on the perceptions of prestige associated with use of standard orthographies, and conversely the negative perceptions of the representation of non-standard speech through text– especially when it is not by the speaker themselves but by a transcriber, rendering such texts risible for most readers (Sebba, 2007, 103). Preston (1982), in the context of folklorist renditions of non-standard speech using non-standard writing, remarks that 'almost all respellings [...] have as their primary effect on the reader a demotion of the opinion of the speaker represented' (Preston, 1982, 323). For Jaffe, the problem lies in 'marking the "orality" of some speakers and not others', whereby only

certain dialects are marked orthographically, while other, usually less detectable dialectal variants are simply written in the standard (Jaffe, 2000, 507). Short of rendering every instance of speech with non-standard transcription, the standard writing of even the most discernibly 'dialectal' speech is the only way to allow such speakers to be perceived as '*the same* as all the other voices in the text' (Jaffe, 2000, 507). Jaffe and Walton (2000) conclude their study of untrained subjects reading respelled texts by saying that 'it is almost impossible to avoid stigma in the non-standard orthographic representation of others' low-status speech' (Jaffe and Walton, 2000, 582). We note that this discussion primarily lies within the context of Type 1/SR non-standard writing, where there is a standard orthographical variant in use within the community, which is diverged from to transcribe dialectal forms. In the case of Type 2/NSR writing, where there is no standard written form to diverge from, atypical transcription is the norm, though as per our discussion in 3.3.2, orthographical variability is almost always perceived as undesirable within SLC, whether for developing or reforming orthographies, or even in communities that neither possess nor are in the process of developing a standardised writing. Where Type 2/NSR writing coexists with a standard form written in a different script, such is the case with LQA, where the standard Arabic script is used for writing SA, there is a negative prestige perception for the writing of an Arabic language (even if it is a QA dialect) using a non-Arabic script, with such transcriptions deemed *improper*, even if they are in some cases the only way to orthographically render the colloquial form (see also 5.3.2). Finally, we note that it is not only the non-standard orthographic form that is associated with low prestige, but also the non-standard spoken form that it represents. Considering our discussion in 3.2.1 of the ideological importance of prestige in the formation of standard languages, it is hardly surprising that non-standard language and non-standard writing both bring with them such judgements of non-prestigiousness, whether in light of their own standard reflexes or of external standard orthographies of other languages.

4.3.2 Transcriptional versus Conventional Writing

We have seen how the expressivity afforded by non-standard orthographies usually comes at the cost of the lower status associated with the use of non-standard writing within SLC. For Jaffe, this is a 'tradeoff between power and intimacy' (Jaffe, 2000, 507). Sebba gives the

example of the German dialect of Alsatian, where each individual possesses an individual ad-hoc graphemic system and where readers must 'sound out' texts on a word-by-word basis (Sebba, 2007, 105). These cases of what Sebba calls *personal orthographies* give their users access to intimacy at the expense not only of prestige, but also by sacrificing a degree of communicability through prioritising 'the representation of the phonetic details which separate their varieties from others', despite it being counterproductive for the efficient reading of their texts (ibid: 106). Jaffe too notes that there exists a second trade-off, this one being between communicability and expressivity (Jaffe, 2000, 501). We thus label the Alsatian writing described by Sebba as *transcriptional* writing*,* understanding the act of transcription to not only be the non-standard transcript produced by a third party of another individual's dialectal speech, but as the very means by which a writer themselves produce non-standard orthographical forms of this kind. What we understand to be transcriptional writing is the non-standard writing that Jaffe describes as interruptive to the 'habitual visual scanning and processing' which usually allows for a 'seamless experience of meaning through text', and whereby the reading of heavily transcriptional non-standard writing 'puts adult readers into a relationship with text that most of them have forgotten in the acquisition of literacy: a decoding mode', with the ultimate consequence being that 'a text that becomes too opaque is simply not read' (ibid: 510-11). Thus we understand the transcriptional writing typical of non-standard orthographies to reverse– to some extent– the advantages of literacy afforded by standard writing, as for Jaffe, 'becoming literate is not just the acquisition of orthographic decoding skills, but also involves the development of a (culturally conditioned) graphic sensibility' (ibid: 509) on the basis of the sound-symbol correspondences of a standard orthography. For Jaffe, learning standardised spelling is 'actually about acquiring a written system that is *divorced* in many ways from speech' (Jaffe, 2000, 502, referencing Kress, 2000, 18; my italics). In this way, we also understand the use of transcriptional non-standard writing as serving to reinstate some degree of the link between written and spoken language.

Sebba describes the non-standard writing of Alsatian to be 'designed for readers whose first language of literacy is German' (Sebba, 2007, 106), marking it as Type 1/SR writing, given that its users rely on standard German orthographical resources to sound out the written dialectal speech. This provides its users with a potential limit to the transcriptional

divergence from the standard form, where variation is optional depending on how much of the spoken dialectal form any given individual chooses to represent orthographically. In contrast, writing transcriptionally is not optional in Type 2/NSR orthographies, given the lack of standardised conventions and sound-symbol correspondences that can variably be strayed from for the sake of for phonetic and social expression (or retained, for prestige, formality or readability). In these cases, transcription is the only means of using such a writing system. Unlike Type 1/SR, variability is not *re-introduced* to Type 2/NSR writing, but rather remains *uneliminated* by any standardisational process. The reduction of variation, however, can also occur through *conventionalisation*, offering a means of limiting transcriptionality through emergent written conventions without the need for either standardisation or SLC, as discussed in 3.2 in the context of Milroy's (2001) proposed purely linguistic homogeneity outside of SLC, Joseph's (1987) *relative language standards* and Gal's (2018) view of normalisation outside the axis of standardisation (all of which we develop further still in 5.2.1). Thus, instead of a dichotomy whereby non-standard writing is always expressive and fully transcriptional, and where only standard writing is non-transcriptional and conventional, we reimagine this as a continuum ranging between transcriptional and conventional writing, with standard writing being the invariable extreme within which correctness is codified and always expected, and less conventionalised non-standard writing being more extremely transcriptional (within which the divorce between speech and writing is– to a certain extent– reconciled). Most non-standard writing is likely to exist in between these extremes, the degree to which it is transcriptional or conventional varying depending on how much conventionalisation has occurred in the case of Type 2/NSR, and additionally the degree to which it diverges from the standard reflex, in the case of Type 1/SR.

Finally, we can also understand the transcriptional-conventional continuum of non-standard writing in terms of distinction versus inclusion. Our discussion (in 4.2.1) of the *engineered emergence* of a standard orthography in the form of a synchronous act of language planning involved navigating a tension between *distinction* (orthographic rules designed on the basis of a single language, affording the orthography greater phonetic detail) and *inclusion* (where an orthography is designed for the writing of multiple variants at the cost of lower encoded phonetic detail). Sebba conceives this as an 'inverse relationship between the amount of phonetic detail in an orthography and the coverage of the orthography' (Sebba, 2007, 110),

which is to say the range of dialects or alternate forms that it is compatible with. This same tension also exists in a diachronous manner within non-standard writing, where phonetic detail marks *expressivity* instead of communicability, given that phonetic detail is continuously variable due to the indeterminate, uncodified nature of non-standard writing. High phonetic detail is likely to be marked in transcriptional writing, while the use of conventional forms can be understood as analogous to *inclusive* orthographies, being less dialectally expressive and instead representing a wider range of dialectal and phonetic forms. We therefore conclude with an understanding that transcriptional writing marks lower prestige, lower communicability, but higher phonetic detail, expressivity and authenticity, and the inverse of these values is marked by the use of more conventional (and in some cases conventionalised) forms. Moreover, outside the perceptions of power and prestige, we can anticipate the continuum between transcriptional expressivity and conventional communicability to hold even outside of SLC.

4.3.3 Transcriptional versus Logographic Writing

We briefly discussed Chinese logographic writing in 4.1.2, as well as cases where standardised orthographies such as those of English or SA come to be partly logographic over time. Reversion to logography is possible even in the case of non-logographic writing systems, particularly when they 'fail to keep up with linguistic change', instead becoming locked into an outdated phonological structure, such as English, which is based on 'a phonological structure which has been obsolete for hundreds of years' (Joseph, 1987, 66). Change is less readily perceptible in spoken language than it is in written language, and within SLC, any visible change usually faces heavy resistance on the basis of the principles of standard ideology, bolstered by the readily-available historical record of what standard writing has *always looked like* and premised on graphicentric ideals which lend authority to written language (ibid: 66). Writing is the holdfast of the standard: for Jaffe, standard language is 'only imperfectly realized in everyday speech' but has a 'palpable existence in writing' (Jaffe, 2000, 500). Even if the spoken standard changes, standardisation is safely stored in writing.

Standard English writing allows the representation of highly distinct language varieties within the same standard orthography, in some cases even varieties that can be considered mutually unintelligible (Sebba, 2007, 110). Such a weakening of the phonetic-graphemic link leads to a lessening of the phonetic detail that we would typically expect to find in alphabetical writing systems, such as that of the Roman script, and thus indicates a shift towards a logography. We see the same factors at work in the Arabic script, often used to represent both SA and LQA with minimal change in orthographic forms (as discussed in 4.1.2). The sound changes that occur over time in any language combined with the graphicentric ideology of SLC means that, over time, we might expect any widely-used written standard that is not already logographic to grow increasingly logographic short of major spelling reforms (Joseph, 1987, 66). Thus, though we have labelled the degree of phonetic expressivity allowed by a standard orthography to be synchronous decision that takes place during codification, ultimately as the spoken standard undergoes change, standard orthographies either lose more of whatever expressivity they originally allowed for, or undergo a repeat of the language planning process in new orthographic reforms. In this way, the shift to logography is one of the long-term *consequences* of standardisation, and as a result, becomes itself a meta-marker of prestige, where highly phonetic orthographies, by contrast, signal lower prestige because of their association with non-standard expressivity through writing. Non-standard orthographies, on the other hand, can be adapted by their users to continue to (optionally) reflect the spoken forms even as they change. Even where conventionalisation occurs, the written conventions available to users are not fixed in the same way standardised spellings are, but optional resources that can be altered or replaced. Though non-standard orthographies, being uncodified, are neither inclusive or exclusive, their inherent variation allows their users to represent a variety of dialects and registers variably through either conventional or transcriptional forms, whether transcription marks historical (i.e. sound change) or geographical (i.e. dialectal) difference.

4.4 Two Models of Non-Standard Writing

4.4.1 The Possibilities of Unstandardised Expression

It is only within SLC that writing which is neither codified nor standardised can be labelled *non-standard*, and only because it does not fulfil the requirements of *standardness*. But

non-standardness itself, as an inherent category, cannot be conceived to exist without the notion of *standard* existing prior to it, as delineated by Gal's (2018) *axis of differentiation* (3.2.2). Within the wide realm of non-standard writing however, there exists a great deal of possibility, wherein perceptions of prestige can be traded off in favour of expressivity (dialectal or social, or both), and even outside of SLC, where prestige need not be sacrificed, only communicability in cases where this writing is heavily transcriptional without extensive conventional resources available to its users. Where there is no pressure for uniformity, phonetic expression becomes a diachronous choice in the hands of the user of the orthography, and not a synchronous process of one-time codification, and so too is the extent of transcriptive writing decided on an individual basis where conventionalised resources are available to limit orthographic variation and increase readability and communicability, whether these conventions are retentions of the standard reflex (in Type 1/SR) or else newly-developed on a grassroots level within the writing of the non-standard (as might occur in both Type 1/SR and Type 2/NSR writing). Though the *divorce* of speech and writing cannot be entirely reversed, given that the very act of writing necessitates a narrowing of the richness of spoken data into limited depiction, writing and speech nevertheless interface more intimately outside of the rigour of the standard and its associated notions of superiority and correctness.

4.4.2 Type 1 & Type 2 Writing in Academic Context

The distinction we have made between Type 1/SR and Type 2/NSR non-standard orthographies is seldom established with any great care or clarity in literature on non-standard writing, even in cases where descriptions apply for one type and not the other, or apply to each in a different manner, while we ourselves have made this distinction clearly throughout our discussion, not least because of its relevance to LQA CMCR as a Type 2/NSR non-standard orthography. Jaffe (2000) identifies two types of non-standard writing studies in the literature, the first of which comprises studies of the 'development or standardisation of previously unwritten minority languages', while the second concerns transcription practices and textual representations of speech (Jaffe, 2000, 500). Jaffe however makes no differentiation between the different ways users of previously unwritten languages develop non-standard writing, nor how this impacts the relationship between the non-standard and

the standard. The studies of transcription that Jaffe discusses pertain primarily to how the social and ideological beliefs of the *transcriber* rather than a user of an orthography are reflected, and is thus unrelated to what we have termed *transcriptional writing* as a mode of writing within non-standard orthographies and their own users. Since Jaffe (2000), this academic landscape has changed most visibly with the significant rise in CMC studies wherein non-standard writing is prevalent and non-standard orthographies have been developed or popularised for a great variety of dialects and other non-standard forms of expression, which we discuss in Chapter 5 to follow. Turning to Sebba's (2007) landmark work on non-standard orthography, we find a categorisation of three primary forms of non-standard orthography, reproduced below:

> 1. vernaculars, in the conventional sense of 'dialects' of an identified standard language
>
> 2. contact varieties and intermediate varieties which are characteristic of situations where creole languages are in contact with their (standard) lexifier languages.
>
> 3. other situations where closely related language varieties exist with a continuum between them
> (Sebba, 2007, 102)

Sebba's first type is analogous to our Type 1/SR, where a non-standard orthography is used for the writing of dialects within the context of a standard language, such as AAVE as a dialect of standard English, or LQA as a dialect of SA. Sebba, however, does not distinguish between whether a dialect such as LQA is written with the Arabic or Roman script; for us, LQA is Type 1/SR when written in the Arabic script (CMCA), but Type 2/NSR when written with the Roman script (CMCR). Our distinction is thus partly concerned with *the manner of writing* of a non-standard form*,* and whether its orthography is a reflex of the one with which its standard language is written (Type 1/SR), or whether it uses an imported script with no established standard orthographical indications for how to write the language using this new script (Type 2/NSR). Sebba's second type of non-standard writing concerns creoles and their lexifiers, and can variably describe either of our types, depending on the nature of the creole in question, the manner and degree of relationship between the languages that inform it, and whether it is written using the same script as that of its lexifier. Jamaican

Creole (JC), for example, is a case of Type 1/SR, where vernacular forms are produced by optional divergence from the standard English orthography (and Roman script) for the representation of phonetic and lexical features unique to JC. We also note that Type 1/SR writing can occasionally include elements more typical of Type 2/NSR writing, where words in a Type 1/SR non-standard orthography have no equivalent in the standard reflex of the lexifier, necessitating users to utilise strategies for writing such forms similar to those we associate with Type 2/NSR writing (such as occurs in Jamaican Creole; see 5.2.2 I). Finally, Sebba's third type describes dialectal variation but does not determine the manner of writing used to represent it, thus making both Type 1/SR and Type 2/NSR viable depending on whether there is a standard orthography written in the same script to act as a *standard reflex* for the non-standard writing in question. Thus we conclude that Sebba, too, does not present the same distinction that we propose.

Jaffe says that 'all "new" codes must choose from a finite number of orthographic conventions and thus, establish relationships with the languages these conventions have been used to codify' (Jaffe, 2000, 505). In the simplest terms, what our two-type model provides is a distinction between the two primary forms this relationship can take, where Type 1/SR indicates a close relationship with a standard orthography used within the same community, usually to write a language *native* to this community, and which is written using the same script as the new non-standard. Type 2/NSR, on the other hand, defines a relationship between a new non-standard orthography that draws on an external standard orthography, usually derived from outside the language community in question and often written using a different script, from which users of the new non-standard writing can derive a limited set of written conventions, but do not inherent a full domain of orthographical forms to which they can resort when not intending to write dialectally. Within SLC, non-standard writing serves the purpose of representing non-standard varieties, such as non-standard dialects and registers that are more usually confined to spoken communication. Type 2/NSR forms can be used for the same purpose, and certainly do allow for expressivity when utilised, though given that they usually emerge in cases where there is no alternative written standard available, their use is not only expressive, but also necessarily communicative. This is the case for the diverse non-standard orthographies in use in West Africa, where both Roman and Arabic scripts are used, often to write the same

languages and where these non-standard orthographies serve different purposes in different contexts (Lüpke, 2008, 12).

We have discussed the categorisation of non-standard as only being meaningful within SLC, where it exists in contrast to the notional standard. Jaffe takes this further, saying that non-standard orthographies themselves are only 'meaningful in comparison and contrast to the standard orthographies that they manipulate or "violate"' Jaffe (2000, 511). Such a view is certainly valid within SLC, where non-standardness will always be perceived in contrast to standardness. It becomes problematic, however, when we define languages or orthographies by necessity as *non-standard*, on only the basis of our own subscription to the standard/non-standard axis of differentiation, even if they exist outside of the duality of SLC and instead within their own axis of differentiation, such as the artisan's-language and farmer's-language of Gal (2018; see 3.2.2), which we reinterpret as *non-standard* only by inducting such forms into our own axis of standard versus non-standard. To their users, these forms are cannot be said to be meaningful in manipulation or violation of a standard which does not exist, even while they retain much in common with the features we have ourselves defined as characteristic of non-standard language and writing. In terms of our two sub-types, we find that Type 1/SR non-standard fits Jaffe's description well, being formed through precisely the manipulation of the standard that Jaffe describes. Type 2/NSR writing, however, is much less clearly defined against the standard form, given that the standard orthography it is usually based on has not been previously used to write the same language, and exists outside of the native language community in question (except as an external, non-native language resource available to members of the community). While there does exist a relationship between Type 2/NSR writing and the external standard forms that inform it, it is a much weaker one than that of Type 1/SR, which is defined by ongoing, optional *deviation* from the standard form, at least until conventions develop for it, and even then those conventions remain in direct contrast to the standardised conventions of the standard orthography. However, that Type 2/NSR writing is less closely associated with a specific standard orthography relative to Type 1/SR does not allow it to escape association with the *conceptual standard*, even if one does not exist within the same language community at all, as long as the ideology of SLC is present. For Jaffe, 'every use of a non-standard form silently invokes the prescriptive power of the standard language myth', to

which we add the qualification that this is true in communities where SLC is present, and we recall our discussions from 3.3.2 and the Corsican desire for the prescriptive power of standardisation and the absolute judgements of *correctness* it brings with it. In this way, it is entirely possible for Type 2/NSR language to be negatively perceived in terms of prestige; that there is no alternative, standardised writing available does not preclude the desire for one.

4.4.3 Conclusions & The Road Ahead

We have developed in this chapter our understanding of the interplay between standard and non-standard, specifically in the context of writing and orthographies, thereby bringing together our work in Chapters 2 and 3 respectively and forming an understanding of how non-standard orthographies develop, for which we have elaborated our own model, as well as understanding how non-standard writing functions in light of the standard, both linguistically and socially. We have also developed a framework for understanding a distinction central to the context of our study of Tripolitan LQA CMCR, distinguishing and defining Type 1/SR and Type 2/NSR non-standard orthographies in a manner never fully delineated in the literature thus far. The Roman script writing of LQA CMCR is a Type 2/NSR non-standard orthography, being not a reflex of a standard written form natively written using the same script, whereas LQA CMCA, using the Arabic script, is defined as a Type 1/SR non-standard orthography. The next chapter will be the final one in which we develop our theoretical groundwork for understanding and approaching our research question, adding a sociolinguistic understanding of the digital media of CMC, and thereafter combining the understanding we have developed across all chapters so far to review the literature that exists in the field of standardisation in the context of CMC writing. Using the concepts and frameworks we have developed in this chapter and those before it, we will reconsider the use of the word *standardisation* to describe both the work done in the field so far, and crucially also how we understand our own thesis and analysis to come. We then return to our work from in Chapter 1 on the sociolinguistics of Arabic and review recent developments within the specific field of QA as written online in light of our sociolinguistic understanding of CMC, and in doing so we thus develop a full framework for undertaking our own analysis of LQA CMCR in Chapter 6 onwards.

Chapter 5: From Writing & Standardisation to CMC & Conventionalisation

We first introduce in this chapter a sociolinguistic discussion of CMC, particularly in relation to questions of vernacular expressivity, convention and identity-performance (5.1), before moving to re-interpreting our understanding of standardisation into one of *conventionalisation,* which we do in context of the prior studies that examine grassroots conventionalisation in online CMC contexts (5.2). Finally, we focus on the CMC writing of QA dialects (5.3), expanding our understanding of the sociolinguistics of Arabic developed in Chapter 1 in light of our new understanding of CMC, as well as critically examining recent work on the CMC writing of LQA from a perspective informed by our understandings of SLC and the writing of non-standard orthographies. By thus completing our review of the various central parts of our thesis question, we can then move on to our preliminary analysis in the next chapter carrying with us a wide-ranging understanding of how our work fits into the various relevant fields.

5.1 Computer-Mediated Communication

5.1.1 The Sociolinguistics of CMC

I. First Wave & Second Wave Studies

The growth in recent decades of computer-mediated communication (CMC) has provided both new avenues for study as well as new challenges to our understanding of language and communication. Androutsopoulos (2006) describes two primary waves of linguistic thought with regards to CMC (Androutsopoulos, 2006, 420), the first essentially summarised in Crystal (2001) who coins the term *Netspeak*, though it would only take three years for academics to begin speaking, instead, of the *Netspeak Myth* (Dürscheid, 2004, in German, *Mythos Netzsprache*). This *myth* encapsulates the core of the first wave of linguistic study of CMC, whereby the use of language on the internet was perceived to be 'distinct, homogenous and indecipherable to "outsiders"' (Androutsopoulos, 2006, 420), taken to be its own distinct *medium* and distinctly differentiable from other communicative media. Crystal defines rigid categories such as *the language of email* and *the language of chatgroups,* with the implicit assumption that these are uniform and free both from

variation within them as genres, as well as variation with regards to who is using them and for what purpose (Crystal, 2001, 148; Androutsopoulos, 2006, 420). In addition to the narrowly medium-centric approach in which 'language of the internet' was seen to be monolithic, invariable and distinct from other language, linguistic work on CMC in the 1990s also commonly suffered from small sample size and a reliance on anecdotal evidence (ibid: 420). Androutsopoulos challenges the existence of such homogenised categories, giving the example of chat being 'more than just the informal setting in which chat-language is described' and pointing out that, for example, political chatrooms are far less likely to use non-standard language as compared to informal ones, while educational chat contexts are likely to encourage other kinds of conventions, such as turn-taking (ibid: 420-421). The second wave of CMC research emerges through the literature which addresses these initial shortcomings, being more concerned instead with the 'interplay of technological, social and contextual factors' and the 'role of linguistic variability in the formation of social interaction and social identities' online, taking a sociolinguistic approach that prioritises the diversity of primarily social elements; thus, the shift from first to second wave reflects a shift from medium-centric to user-centric analysis (ibid: 421). This transition echoes, to some degree, the original development of the sociolinguistic study of both language and of writing in their original, offline contexts, and Androutsopoulos notes a rejection by second wave CMC studies of 'technological determinism' (ibid.) that broadly reflects the same rejection of *autonomous* approaches to writing itself that we discussed in 2.1.2 and 2.2.1. Moreover, modern linguistic study understands the features of CMC as *resources* that users can to draw upon to varying degrees (ibid.), again reflecting the resource-based approach we have understood in the context of language in general (3.3.1) and writing specifically (2.1.3 & 5.2.4 to follow). We now summarise the most pertinent elements of the modern sociolinguistic study of CMC, after which we move on (in 5.1.2) to further develop our understanding of how CMC interfaces with the notions of non-standard writing, expressivity, transcription and convention discussed in the previous chapters.

II. Current Sociolinguistic Perspectives of CMC

Online language communities are defined variously, ranging from lax definitions such as that of Preece et al (2003, 1023), who take any 'group of people who interact in a virtual environment' to be an online language community, while others like Baym (1998, 2003) and

Herring (2004) propose a set of criteria to be fulfilled in order for the label to apply. Others, like Appadurai (1996) and Castells (2000) determine that online communities do not function in the same way as other language communities, being ephemeral, difficult to predict and difficult to define (Androutsopoulos, 2006, 422); for Castells, they 'work in a different plane of reality' (Castells, 2000, 389). We see within our own work on LQA CMCR a hybridisation of conventional and digital language communities, given the physical geographical area denoted by the online communities we examine, (see 6.1). For Androutsopoulos, the social profile of online communities can be visible in 'region-specific chat channels' (Androutsopoulos, 2006, 425), and as such, the online groups we use in the first part of our study (Dataset 1; see 6.1.3) function as windows into the CMC-usage of the wider physical community of Tripoli and the orthographical reflections produced within it, which we also approach by other means, such as the collection of CMC data from individuals in an interview context (Dataset 2; 9.1). Online language variation studies, meanwhile, have been challenged by the lack of phonetic data in CMC, but for Androutsopoulos the traditional phonetic approach to variation can be replaced instead by a primarily orthographic approach, including the study of emoticons, unconventional spellings, the representation of spoken features, the use of obscenity and the employment of code-switching (Androutsopoulos, 2006, 425). Our own work approaches variation from the perspective of the development of orthographic conventions within a non-standardised CMC context, within which regional variation and other factors (such as age and gender) also play a role. Though our aim is not the social categorisation of any quantitative variation we find, there is nevertheless much crossover between that work and ours. Moreover, in utilising voice recordings in conjunction with CMC-based orthographical productions by the same individuals, we are able to examine both orthographic and phonetic data, and variation between them. In a similar vein, Herring (1993, 2000, 2003) has conducted much work on gender in CMC, demonstrating for example how male users, relative to female ones, produce longer messages with strong assertions, use exclusive 'we', have a higher tendency to disagree and lower tendency towards politeness. Later work challenges some of these assertions, such as Huffaker and Calvert (2005) who find that gender is *performed* rather than inherent as a category, and is capable of being performed by members of either gender (recalling similar conclusions for gender performance in a non-CMC Arabic context by Al-Wer, 2014; see 1.3.3), while Herring and Paolillo (2006) find add that certain CMC

genres themselves are gendered, where members of either gender adopt the gendered features of the genre they write. Another facet of the study of CMC is language-choice, and though English dominated CMC in the 1990s, the past decades have seen ever-greater digital linguistic diversity (Androutsopoulos, 2006, 428). Within this, *translanguaging*, similarly to how it is applied for spoken language, has grown in popularity as a means of understanding the use of multilingual resources to produce digital text, allowing for the construction of distinct, trans-lingual online identities (Tagg, 2015, 204). Within this context, central to our work are the studies of the Roman script writing of languages that are not usually written using it, and which we address in 5.3. Digital communication, beyond offering a new landscape for traditional sociolinguistic approaches, also opens up new ones within the field, such as the study of vernacular language in the construction, negotiation and performance of local, ethnic and communal identity whereby dialectal features and spoken accents are utilised in a number of ways through CMC and provide a rich new branch of sociolinguistic interest. What Androutsopoulos calls 'the lack of institutional constraints', and Pietrini (2001) calls 'the triumph of informality' makes the landscape of CMC a rich field of study for the use of vernaculars online, and the 'literalization of varieties that were traditionally confined to spoken discourse' (Androutsopoulos, 2006, 429). The discussion of the vernacular writing of CMC is central to our work and continues in the following section.

5.1.2 Non-Standard CMC: Vernaculars, Expressivity & Convention

I. The Non-Standard Nature of CMC

As a result of a 'lack of institutional constraints', the language of CMC is often highly non-standard in nature, especially in informal contexts such as synchronous chat and even asynchronous communication such as non-business emails between close acquaintances (Androutsopoulos, 2006, 429). Crystal's view of 'highly colloquial constructions and non-standard usage' in chat messages (Crystal, 2001, 165), though modified by the second wave or sociolinguistic approach to CMC studies, is largely retained, even if non-standard features of writing are now understood not on the basis of genre, but individual choices of performativity, prestige and expressivity. Tagg (2015) says that 'the internet is blurring the

line between traditionally private and public spaces, providing a public place in which unregulated vernacular writing can reach a wider audience' (Tagg, 2015, 198). For Coulmas (2013), 'non-standard spellings are some of the most conspicuous features of some kinds of CMC' (Coulmas, 2013, 130). Coulmas also proposes that the *visible* nature of writing (compared to spoken language) is the reason non-standard forms of writing are more objectionable from a popular perspective than non-standard utterances, but this is 'breached' in CMC because it is characterised by a 'quasi- or conceptual orality', indicating another blurring of the lines, this time between spoken and written communication, where '[new] forms of written communication evolve in ways that resemble those characteristic of vernacular speech' (ibid: 130). For Themistocleous (2010b), CMC is a mode that 'shares both spoken and written linguistic features', being 'neither totally speech-like, because the interlocutors cannot see or hear each other, nor totally written, as although it is typed, it lacks planning and editing strategies' (Themistocleous, 2010b, 321, citing Collot and Belmore, 1996). Within this non-standard and speech-like context, we understand Androutsopoulos's statement that CMC encourages the 'literalization of varieties that were traditionally confined to spoken discourse' (Androutsopoulos, 2006, 429). We have previously seen (4.3.2) that non-standard writing even outside of CMC is closely linked with *vernacular* writing, and this connection is only reinforced by the use of CMC. While vernacular writing had been traditionally limited to humour, poetry and the occasional local newspaper publication (Sebba, 2007, 106), with the advent of CMC it has become a natural form of expression, even a pragmatic one considering the oral-like qualities of digital communication. We understand therefore a three-way link between non-standard writing, vernacular writing, and CMC writing.

II. The Use of the Vernacular in CMC

Siebenhaar (2006) studies the use of Swiss German through a corpus of chat logs from 2002 to 2005 from IRC (Internet Relay Chat) networks, finding a high proportion of dialectal features in the written CMC of speakers in Swiss-based regional channels and concluding that 'interactive modes such as chat and e-mail appear to promote the use of [linguistic] varieties which have rarely been used previously for written communication' (Siebenhaar, 2006, 482). Siebenhaar also finds that, where 'there is no standard for dialectal orthography, personal orthographic preferences prevail, and can be inconsistent' (ibid: 483),

echoing Sebba's concept of *personal orthographies* (see 4.3.2). What Siebenhaar describes is, in effect, the same *transcriptional* writing we discussed in the context of Alsatian, wherein Sebba describes each writer of Alsatian as having 'their own graphemic system' (Sebba, 2007, 106), though Siebenhaar finds that the extensive variation found in his study does not generally impede comprehension, meaning either that there are emergent conventions within this non-standard system, or more likely, given its Type 1/SR nature, that there is a tendency not to diverge too distantly from the standard German orthography as to put other readers in the 'decoding mode' described by Jaffe (2000, 510). Here we encounter again the tension between the *transcriptional* nature of expressive writing versus the higher legibility (and lower phonetic detail) of more conventional (or conventionalised) writing. The Swiss German of Siebenhaar and the Alsatian of Sebba are both Type 1/SR non-standard orthographies, both variably diverging from an initial position rooted in the standard writing of German. The major difference between the two is the new scope provided by CMC: Sebba discusses written Alsatian in the context of localised communication such as plays, poetry, articles and humourous pieces, and even demonstrates the breadth of its usage by stating that 'print runs of 1000 sell out readily' (Sebba, 2007, 106)- a number only really impressive in a pre-CMC context. The use of vernacular writing in CMC, on the other hand, can be expected in the case of speakers of virtually any written language using CMC. Themistocleous (2010a) relates how CMC has revitalised the use of (non-standard) Cypriot Greek writing, which, like Swiss German, had previously existed only within the narrow confines typical of pre-CMC non-standard writing. The advent of CMC has therefore produced a virtual landscape in which the use of non-standard vernacular writing is writ large.

III. Vernacular Expression or Identity Performance?

The view of CMC as a reflection of spoken language wherein users use their own vernacular forms has been challenged in recent scholarship, with a growing focus instead on the performative nature of CMC participation, particularly in semi-public settings such as Facebook. Androutsopoulos (2015) finds the 'tendency to view language use in CMC as a reflection of spoken language choices' to be limiting (Androutsopoulos, 2015, 202), instead viewing such writing in the terms of Papacharissi (2009, 211) as 'an ideal environment for the *performance of the self*' (my italics). Hillewaert (2015) describes vernacular features not

as representations of spoken language but as symbols of identity, used strategically and in an *indeterminate* manner based on Jaffe's (2009) notion of indeterminacy that allows for the avoidance of severely negative prestige perceptions by making it unclear whether these representations of ultra-local features are written purposefully or in error. This discussion, however, must be understood in the context of the social variation between different genres and sub-genres of CMC: within the semi-public nature of open sites such as Facebook, performativity and identity-creation are at their strongest, particularly because there is usually a very clear representation of self through the public profile (the Facebook *wall*), use of one's real name, and the ability of one's real-world social circle to glimpse the orthographical productions that take place in this setting. On the other hand, Siebenhaar's study of the use of Swiss vernacular features takes place within IRC (Internet Relay Chat), where anonymity is much more prevalent and so where expressivity might in some cases be preferred over prestigious self-presentation. Bolander and Locher (2010) find that Swiss German users on Facebook make much more extensive use of *implicit* communicative strategies to mark their identity, pointing to how *wall* content on Facebook is produced with the expectation that it is visible to and read by friends, meaning that specific acts of identity construction take precedence over communication, in contrast to Siebenhaar's study of Swiss German in the explicitly expressive communicative context of IRC, where identity performance certainly also takes place, but not necessarily precedence. The Facebook groups we use in our own study share some similarity with IRC chatrooms, since despite being public pages, the writing produced by their members is not automatically displayed to non-members of the groups unless they too subscribe to them (or else if they specifically seek out the groups and view their public content). The writing produced within these pages is not automatically presented to the rest of an individual's wider Facebook community, making these groups more akin to (mostly asynchronous) chatrooms than they are to the traditional Facebook *wall* where semi-public writing takes place in direct contact with one's circle of Facebook friends and family. Rather than necessarily reflecting the general expected nature of semi-public Facebook writing, we instead understand the writing in these groups to share similarities with a medium such as IRC. The communication within these groups can also be understood to be semi-synchronous, as it takes places in the form of comments on the latest news provided by the group, which is updated on a regular basis. As such, there is an immediacy to the comment threads that form beneath each news post,

with shorter amounts of time separating comments than on the Facebook *wall*, and these comment threads are transient, quickly moving out of sight as more news posts are made, where new discussions then emerge.

Finally, the view that limited vernacular forms are used for identity performance in a public or semi-public CMC setting, crucially, also presupposes the use of Type 1/SR non-standard writing, which we have understood to provide resources from a direct and immediately available standard form which can be retained when users do not choose to perform a vernacular identity. In the study of Type 2/NSR writing, where no such standard exists to fall back on, identifying and interpreting such performance is considerably less straight-forward. Indeed, in a Type 2/NSR context, the use of 'personal orthographic preferences' in a 'quasi-oral' manner that ultimately reflects vernacular speech is not optional, but in most cases (where conventional forms are not available) a *necessity* of transcriptional writing. The use of dialectal orthographic features in an orthography such as Tripolitan LQA CMCR cannot therefore be limited only to performative acts because the very framework of the orthography is based on the vernacular spoken in Tripoli, produced transcriptionally and diachronously by its users, though we can of course nevertheless still observe the function of identity-performance even within this, for example in the degree to which particular dialectal features symbolic of certain lower-prestige registers of Tripolitan LQA (in contrast also with higher-prestige Beiruti LQA) are utilised orthographically, as well as the occasional retention of spellings which even in the Roman script continue to reflect etymological SA forms. Ultimately, the use of LQA CMCR as a transcriptional Type 2/NSR orthography means that most orthographical productions are– to some extent– vernacular representations, and as we see in 10.3, very often a direct reflection of the speech of the individual producing them, with the only exceptions being newly-introduced conventional orthographic forms by a process of *grassroots conventionalisation* (5.2 below) which allow for a certain degree of non-transcriptional writing, the most telling LQA CMCR example of which we discuss in 10.2.1. Finally, we see again in this context the importance of distinguishing the *types* of non-standard writing, considering the difference in analytical approach necessitated by whether Type 1/SR or Type 2/NSR writing is the subject of study.

IV. Convention and Expressivity in CMC

Where there are newly-emerging conventions such as in a CMC environment, we observe an interplay between both expressivity and identity as well as transcriptionality and convention, in particular where initially expressive, vernacular written features become conventions over time, used infrequently for the representation not of a transcribed spoken form but a performance of intended identity– something which can occur in both Type 1/SR and Type 2/NSR contexts. Shaw (2008) groups non-standard features of vernacular English CMC (Type 1/SR) into seven types, the final two being of the greatest interest to us and which we re-label as follows:

> **(1): apparent representation of spoken forms -** <gonna>, <bein>, <da>, <fink>
> **(2): irregularisation of regular spelling -** <nite>, <coz>, <cuz>
> (Shaw, 2008, 43)

For Shaw, there is overlap between these two categories, for example where apparent spoken representations from category (1) such as <gonna> and <bein> give very little information about 'accent' and are instead stylistic, their use only indicating that 'this person has adopted the low, covert-prestige variable' as a manner of performance (Shaw, 2008, 43). Though such forms have a basis in spoken language, their use in CMC has become as largely conventionalised markers of purely non-standard *identity* more than they are markers of non-standard *speech*. In this way, these forms too can become *respellings,* with little difference to the irregularly respelled forms of Category (2), which consists of forms such as <nite>, indicating no phonetic difference from standard form <night>, but again indicating an informal or non-conformist identity. Shaw (based on Sebba 2003) calls this *rebellion spelling* (Shaw, 2008, 34), and we immediately understand this as another manifestation of the *respellings* we discussed in in the context of the German fanzines of Androutsopoulos (2000; 4.1.1), and thus instances of what we have called *standard language written with a non-standard orthography* (type B1 in Figure 4.1). Though this kind of *respelling* is naturally only possible in a Type 1/SR context, previously phonetic forms can also come to be conventionally written in Type 2/NSR writing; conversely, transcriptional writing that reflects phonetic and vernacular forms is not confined to Type 2/NSR writing, but can also occur in Type 1/SR writing too, even if it is not *necessitated* in that context.

Shaw describes such instances as *self-revelation,* where the writer 'reveals some assumptions about pronunciation which give information about their actual speech', giving examples of forms like <cuz> and <coz> which 'can show the actual variant used by the speaker', in this case indicating something like /kʌz/ instead of /bikʌz/ (Shaw, 2008, 43-44). The crucial difference is that in Type 1/SR, these are optional choices, rather than inherently part of the creation process of orthographical forms. As such, these Type 1/SR instances are part of the repertoire of identity-creation, being optional and infrequent markers of speech (rather than fully transcriptional sentences), but nevertheless, ones which are not only socially but also phonetically meaningful, presenting 'the possibility of representing one's identity through "accent"' (ibid: 44). Outside of this kind of social-phonetic performativity, we see in Shaw's work something like a process of conventionalisation, whereby certain forms begin as a transcriptional reflections of speech and come with time, on a grassroots, user-driven level, to be used instead as conventional markers of identity. We will apply a similar approach to our own work, but must first develop a richer understanding of how this conventionalisation occurs, which we do through the work conducted thus far on the development of written conventions within CMC under the label of *grassroots standardisation*– something which we now understand to be in need of relabelling.

5.2 Grassroots Conventionalisation

5.2.1 Conventionalisation or Standardisation?

The growth of CMC has allowed access to an unprecedented corpus of non-standard writing for analysis, while at the same time, this same high-frequency use of non-standard writing online has allowed the process of conventionalisation to take place at higher rates than it would have, such as Hinrichs (2004) describes for the CMC writing of JC (Hinrichs, 2004, 93), and which leads Coulmas (2013) to conclude that '[digital] media have different implications for standardization from their predecessors' (Coulmas, 2013, 130). While these new possibilities have yet to lead to a fully-fledged field of study, there are significant works that precede ours, such as Hinrichs (2004), Deuber and Hinrichs (2007) and Rajah-Carrim (2008), which form the basis for our approach to the question of standardisation in the non-standard LQA CMCR writing of Tripoli– or rather, the question of *conventionalisation* therein, for while these studies, like Coulmas, speak of *standardisation,* our discussions in

Chapters 3 and 4 have made the complex, specific and involved nature of the process of standardisation abundantly clear. What these studies address is, in fact, a fraction of the full standardisation process, specifically the emergence of *written conventions,* which we equate with the initial synecdochal emergence of *relative language standards* described by Joseph (1987, 7; 3.2.2), something that occurs even outside SLC and the pressures of standardisation. Just as Joseph's *relative language standards* can develop into *standard language* given the right pressures, so too can such written conventions become part of a codified *standard orthography*, and as such, while their organic emergence in a CMC context *can* be understood in the broader context of standardisation, it remains crucial to note both that such developments first play only a minor role in the decidedly more complex, overarching process of standardisation, and secondly that their emergence need not (and should not be expected to) inherently lead to standardisation of any kind, absent the extensive requisite measures and pressures described in 3.1.1. While the studies we examine in the section to follow do strive to understand the emerging conventions they describe in light of a potential, desired standardisation to follow, only Rajah-Carrim (2008) makes the distinction between the two as different processes, and even she does so only in passing. For us, the organic development of written conventions through conventionalisation occurs not as the result *imposed uniformity* (as in Milroy, 2001; 3.2.1), but such conventions instead are instances of what we might label *emergent uniformity*. Moreover, the uniformity afforded by conventionalisation is necessarily limited, certainly relative to standardised uniformity, and can be understood as uniformity only in contrast with unbounded transcriptional variability (and as Milroy himself points out, *absolute uniformity* is not possible even within standardisation itself; Milroy, 2001, 534). Thus such written conventions function in direct contrast to the transcriptional writing that characterises Type 2/NSR non-standard writing, and are the only real means by which more efficient communication can develop within it, relative to fully transcriptional and difficult to decode writing predicated on highly variable *personal orthographies*. We understand, ultimately, the conflation of the *development of conventions* with *standardisation* to be misguided, both for spoken as well as written language, and so studies that discuss *grassroots standardisation* in a CMC context are more accurately re-labelled studies of *grassroots conventionalisation*– which applies both to our own work, as well as to the previous studies in this field, which we now move to discuss.

5.2.2 Studies of Grassroots Conventionalisation

I. Jamaican Creole

Hinrichs (2004) examines conventionalisation within Jamaican Creole (JC), in a linguistic context where different pressures exist towards the creation of a standardised orthography, with most proposals based on the Cassidy/LePage system used by linguists but not actual speakers of JC. Hinrichs thus promotes the use of organically developing CMC-based conventions for subsequent use as a basis for standardisation, even if he does not explicitly express this distinction in the manner that we do here. What we will henceforth call JC CMC is based on the 'Chaka-chaka' writing of JC, which, due to its relative communicability, has become widespread in CMC use by speakers of JC, within which system 'each writer makes his or her personal spelling, except where spellings of creole words have become established' (Hinrichs, 2004, 92). In this we see a reflection of the *personalised orthographies* of Sebba's (2007) Alsatian and Siebenhaar's (2006) Swiss German, as well as the familiar tension between transcriptional and conventional writing (4.3.2). We categorise JC CMC as Type 1/SR, given that JC is lexified by standard English (hereafter StE), the orthography of which therefore acts as a direct standard reflex for JC CMC. For Hinrichs, deviation from StE in JC CMC is 'employed wherever convenient' (Hinrichs, 2004, 93). Its users therefore have two distinct sets of conventions (or *conventional resources*) to draw upon: either retention of the orthographical conventions of StE orthography or the use of newly-emergent JC CMC conventions, with both strategies allowing users of JC to limit orthographical variability. Outside of CMC, Hinrichs considers what he calls Chaka-chaka writing to be less characterised by deviation from StE compared to other orthographies used by speakers of JC, though this conclusion cannot hold when we speak of JC CMC, wherein Chaka-chaka is no longer an orthography with a set genre and describable characteristics, but must be seen in a sociolinguistic view as a resource that both allows communication and the encoding of identity. In this way, Chaka-chaka becomes a distinct entity when adopted for the writing of JC CMC, defined not as a genre but in the context of the dynamic choices of its users.

Hinrichs observes conventionalisation in JC primarily where semantic confusion is caused by the co-existence of a JC form and its StE cognate (Hinrichs, 2004, 93). Many JC words derive from StE with a shifted meaning in the JC context, while the original StE word from which they derive is also retained in use within JC (ibid: 94). The primary example Hinrichs uses is <yard> and <yaad>, where <yard> is the StE word meaning "*garden*" whereas JC <yaad> indicates "*home*" or even the country of Jamaica itself (ibid: 95). For Hinrichs, this is a prime example for how conventionalisation takes place under the pressure of distinguishing the two words in writing, where StE <yard> and JC <yaad> come to form a contrastive pairing from which arises the orthographical JC CMC convention for distinguishing two semantic forms. Such communicative pressures mean 'new standardized spellings for Creole items arise most quickly', as individuals must 'choose a deviant spelling [...] in order to avoid being misunderstood' (ibid: 96). Such forms become more regularised as users are more likely to 'opt for the spelling of the term they have most frequently seen', and Hinrichs even predicts the possibility of less-frequent spellings <yawd> and <yaard> for the JC variant disappearing entirely in the future, leaving only a 'standardized' spelling of <yaad> (ibid.). We make here a final note upon the terminology, whereby the range of meanings and connotations we understand for the term *standard* precludes its use in such a context, and instead the emergence of such forms is a clear example of *conventionalisation,* characterised by preference and tendency but not strictly enforced regularity. We retain the terminology of the works we analyse in direct quotations from them, but in nearly all cases we understand such descriptions of *standardisation* to be *conventionalisation*. Moreover, we understand Hinrichs' expectation of further regularisation of <yaad> over other forms to their eventual exclusion to be unlikely outside of external standardisational pressures, given the inherent variability of non-standard writing, here illustrating the importance of distinguishing between the two processes beyond mere terminology, but rather as a vital means of understanding the workings and consequences of the process being described.

Hinrichs finds the case of <yaad> is replicated in other forms, too, such as JC <neva>, which though related to StE <never>, functions uniquely to it in JC, with the additional capacity to mark past tense negation where standard English would employ "*did not*" (ibid.). Hinrichs' data demonstrates very clearly that when the word is used to mark the unique JC meaning, <neva> is significantly preferred over <never> (where 72.3% of spellings show <neva> and

27.7% show StE form <never>; ibid.). On the other hand, where the word is used in a context also applicable in StE, there is an even stronger preference this time for the StE form <never> (showing 90.5% use, versus 9.5% for <neva>; ibid: 96). Hinrichs finds in this clear evidence that 'a convention is emerging based on the meaning difference between the JC and StE heteronyms' (ibid.). These orthographic conventions do not mark phonetic differences, but semantic ones, as <never> and <neva> are pronounced alike (ibid: 97), meaning that this is neither transcriptional writing, nor are <neva> or <yaad> identity-indexing performative deviations from the StE forms, which are retained as <never> and <yard> in the relevant semantic positions. The strong preference for each respective form in the appropriate semantic context is clear indication that these choices are motivated by communicative pressure. Hinrichs offers further examples such as <seh>, which shares functions with English <say> but also functions in JC as a conjunction like StE "*that*" or as a quotative indication of speech, all of which are homophonic and not distinguished in speech (ibid.). The StE form <say> is used with high frequency for marking the StE verbal function (93 appearances versus 29 instances of <seh>), whereas when the JC conjunctive function is indicated more frequently by <seh> (appearing 43 versus only 4 instances of <say>). Hinrichs concludes that these non-standard spellings are systematic results of individuals avoiding semantic confusion, in contrast to forms that are optionally respelled in JC to indicate phonetic rather than semantic shift, such as <rispek> for "*respect"*, or <yuh> for "*you",* where the use is performative and similar to the forms discussed by Shaw (2008; see 5.1.2 IV above), and where the use of such words need not necessarily be a phonetic reflection of speech but can also be a simple marker of identity. Here, however, the lack of communicative pressure for differentiation does not lead to the same kind of conventionalisation in JC as occurs for cases where semantic confusion is possible (Hinrichs, 2004, 98).

Finally, for JC words with no etymological link to English, such as second person plural <unu>, variation occurs along the basis of the different sound-symbol correspondences that are available, such as <oo>, <o> and <u> for /u/ and <n> or <nn> for /n/ (ibid: 100). Forms like <oonu>, <unu> and <oono> appear at rates ranging between 5.7 and 17.1% with no distinguishable conventionalisation. Hinrichs demonstrates a case where a single individual writes three forms for this same word: <unuh>, <oonoo> and <unu> (ibid: 101). This forms

another example of *transcriptional* writing, where speech is marked variably using whatever graphemic resources are available, inevitably leading to variation where there are variable graphemic resolutions for any given phoneme. This is more typical of Type 2/NSR orthographies in which all writing necessarily takes place in this manner (at least until conventions develop); in Type 1/SR writing, the same occurs for words that do not derive from the lexifier (as we anticipated in 4.4.2), and thus for which the standard reflex (in this case StE) cannot provide a stable, standardised form. Hinrichs concludes that where there is no 'interference from any English cognate, and because creoleness is sufficiently indicated in the mere choice of the lexical item, writers have no need to resort to any alternative' (ibid: 100). Because Hinrichs' two primary motivations for conventionalisation are absent– distinction from StE forms and the marking of JC identity– he does not anticipate conventionalisation for such words, which instead remain orthographically variable. While this kind of variation is largely unproblematic in the Type 1/SR context, particularly given that variation is typical of non-standard writing, this is only because such forms are generally uncommon within this kind of orthography; in the case of Type 2/NSR, where this is the norm, we will find in our own work that this kind of variable-grapheme variation can also be resolved by means of conventionalisation, and we see something similar even in a Type 1/SR context in Deuber and Hinrichs (2007) below.

II. Jamaican Creole and Nigerian Pidgin

Deuber and Hinrichs (2007) combines Hinrichs' work on JC with a similar approach for Nigerian Pidgin. Much of the same analysis is conducted for JC, with some minor but notable differences, such as <yawd> and <yaad> now presented as competing conventions (Deuber and Hinrichs, 2007, 29), in contrast to Hinrichs' (2004) positioning of <yaad> as near-unanimous, and his prediction that it might become the exclusive orthographic form, with the reversion to variability in the newer study being more in line with what we expect of the natural flux of non-standard writing. Semantic clarity is reiterated as the prime determinant for conventionalisation in JC, though it is made clearer that it is not the only means by which conventionalisation occurs. Deuber and Hinrichs show how the form <mi> is used instead of <me> most frequently when it signals the subjective first person singular pronoun (a role it does not play in StE), but that it is also becoming established as an orthographic variant even in cases where there is 'little danger of misunderstanding or semantic overlap',

including marking the objective, as it does in StE, and yet where the JC form <mi> is nevertheless just as popular as StE form <me> (Deuber and Hinrichs, 2007, 30). Deuber and Hinrichs explain this as a consequence of <mi>/<me> being a high-frequency item, and so conclude that conventionalisation is led not only by semantic clarity (as in Hinrichs, 2004) but also by frequency of use. In spite of this broader view of conventionalisation, however, the discussion of JC in Deuber and Hinrichs (2007) loses a lot of the impact it had in Hinrichs (2004) due to the fact that the sociolinguistic elements are downplayed relative to the earlier paper, with little to no discussion of identity-marking and social intention.

We find in Deuber and Hinrichs also a new perspective through the examination of Nigerian Pidgin, through which work we see that the same pressures and tendencies do not apply to all contexts, even ones as similar as JC and NP, both of which are Type 1/SR non-standard orthographies, lexified by English, used widely in CMC and which signal local identity in contrast to English colonial histories. In NP, even more than in JC, frequency of use is a powerful vector of conventionalisation even where there is no semantic motivation for maintaining clarity. Additionally, many of these emerging conventions are also not based on StE phoneme-grapheme conventions. The form <pikin> (meaning "*child"*) appears 101 times in the data for NP, with the alternative spelling <pickin> (using a more distinctly English <ck>) appearing only 7 times. Not only is the most common form not based on StE spelling, it is also highly conventionalised despite there being no risk of readers mistaking the word for another (ibid: 35). The same goes for <sabi> ("*know"*) and <abi> (a question marker), which appear 213 and 145 times respectively with no notable alternative forms, despite both being at no risk of being misread, and despite possible alternatives such as <sabby> (ibid.). These are, in fact, based on the standard Yoruba spellings of the same words, from which language these words themselves derive: <abi> is simply the standard Yoruba <àbí> written without diacritics, and another conventionalised NP form, <sebi> (another question marker), is similarly derived both in meaning and spelling from standard Yoruba <ṣebí> (ibid.). This introduces a compelling case where other orthographies can also contribute to the process of conventionalisation, and so despite its Type 1/SR nature, NP writing can derive forms from multiple orthographic sources (though both orthographic contributors- StE and Yoruba- are written in the same Roman script as NP), even while remaining clearly predicated on the standard orthographical reflex of its StE lexifier.

Of the NP forms that derive from StE, <dey> ("*there*") and <wey> ("*where*") are particularly interesting for being 'completely dissociated from these etymons' (ibid: 36), and yet are very still heavily conventionalised, being perhaps the most clearly conventionalised forms in the study, appearing 2,495 and 1,546 times respectively (with the second-most popular forms appearing 122 and 110 times respectively; ibid.). According to Deuber and Hinrichs, this convention derives not from the writing of etymological StE forms <there> and <where>, but from other StE orthographical rules where <ey> reflects the /ei/ of StE words such as in <they>, which sound, in turn corresponds to NP pronunciation /e/, leading the correspondence between <ey> and /e/ in NP writing. This is a compelling case of conventionalisation occurring on the basis of StE and yet not on the basis of the StE spelling for the words in question; rather, a sound-symbol correspondence based on the NP pronunciation of StE words is reappropriated for the writing of NP words derived phonetically from spoken StE but not directly orthographically from StE writing. Conventions in NP can therefore arise in multiple manners, as individuals make wide use of the rich linguistic resources available to them, including both the StE and standard Yoruba orthographies. Semantically motivated conventionalisation, however, which was demonstrated to be the main means of conventionalisation in JC CMC, does not appear to occur at all in NP. Deuber and Hinrichs cite one instance of a potentially semantically motivated convention in the case of <don> being used in its NP function as a preverbal perfective aspect marker while <done> is used preferentially in its StE function as a past participle (ibid: 37). Deuber and Hinrichs, however, dispute the semantic motivation behind this contrastive pairing by demonstrating that this does not occur in other places where it would be expected, such as <say> which appears in its StE orthographical form for both NP and StE functions, and finding this to be the case same across all other potential contrastive pairings (including <make> and <them>; ibid.). While it does appear to be the case that semantical motivation plays a more minor role in NP, the fact that it seems to motivate orthographic choices for some words but not others requires further investigation, and perhaps an alternative explanation for the semantic-orthographic pairing of <don> and <done>. Deuber and Hinrichs do propose an explanation for why semantic clarity plays a more minor role in NP more generally, based on the differing approaches to identity-marking between users of JC and NP, where the latter see the use of to use forms that mark

pidgin-ness to be inappropriate in the writing of StE-derived words, and instead reserve orthographic forms that mark it for the writing of words not derived from and unrelated to and StE (ibid: 38), though this does raise the question of why *"there"* and *"where"* are rewritten with distinctly pidgin spelling as <dey> and <wey>.

In summary, we have thus far observed conventionalisation occurring as a result of the marking of semantic clarity (for JC but not for NP), as a result of high usage frequency (in both JC and NP), and as a result of the orthographical resources available to individuals (in the case of NP, the use of Yoruba-derived spellings for Yoruba-derived words). Additionally, we have also observed that what is true for one language community is not necessarily true for another. Finally, Deuber and Hinrichs introduce for NP what Hinrichs (2004) very much argues against for JC: phonetic motivations for orthographical variation. The NP form <dis> appears 364 times, more than the 318 tokens of StE <this>, while NP <dat> appears 312 times, just under StE <that> at 325. These are presented as examples of individuals indicating their pronunciation, though patently missing is a more involved sociolinguistic discussion of the motivations behind individuals choosing to make these representations and what they are choosing to indicate about themselves and their identities through them, or indeed whether they are intended as phonetic transcriptions or instead are more akin to Shaw's (2008) examples of identity-markers that say little about phonetic pronunciation. Given the Type 1/SR context of this writing, these are not unmotivated choices but *respellings*, and in this case, ones not explained by the motivations of semantic pressure.

III. Mauritian Kreol

Unlike the previous two studies, Rajah-Carrim (2008) is not primarily a quantitative study of tokens within a corpus, but instead more directly concerned with the attitudes of individuals towards the orthographical resources available to them. Mauritian Kreol (MK) is lexified by standard French (StF) rather than StE, and despite being the first language of 75% of the population of Mauritius, it is perceived negatively and sometimes associated with the Afro-Mauritian ethno-religious group in particular (Rajah-Carrim, 2008, 485). Rajah-Carrim cites the lack of a standardised writing system as one of the reasons that Kreol is negatively perceived, and describes attempts at producing this, including *Ledikasyon Pu Travayer* (LPT) which is 'based on phonemic principles [...] without reference to the lexifier', and *grafi legliz*

(GL) which she takes to be an 'intermediate phonemic orthography' that is 'largely based on phonemic principles but does make some concessions to French spellings' (ibid: 486). What these inherent 'phonemic principles' might actually be is not specified, but in the context of Rajah-Carrim's discussion they are better described as principles grounded in the StE orthographic system, which thus mark *distance* from the local lexifier of StF (see 2.3). In this we see a reversal of the Haitian situation, where closeness to and retention of StF features was preferred to StE features associated with neo-colonial incursion (see 2.3.3). Though neither LPT nor GL retain the original StF spellings, they vary in the degree to which they reconstrue the StF forms orthographically, where StF <boire> ('*drink*') becomes <bwar> in LPT but <boir> in GL, and StF <cuire> ('*cook*') becomes <kwi> in LPT but <koui> in GL. In these examples we understand the use of <w> instead of <oui> or <oi> as a distancing strategy from StF writing, utilised in GL but not LPT, while <k> instead of <c> is a distancing strategy that both GL and LPT adopt. In this way we further understand Rajah-Carrim's *phonemic spelling* in terms of phonemic distance from StF, such as in the use of alternative resources like <w>, largely deriving from StE orthographical convention, though <k> appears to be a wide-ranging marker of difference not only marking difference from lexifiers like StF or standard Spanish (such as in Basque; see 2.3.2), but even distance from StE orthography itself, such as in the NP use of <pikin> instead of <pickin>. Ultimately, Rajah-Carrim's clear distinction between 'etymological' (i.e. StF) and 'phonemic' (i.e. StE) derivations (ibid: 487) is somewhat problematic, not least due to the difficulty of determining what a truly *neutral* phonemic writing system might be.

In Rajah-Carrim's description of MK as it is used in CMC writing, we see that variation is largely orthographical rather than phonetic or semantic. Hinrichs (2004) defines a *creole continuum* in the use of spoken JC as the range between basilect and mesolect, which is to say, different degrees of *creole*-ness*,* which can also be reflected in writing of JC CMC (Hinrichs, 2004, 93-94). We see something similar in the CMC writing of MK, except instead of ranging between indicated spoken forms, the continuum is purely orthographic, and different degrees of closeness or distance are marked in spellings that ultimately mark the same phonetic forms. We see this most clearly in Rajah-Carrim's example of a sentence written in three orthographic styles:

(1) mo cause creole avec toi

(2) mo koz Kreol avek twa

(3) mo cose creol avek toi

(Rajah-Carrim, 2008, 487)

These sentences vary only orthographically, based on the degree of distancing (or otherwise) from StF writing conventions, unlike JC, where there are relatively fewer orthographical options and where the variation is instead dialectal– indeed, even in cases where a word's JC function is marked by an alternative spelling, the spelling does not vary for orthographical purposes, but instead for marking different meaning. The core concern in JC is not distance from the lexifier as it is in MK, but clarity of meaning, while in NP non-English identity is often marked for non-English words, but is of less concern in the case of English-derived words; in MK, however, variation appears to be primarily stylistic and orthographic.

In addition to the usual label of standardisation rather than conventionalisation for organically arising conventions, Rajah-Carrim also states that MK 'is still largely perceived as a nonstandard language by its speakers' (Rajah-Carrim, 2008, 486), a perspective that can only follow from the conviction that standardisation is possible merely through conventionalisation: from our perspective, it is hardly surprising that MK is not in any way standardised given that it has not gone through any process of standardisation, even if certain written conventions might be in use within its CMC writing. Rajah-Carrim does cite Joseph (1987) and gives an admittedly brief summary of the process of standardisation, including questions of which variety of a language is chosen for standardisation as well as discussing the prestige, legitimacy and status that standardisation confers upon a variant (ibid: 487). Nevertheless, we must again re-interpret most of what she calls *standardisation* to be grassroots orthographic *conventionalisation*. For speakers of MK, CMC is important not only as a platform where high frequency use of the language in writing is possible, but also as a space where Mauritians of all ethno-linguistic backgrounds use this shared code in order to communicate across ethno-linguistic boundaries. Rajah-Carrim finds 'interesting parallels between the *standardization'* (my italics) of Kreol and that of JC (as related in

Hinrichs, 2004) due to the fact that both are ‘user-driven and indexed as markers of a local identity’ (Rajah-Carrim, 2008, 489). This similarity is in fact better described in terms of *grassroots conventionalisation*, of which both orthographies are examples, and as such are user-driven by their very nature (and so that overlap is no coincidence). Rajah-Carrim poses the questions ‘can CMC promote the standardisation of the language? Can standardisation be brought about by people at the grassroot level[?]’ (ibid: 489), to which the answer is *no* when taken at face-value, but which become more interesting when re-interpreted as questions of how conventionalisation can occur on a grassroots level, and, should there be a desire to, can then be used as a basis for constructing a standardised orthography.

Rajah-Carrim's questionnaire approach (including a clear and self-reported record of the group to which any given respondent belongs) makes sense from the perspective of Mauritius and the importance of the ethno-religious sub-groups within the island. Her participants (from a number of ethno-religious backgrounds) were presented with six extracts written in a variety of registers, ranging between what she calls *etymological*, *phonemic* and *mixed*. They were asked to rate these by readability, learnability and closeness to StF, and the results were largely as expected, perhaps due to the fairly straight-forward questions posed: the more etymological spellings closer to StF were considered more learnable and legible by the respondents, all of whom speak and write StF, whereas the ‘phonemic’ extracts were rated more difficult to decipher (ibid: 493). Rajah-Carrim's follow-up within the same study, however, is of greater interest, resembling as it does the work of Hinrichs (2004) and Deuber and Hinrichs (2007) a great deal more. Collecting MK passages from a CMC database, Rajah-Carrim examines the attitudes reflected by the orthographical productions she encounters. For her first example, the retention of StF orthographical forms leads Rajah-Carrim to conclude that its writer implicitly maintains the 'traditional linguistic hierarchy' in which MK is seen as ‘a derivative or an inferior form of French’, while concluding that the use of ‘phonemic’ spelling in her second example ‘indexes Kreol identity by obscuring the French origins of the words and highlighting their uniqueness’ (Rajah-Carrim, 2008, 502). Rajah-Carrim finally examines the five most commonly used words in the same manner as Hinrichs (2004) and Deuber and Hinrichs (2007), and finds four words which tend towards ‘phonemic’ spelling and only one that tends towards ‘etymological’ spelling, though as she herself says, a much wider study of this

kind is required before further conclusions can be drawn (Rajah-Carrim, 2008, 501). As such, it is a shame that the work is focused more on the questionnaire that produces relatively unsurprising answers rather than on more of this kind of work, though it is certainly telling that the respondents generally claim a preference for the 'etymological' spellings that indicate a closeness to StF, which is belied by how MK is actually used within CMC, where it would appear that in reality what Rajah-Carrim calls phonemic spelling is potentially more popular (ibid.). Rajah-Carrim concludes that MK allows writers 'the convenience of a nonstandard language which can be written in various creative ways' (ibid: 504), echoing our discussions of expressivity in non-standard writing, and raises– for us, at least– the question of the value of a standard orthography if it is to come at the expense of the convenience and creativity afforded by non-standard MK CMC and the flexible possibilities for identity-indexing that the range of orthographical forms allows for, though we must also acknowledge the effect of SLC and the negative perceptions of prestige and value users of MK will likely continue to ascribe to this orthography if it remains unstandardised. Rajah-Carrim notes that 'Mauritians have devised their own orthography and interestingly, orient towards some specific phonemic forms' (ibid: 505), in which we glimpse a potential process of conventionalisation, and so a broader quantitative study of word-frequency of MK words in the fashion of Hinrichs (2004) and Deuber and Hinrichs (2007) would be of great interest, and would be expected to further reveal the manner in which conventions ('etymological' or 'phonemic') develop and are used by speakers of MK.

5.2.3 Grassroots Conventionalisation Summarised

Variation in MK primarily occurs through variable orthographical representations of largely the same words, such as in <boire> and <bwa>, where there is no clearly indicated difference in phonetic realisation, in contrast to JC where respellings are utilised not for the purpose of representing vernacular forms but rather to delineate meaningful semantic distinctions. Variation in MK is therefore attributable to social motivation for difference-creation on a purely orthographical basis, whereas NP does the same on a phonetic basis in vernacular-based respellings like <dis> and <dat> that are not semantically significant but used instead to represent vernacular speech and likely index identity. The writing of MK, like that of JC and NP, is Type 1/SR, with the StF orthography of its StF lexifier serving as its

standard reflex. The case of NP is complicated by the influence of the orthography of other indigenous languages spoken in Nigeria, such as Yoruba, and yet nevertheless remains clearly a language with a Type 1/SR writing system derived from English, just as much of the language's lexicon is. Also notable in NP is the fact that localised orthographical realisations of words (on the basis of the equivalent NP pronunciation) take place primarily for non-English derived words, for which a non-standard spelling is seen as appropriate, whereas English-derived words tend to retain their English spellings, even where there is risk of misunderstanding through semantic overlap. On the basis of these works, we now produce the following primary motivations for the occurrence of conventionalisation in online orthographies in the case of Type 1/SR non-standard orthographies, all of which are encouraged through the high-frequency use of non-standard orthographies such as takes place in CMC:

A. **Orthographical differentiation** from a lexifier or dominant colonial language for purposes of national or local social identity (such as MK <bwa> instead of <boire>).
B. **Phonetic realisation** of the non-standard language for purposes of expressivity and social identity (such as <dis> and <dat> in NP, or <rispek> in JC).
C. **Avoidance of semantic confusion** where words have been adopted from the standard with a modified meaning that co-exist with their etymons (such as <yard> and <yaad> in JC).
D. **Minimisation of variability** which can either occur on the basis of the language these words derive from (such as Yourba for NP, which guides the conventionalised spellings of <sebi> and <abi>), or more generally through high-frequency selection of a preferred form on the basis of the orthographical rules, often premised on the writing of the lexifier language, such as <dey> and <wey> in NP.

The existence of a clear lexifier or other primary language that shapes the writing of the non-standard language in question is central to many of these processes. Type 1/SR non-standard writing makes possible the creation of distance on the basis of respelling, and is the primary cause for the semantic overlap that leads in the case of JC to conventionalised solutions. We would expect to find some of these same effects in LQA CMCA, being the non-standard Arabic script writing of LQA in a CMC context, as it too is a Type 1/SR non-standard

orthography. In the case of LQA CMCR, however, while we still expect the same pressures to be present, we also expect them to function differently due to the different nature of its Type 2/NSR writing. While the LQA lexicon derives primarily from SA (substratal elements and historical borrowings aside), LQA CMCR is written using the Roman script, and borrows variously (and to various extents) some of the orthographical associations of StE and StF. While semantic confusion (C.) may also play a role in the orthographical choices of users of LQA CMCR, it must necessarily function differently to how it does in JC as the majority of LQA CMCR lexemes do not derive from the languages that inform its orthography (StE and StF) in the way that JC derives words orthographically and semantically from StE. There also can be no motivation in LQA CMCR to differentiate between standard writing and locally-indexed non-standard writing (A.), given that there is no standard way of writing this language in this script in the first place, as a result of which we conversely can expect phonetic realisation (B.) to become heightened, given that the orthography itself is inherently transcriptional and has there is no real 'etymological' writing to draw upon aside from minor instances of transliterative imitation of the SA orthography where such is possible (and which in turn leads not to more uniformity, but in the competition with other forms, more *variability*). This, in turn, is likely to lead to problematic circumstances of variation (D.) within the writing, and it is on the basis of this that we expect most conventionalisation to be driven: as grassroots resolutions to the difficulty of reading transcriptional writing. Our study is unique in addressing a Type 2/NSR non-standard orthography of a language which is neither a creole nor a pidgin, and which is not tied to any standard orthography at all. Though the emergence of orthographic conventions through what we have labelled grassroots conventionalisation has also been put forward as a means by which a *standardised* orthography might be ultimately produced for a number of languages, such as Romani (Matras, 2005) and Bahmian Creole (Oenbring, 2013), to my best knowledge no study of grassroots conventionalisation has been conducted for any orthography qualifying as Type 2/NSR on the basis of our criteria. This makes any evidence we find for conventionalisation within LQA CMCR not only unique, but also a further affirmation of the viability what Deuber and Hinrichs (2007) call *grassroots standardisation* and which we have re-termed *grassroots conventionalisation*.

5.2.4 Grassroots Orthographies

I. The Resources of Language, Writing and CMC

We have observed the growing prevalence of the notion of *linguistic resources* for understanding both languages and orthographies. Instead of distinct languages, this model instead considers molecular *pools of linguistic resources* upon which individual speakers draw, offering fluid alternative views to rigidly defined and arbitrarily delineated boundaries between languages, introduced by Grace (1981) in the context of the Austronesian languages he understood to be unbounded and uncategorisable and espoused by Milroy (2001) as part of a hypothetical new system of analysis (3.3.1). For Tagg (2015), 'the set of language resources which any one individual has access to is *emergent*', meaning that these resources build up and also vary over time, based on the interactions of an individual with any community or set of communities (Tagg, 2015, 10). Blommaert (2013) brings the concept of resources firmly into the field of writing, identifying the most prominent resources that any writer can (or else is unable to) draw upon, ranging from simple resources such as the actual physical possession of pen and paper, to highly social and cultural resources such as ritualised expectations for different forms and in different genres (Blommaert, 2013, as discussed in 2.1.3), which Blommaert uses as the basis for what he calls a 'mature sociolinguistics of writing'. The resources of CMC writing are, too, an extension of the resources of writing, just as the resources of writing are themselves an extension of the resources of language. For Androutsopoulos (2006), what were initially understood to be 'characteristic features of "*the language of CMC*" are now understood as *resources* that particular (groups of) users might draw on' (Androutsopoulos, 2006, 421; my italics). The resources of CMC consist not only the social expectations for various CMC genres, but even basic questions of access to the internet, and not so basic questions such as access to one's native script, the lack of which *resource* was a major reason for the emergence of Roman script writing such as that of LQA CMCR (see 5.3.1 ahead), and which in turn continue to be used because such non-standard scripts have themselves *become* useful resources, for example for the expression of a vernacular orality and localised identity, or even just the simple expressivity of writing in a manner closer to an individual's own dialect than had not been possible prior. The resources of non-CMC writing, despite overlap with those of CMC-writing, nevertheless in theory offer a less limited scope for

shape-creation by free-hand, though of course even in writing (rather than typing) this is a resource limited by the actual scripts in use, and which graphic productions can actually hold meaning for any potential reader. The use of CMC therefore partly modifies the resources of writing, adding new possibilities but also new constraints. For Coulmas, 'CMC-induced changes are more likely to supplement rather than replace established features and modes of writing' (Coulmas, 2013, 132), reinforcing the view of CMC writing as an extension of non-CMC writing, where the use of CMC is dependent on drawing upon traditionally written resources (primarily *literacy* itself), but at the same time, it is also a space where different resources are available (and sometimes, required).

II. Grassroots Orthographies

The term *grassroots literacy* was first introduced by Fabian (1990), for whom it is a form of writing 'rooted in orality' and which 'cannot be read (understood, translated) by outsiders except 'ethnographically', by way of 'performing' the written script according to the rules that govern oral communication in this culture' (Fabian, 1990, 2). Fabian later supplements this with the idea of 'a literacy which works despite an amazingly high degree of indeterminacy and freedom' (Fabian, 1993, 90). Blommaert (2008) adopts this term for a broad range of meanings: for him, *grassroots literacy* is a non-elite form of writing produced by mostly marginalised individuals 'who are not fully inserted into elite economies of information, language and literacy' (Blommaert, 2008, 10-11). Aspects of these literacies include the use of symbols, often hand-drawn, that 'defy standard *orthographic* norms', with writing in some cases becoming *drawn* rather than written, calling to mind the use of numbers as orthographically-meaningful symbols in Arabic CMC writing and certainly that of LQA CMCR, such as <3> for the voiced pharyngeal fricative /ʕ/. Grassroots literacies also often show difficulties with spelling, including highly variable forms which 'very often reflect 'accent', the way in which they are pronounced in spoken vernacular varieties' (ibid: 10)– in essence, utilising what we have called *transcriptional* writing and, in existing outside of SLC, undoing in part the *divorce* between language and writing (4.3.2), something we also expect to occur in the Type 2/NSR writing of orthographies such as LQA CMCR. Users of grassroots literacies often 'write in local, so-called 'substandard' varieties of language use code-switching, colloquialisms and other 'impurities' in their written texts'; the texts produced are ones of 'constrained mobility', being 'only locally meaningful and valuable', and lose

value and legibility once moved outside their local place of origin (ibid.). In many ways, we can understand LQA CMCR to be a *grassroots orthography*, whereby the adoption of the Roman script has severed– at least partly– the orthography's rooting in the Arabic script and the orthography of SA, losing in the process the standard or standard-like features of Type 1/SR dialect-writing. There is a certain marginalisation of Tripoli (and the north of Lebanon) relative to the capital of Beirut (see 6.1.1 ahead), on top of which the adoption of the Roman script further *marginalises* the writing of LQA CMCR as it leads to unorthodox sound-symbol correspondences such as the use of numbers to write, a free, initially unstructured variability and a vernacular orality. In being a user-driven writing of a non-standard, partly marginalised dialect, existing outside the realm of SLC, prestige and orthographic norms, with high variability, indeterminacy and freedom, we understand LQA CMCR to belong– at least partly– to the paradigm of grassroots literacy.

For Lüpke, 'performativity and fluidity hold for grassroots literacies in the Latin script' just as they do for the writing of the same West African vernacular non-standard languages using other scripts such as Ajami; for him, the non-standard writing produced by speakers of West African languages in these scripts is very much an example of grassroots writing, though it does not follow for Lüpke that they are cases of 'stand-alone literacy', but rather he considers them rooted in *lead languages* that provide the sound-symbol correspondences and serve as the orthographic lead (Lüpke, 2018, 12). In our own terms, we understand the *lead language* as the collection of orthographic correspondences from which Type 2/SR writing derives, where in the case of LQA we expect the lead-languages to be StE and StF, the orthographic correspondences of which are resources that individuals are likely to draw upon in producing their writing, alongside other linguistic resources including both the writing and language of SA, as well as their own non-standard spoken LQA dialect (and within it, a variety of registers including both regional differences and degrees of *locality*; see 6.1.2). Conversely, just as the adoption of the Roman script produces LQA CMCR in the mould of a *grassroots literacy*, the process of *grassroots conventionalisation* will do the opposite, reducing the transcriptional variability typical of grassroots writing through the introduction of conventional and conventionalised resources into the pool available to its users.

5.3 The CMC Writing of Arabic

We have developed clear expectations for how grassroots conventionalisation takes shape in a CMC setting, bringing together the various strands developed throughout previous chapters, including the sociolinguistics of writing, standardisation and non-standard writing, and applying them into a CMC sociolinguistic context. It only remains for us to apply these understandings to the sociolinguistics of Arabic (and the QA dialects) discussed in Chapter 1, which we now also extend into the realm of CMC in order to better understand the specific socio-cultural context in which the CMC writing of LQA occurs, and to which this section and the remainder of this chapter is dedicated.

5.3.1 Roman Script Writing of Non-Roman Orthographies

As a consequence of the ASCII encoding of the early internet, the Roman script has been widely adopted for the writing of languages traditionally written using other scripts, as ASCII is based on an Anglocentric Roman orthography with little capacity for representing anything outside that narrow confine (Themistocleous, 2010b, 319). By the time the Unicode Standard was introduced, allowing the use of non-Roman scripts, various traditions of Roman script writing had developed for the writing of non-Roman script languages (ibid: 320). Such usages are, unsurprisingly, complicated by sound-symbol correspondences of the language being written not matching up with the characters made available by the Roman script. Creative solutions have emerged for these across various languages, and in the past two decades a growing number of studies have been dedicated to examining this phenomenon. Lee (2007) looks at Email and ICQ usage in Hong Kong, and Su (2007) at Taiwanese on BBS (Electronic Bulletin Boards). Mokroborodova (2008) focuses on the Roman script writing of Cyrillic languages, in particular the 'new spelling' of Russian on the internet. The use of Greek in CMC has been especially well documented in such works as Tseliga (2007) and Androutsopoulos (2009), as well as those focused on particular dialects of Greek. Themistocleous (2010a) examines Cypriot Greek (CG) as it appears on IRC, where the Greek script is not available, gathering data for a corpus of some 5,500 words from the IRC channel #Cyprus in which the majority of members are speakers of CG, and analysing how individual phonemes are realised graphemically using the Roman script. She concludes that 'rather than suppressing their own language to conform to the technological constrains of

IRC, Greek-speakers have successfully promoted their language within this global environment' through creative use of orthography (Themistocleous, 2010a, 165). The use of the Roman script in the writing of CG has also allowed for the orthographic expression of particular CG phonemes that are difficult or impossible to render using the standard Greek orthography (ibid: 158), leading to greater freedom of vernacular expression than that available in the use of standard Greek writing, even if it comes at the cost of difficulty in the mapping of sound-symbol correspondences. These are, in turn, creatively resolved, such as by the use of numbers to mimic the Greek script, including the use of <w> for <ω> (/o/) and <8> for <θ> (/θ/). Unlike the Type 1/SR contexts reviewed thus far, the adoption of a new writing system makes the CMC writing of CG another case of Type 2/NSR, whereby the expression of vernacular forms cannot be purely performative as a consequence of the transcriptional nature of its writing. Nevertheless, as we anticipated (5.1.2 III above), identity performance nevertheless remains as much of a motivation for users of Type 2/NSR orthographies, and in the case of CG is visible where users 'sometimes write a Greek character phonetically and in other cases orthographically' (as summarised in this case by Themistocleous, 2010b, 323), recalling the interplay noted by Rajah-Carrim (2008) in MK between *etymological* and *phonemic* spellings. The CG sound /tʃ/ can be written as either as <tz>, etymologically reflecting standard Greek writing, or else as <j>, being a 'phonemic' representation of the sound (Themistocleous, 2010b, 324), which, as in MK, is thus likely influenced by StE graphemic conventions. Such choices are likely to be ideological, marking *distance* and *closeness* from standard Greek and thus how the users of CG perceive their language. Though Themistocleous is not explicitly concerned with conventionalisation, nevertheless we see in her work familiar issues pertaining to this process, and a conventionalisation-focused study of Greek Cypriot non-standard writing would itself be of much interest.

5.3.2 Roman Script Writing of QA: Perspectives & Attitudes

The Roman script writing of QA dialects has been the subject of much study, in no small part due to the innovations employed by speakers of QA dialects in order to communicate effectively without access to the full resources (or indeed, constraints) of their native Arabic script. Of the earliest such works is Berjaoui's (2001) study of regional Moroccan in online

chat based on a database of CMC writing, where Berjaoui determines that emphatic consonants are not typically distinguished in the Roman script writing of Moroccans, with grapheme <s> used for both /sˤ/ and /s/, in addition to a tendency to use a single Roman character to represent geminated consonants, and variation between <j> and <z> for the writing of /ʒ/ (Berjaoui, 2001). Warschauer et al (2002) find that the use of the Arabic script is neglected in online communication by Egyptians, whether it is for the writing of SA or Egyptian QA, replaced instead by a disglossic use of either StE or Egyptian QA written in the Roman script. Palfreyman and Khalil (2007), studying Emirati QA, point out the existence of conventionalised Roman script spellings used in the writing of road signs in the UAE which are written both in the Arabic and Roman scripts. Palfreyman and Khalil call this *Common Latinized Arabic* (CLA), though they also acknowledge that it is not possible to generalise this form across the entirety of the Arab world. This writing must instead be seen within its localised context, considering that Lebanon, for example, uses StF rather than StE as an orthographical basis for the Roman script writing on road signs, such as <Beyrouth> is instead of <Beirut> for the capital. Palfreyman and Khalil themselves also note that /u/ is usually written as <ou> in French-dominated Morocco but as <oo> in the UAE where StE is the primary foreign language, as well as pointing out the duality of StE <sh> and StF <ch>. These national conventions that Palfreyman and Khalil label CLA, though a common sight in the linguistic landscapes of Arabic-speaking countries, form neither an internationally unified code nor a fully-fledged orthography that can be utilised by writers of QA, though we anticipate that each regional or national form of CLA can, through familiarity, contribute to the orthographical productions of users of Roman script QA writing, and thus might play a role (even if it is a limited one) in grassroots conventionalisation.

The resolution for SA and QA sounds that have no Roman script representation is also complicated in the case of CLA, primarily as a result of perceptions of *properness,* where the use of unconventional characters is perceived to be unprofessional in the context of street signage, meaning that for example both glottal /h/ and pharyngeal /ħ/ are both written as <h>, leading to a considerable increase in orthographical ambiguity. These limitations are better resolved in the CMC writing of QA, where the non-standard nature of this writing allows for the use of unorthodox resolutions such as the use of numbers, typically <3> for the voiced pharyngeal fricative /ʕ/, <2> for the glottal stop /ʔ/, and <7> for the unvoiced

pharyngeal fricative /ħ/. Palfreyman and Khalil take the origins of these numerical representations to be (as in the CMC writing of Greek) visual imitations of the corresponding graphemes in the Arabic script, such as <3> being a mirrored visual re-representation of the Arabic character <ع> (Palfreyman and Khalil, 2007). Though some of these resolutions are highly variable on a regional basis, Figure 5.1 below shows the most common instances of numerical representations used in Tripolitan LQA CMCR for the sake of reference (though by the end of our work, we will develop a significantly more nuanced and detailed table; see 11.1.6).

Figure 5.1

Form	IPA	Sound
<3>	**/ʕ/**	voiced pharyngeal fricative
<7>	**/ħ/**	voiceless pharyngeal fricative
<2>	**/ʔ/**	glottal stop
<5> (or <kh>)	**/x/**	voiceless velar fricative
<8> (or <gh>)	**/ɣ/**	voiced velar fricative

The work of Yaghan (2008) is widely cited in discussions of the Roman script writing of QA, which Yaghan labels 'Arabizi' (a portmanteau of <arabi> *"Arabic"* and <englizi> *"English"*). Yaghan characterises Arabizi as variable and contextual, wherein written vowels replace the diacritics of SA writing and can be optionally omitted, governed by factors including clarity (Yaghan, 2008, 42). However Yaghan also makes generalisations about the use of Arabizi which, when conceptually applied to all QA CMC writing, verge on being prescriptivist, such as his clam that gemination is represented with reduplicated consonants (ibid.), which contradicts both Berjaoui's (2001) findings for Moroccan QA as well as our own for LQA CMCR (see 9.2.4). Yaghan attempts to map out all possible Roman characters to their equivalent QA phonemes through a table showing all graphemic resolutions for each, and yet fails to account for example for graphemes that are widely popular in the CMC writing of LQA (and other QA dialects), such as <8> for /ɣ/ (which Yaghan attributes to /kˁ/ instead), made worse by the complete omission of digraphic representation <gh>, instead suggesting the primary reflection of /ɣ/ is <3'> (ibid: 43-44), a grapheme for which we find a grand total of 8 tokens across the thousands of tokens that span both our LQA CMCR datasets. The

shortcomings of Yaghan's attempts at characterising Arabizi stem primarily from the futility of attempting to characterise a highly variable, changeable, non-standard writing used across a vast geographic and cultural space as though it were monolithic in a manner more appropriate for the description of a standardised orthography. Yaghan's work, though often cited as an introductory description of Arabizi, is dated with regards to the very idea that the Roman script writing of Arabic can be generalised for anything more than a single, localised variant, and even within such a variant we expect to find variation that can only be represented probabilistically rather than definitively.

Potentially more problematic still is Yaghan's attempt to superficially resolve the real sociological concerns over the replacement of the native Arabic script with the Roman, particularly in light of the initial colonial attempts to do precisely that (see 1.2.4), which colonial implications are retained in the modern day, even if in an indirect manner as consequences of the resources made available in western technology and more tellingly still the neo-colonial prestige associated with the Roman script. Yaghan reductively dismisses the cultural value assigned to the native Arabic script as merely the product of individuals 'romanticizing about the visual beauty of calligraphy', for which he proposes the solution of simply promoting Roman script typefaces that 'could have an Arabic look', neglecting the real concerns of the major subset of Arabic speakers who value maintaining native linguistic traditions over the adoption of a globalised colonialism (ibid: 47). Such a view runs directly contrary to modern sociolinguistic currents that prioritise the perceptions of speakers, and where self-declared definitions are not challenged by linguistic ones, such as the self-declared perception of SA as a *mother tongue* for many speakers of Arabic not being challenged by academics as *incorrect* on linguistic grounds even while QA is seen to fulfil many of the functions more usually associated with a *mother tongue* (see Albirini, 2016, 33).

As for the users of written QA themselves, we find attitudes that are not unlike those held towards spoken QA reviewed in Chapter 1 (1.3.1), often being similarly negatively charged, primarily on the basis of the disapproval of the use of the Roman script as a 'distortion of the Arabic language' (Albirini, 2016, 276), an attitude informed in part by historical colonial efforts to introduce a Roman script concurrent with the promotion of QA over SA (1.2.4). Yaghan (2008) traces a gradual change in attitude towards the Roman script writing of QA

which he attributes to the widespread use of the internet and the English language by young Arabic speakers, as well as a growing dissociation between the Roman script and the colonial past (Yaghan, 2008, 45). Riegert and Ramsay (2012) add to this a factor of familiarity with the Roman script as a result of its use online and in other media. El-Essawi (2011) cites the positive social image associated with Roman script writing in Egyptian society, reinforcing the views of Warschauer et al (2002) who describe the prestige associated with the use of Arabizi by Egyptians, where in the Egyptian context the ability to read and write the Roman script is closely associated with formal education. For Yaghan, some young users consider the use of 'Arabizi' to be 'cool', alongside advantages of flexibility and even the fact that it is 'free of errors', intuitive and, as a result, concludes that there are 'no typos in this sense' (Yaghan, 2008, 45). Though the literature describing the unrestrictive nature of the Roman script writing of QA seldom makes explicit mention of it, it is specifically the *non-standard* nature of this writing and its existence outside of the SLC paradigm as an uncodified, unstandardised orthography that primarily leads to such positive perceptions of non-prescriptivist expressivity.

5.3.3 Recent Work on the CMC Writing of QA Dialects

The study of the use of QA online has come to primarily revolve around questions of translanguaging and the variable use of QA in contrast with either SA, or indeed with other languages such as StF and StE. Studies such as Taki (2010), Riegert & Ramsay, (2012), and Warschauer et al (2002) focus on whether Arabic speakers predominantly utilise StE or QA in online communication, alongside studies focusing on the interplay between SA and QA, where the QA forms that were once primarily oral now compete with SA within the written domain (Albirini, 2016, 264), with work such as that of Al-Tamimi and Gorgis (2007) and Mimouna (2012) focusing on determining which form of Arabic (SA or QA) is most prevalently used in CMC. In a specifically LQA context, Abdallah (2008) examines the use of Roman script writing alongside other means of identity-construction among a group of Lebanese Christians in Beirut, focusing primarily on how they express identity but taking also into consideration the role of the CMC writing they produce within this. Abdallah also notes the use of written forms that indicate a specifically LQA vernacular, such as LQA CMCR <ekhet> echoing LQA /əxət/ (meaning *"sister"*), as opposed to the transcription the SA form

that would appear as something like <ukht> (see 6.2.4 for a similar discussion in our Tripolitan LQA CMCR context). Interesting work has also been done on the combination of Roman and Arabic script in online Arabic writing, such as Schulthies (2014) who explores the heterogeneity of Arabic CMC writing through analysing YouTube comments and demonstrating the rich variation and cross-orthographic communication of users utilising a wide array of available resources, including their own localised dialectal repertoires (Schulthies, 2014, 55). Taking another approach, Panovic (2018) examines the practice of *script-fusing,* where Roman and Arabic scripts are creatively combined within a single word or expression for purposes of aesthetic but also ideological social indexing (Panovic, 2018, 70). Panovic argues that this is one of the major ways in which the Roman script writing of QA is retained even after Arabic script resources have become fully available within CMC, and sees script-fusing particularly as a marker of cosmopolitan identity and the ambiguous subscription to multiple cultural spheres (ibid: 79).

There also exists, however, a nascent body of literature that examines the specific user-driven choices of individuals in specific QA contexts, driven in part by the work of Abu Elhij'a, though such work unfortunately does not endeavour to break away from the confines of SLC and standard-based approaches to language and writing. Abu Elhij'a (2012) is the most relevant study to our own in this context, examining the Levantine dialects of QA as they are used on Facebook and using data that gives a clear impression of potential conventionalisation and newly emerging conventions, even if this data is not discussed in this context. Abu Elhij'a finds, for example, that 'in some cases [speakers] are not entirely sure how to write a word or letter because the conventions are not fully developed', and goes on to determine that 'this type of confusion is clearly decreasing over time as spelling is becoming more fixed' (Abu Elhij'a, 2012, 73). This view is further supported by Abu Elhij'a's evidence that users do not always write as they speak because (what we understand to be) a conventionalised written form has developed, one example being where SA emphatic /kˁ/ is written as <2> (rather than <k> or <q>), corresponding to the urban Palestinian QA pronunciation of this phoneme as a glottal stop, and yet which convention has also spread to rural speakers who otherwise retain the emphatic pronunciation in their speech (ibid.). In this way, young people 'speak one dialect and write in another' in CMC (ibid: 78), whereby the social pressure of the urban form shows a clear

prestige effect on orthographical productions. It is also important to note that this conventionalised use of <2> occurs in the semi-public context of the Facebook *wall*, whereas retention of the <g> that more closely resembles these individuals' own speech occurs in more intimate synchronous chat messaging (ibid: 79; consistent with our discussion in 5.1.2 III). This is also a gendered convention, where females in Iksal, Palestine who produce /kˁ/ as /g/ in speech nevertheless use the urban <2> form on their Facebook *wall* as it marks feminine prestige, whereas masculine prestige is marked by the use of /g/, and so males use both /g/ in speech and <g> in semi-public Facebook CMC writing (ibid: 78-80; this is consistent with our discussion in 1.3.3, and with the Tripolitan LQA context we discuss in 6.1.2). Abu Elhij'a also finds in what she calls apical pharyngeals (and we call emphatic consonants, /sˁ/, /dˁ/, /tˁ/ and /ðˁ/) 'a clear tendency for people who pharyngealize these sounds less to write them in the same way as they do non-pharyngealized sounds, while those who pharyngealize these sounds more strongly write them differently, using numbers' (ibid: 83), indicating a strong transcriptional link between phonetic and orthographical realisation. She finds that the distinction of emphatic consonants is almost never made in writing in Lebanon and Palestine (where only 3% and 2% of her subjects marked emphatic consonants in any way), though her assertion that this is transcriptionally linked to a lack of pharyngealisation in LQA speech (ibid: 83-85) is one we will challenge, at least in the context of Tripolitan LQA (see 10.3.2). In addition to the relevance of her work on Lebanese QA to our own, by using voice recordings Abu Elhij'a is able to determine differences between spoken and written forms in a way that none of the studies of grassroots conventionalisation have thus far managed, and while she herself does not apply the paradigm of conventionalisation to her work, we will redress this by utilising ourselves a combination of orthographical and phonetic tokens in a similar manner, but firmly within the context of grassroots conventionalisation. While Abu Elhij'a focuses on the break between pronunciation and spelling, she considers instances of individuals not writing as they speak to be a result of complications that lead to confusion (ibid: 100), and because she does not fully engage with the literature of standardisation and conventionalisation, holds the view that 'writing conventions have not fully developed', conflating *conventions* and *codified rules*, the latter of which Abu Elhij'a expects to inevitably develop (ibid: 101). For us, the very break between orthographic and phonetic forms is itself indicative of the grassroots emergence of written conventions, the use of which limits the degree of phonetic

detail produced (in so far as individuals can now choose to no longer *write as they speak*), but which we do not expect to develop into a fully codified system, instead anticipating both transcriptional and conventional writing resources to both be flexibly available in a manner typical of a non-standard writing.

In a follow-up paper, Abu Elhij'a (2014) builds on some of her earlier work, as well as adding a wider scope that includes QA dialects from the Gulf and North Africa. This work, too, maintains the same attitude rooted in SLC that we found in Abu Elhij'a (2012). Abu Elhij'a (2014) begins by drawing a distinction between a dialect and what she calls 'a full-fledged language' on the basis of whether or not the dialect possesses a standardised writing system, with no discussion of the intricacies of the relationship between standard and non-standard (Abu Elhij'a, 2014, 190). She also adheres to the basic disglossic principles of Ferguson (1959b), speaking of a 'gap between the language of literacy' and 'everyday spoken Arabic dialect' (Abu Elhij'a, 2014, 190), absent the considerations we discussed in 1.3.2 with regards to the complex interplay and the disglossic scales developed in more recent academic work. She also draws a comparison between the use of Roman script CMC writing and the establishment of the printing press, which she sees as the means by which the languages of Europe themselves came to be 'full-fledged' (ibid: 191)- a problematic proposition to say the least, within which she fully conflates both writing and standard writing, as well as then conflating both of these with individual concepts of language, standard language and perceptions of language prestige all at once. While she does argue firmly for the importance of ideological factors in the choice of Roman script writing, concluding 'colonialism, prestige and modernity' to be the common factors underlying its use (ibid: 193), we find within this discussion too contentious views, such as the higher-frequency use of the Arabic script in countries like Saudi Arabia being attributed to stronger religiosity and religious identity, predicated as it is on the (unfounded and unsupported) notion that Saudi Arabian Muslims are likely to be more pious (or identify as such) than Muslims in other countries, or even in multi-confessional countries such as Lebanon, where in specific locales such as Tripoli, an identity premised on Islamic piety is a cornerstone of local self-perception. Abu Elhij'a similarly equates the use of the Arabic script with the social meaning of 'being a Muslim', while the use of the Roman script is associated with 'being Christian' (ibid: 193), another problematic conclusion that, at best, is a generalisation based

on what might exist only as a highly localised phenomenon, and so serves as little more than a simplification of a sociolinguistically far more complex web of social meaning and identity-creation, with this view centring quite clearly on the Lebanese capital Beirut with its mixed ethno-religious population, and saying very little not only about the rest of the Arabic-speaking world, but even other places in Lebanon, such as Tripoli and its Muslim majority population that makes extensive use of the Roman script in CMC contexts without any concessions to the self-identification of 'being Muslim'. Abu Elhij'a (2014) identifies age and education as factors in the choice of script, though again the blanket statement that people over 30 'almost exclusively use Arabic' (rather than the Roman script) forms another broad statement that is not consistent with our findings (all our participants in Dataset 2 proved proficient in the use of the Roman script, including thirteen who were aged 26-30, five aged 31-40 and three aged over 41; see 9.1.2 II). Abu Elhij'a also claims that those 'who are over 28 years old frequently use the digraphs <kh> and <gh>' (ibid: 209), which is an unfeasibly precise cut-off point not justified by any historical timeline or process, but instead is most likely a result of her small sample size, given her data derives from five male and five female subjects per country (with additional subjects in cases where data was not sufficient; ibid: 198).

Ultimately, we find here the same issue as we did in the case of Yaghan (2008): broad, blanket descriptions of the use of the Roman script across multifarious social and linguistic communities can never be accurate. Abu Elhij'a (2014) concludes with a table demonstrating the realisation of most major consonants in the different Roman script writings of all of Kuwait, the UAE, Jordan, Lebanon, Palestine, Egypt and Morocco (Abu Elhij'a, 2014, 208), which though doubtless a great improvement over Yaghan's (2008) single table intended to cover all uses of 'Arabizi', it remains the case that meaningful results require a local focus, given how much variation even individual national QA dialects can have (see 1.3.1 and 6.1.2 ahead). The most problematic part of Abu Elhija's table, however, is her clear-cut distinction between proper nouns on the one hand, for which she delineates an invariable use of <h> for the voiceless pharyngeal fricative (and digraphs in the case of the velar fricatives) while purporting that in all other positions <7> is used instead (and numerical graphemes for the velar fricatives)– in direct contradiction to the great variability we find within our own data (see Chapter 8). While we also find that proper nouns and place

names certainly do have an effect on which variant grapheme is used, the situation is very far from one which affords this manner of surgical distinction between where one variant is used and where the other (see 8.2.4.). Ultimately, Abu Elhija's approach is firmly rooted in SLC, which informs her attempts to make clear-cut distinctions for graphemic choice clearly motivated by the desire for the type of invariable classification that is typical of standard ideology, in anticipation of an inevitable 'imposition of unity' upon variation. Indeed, For Abu Elhij'a (2014) it is not only the case (as it was in her 2012 study) that 'writing conventions have not yet fully developed', but she now also adds the statement that this manner of writing is '*not yet standardised*' (ibid: 209, my italics), heavily pregnant as it is with an imminent expectation of inevitability, and so very far from an acceptance of the freely-variable, flexible nature of non-standard writing, which ultimately, and unfortunately, does not allow her to pursue notions of grassroots conventionalisation outside of SLC in a way for which so much of her data and findings are otherwise very highly suitable.

More recent work on the online Roman script writing of LQA has been undertaken in doctoral and masters' theses, such as Bou Tanios (2016) who uses a limited corpus but still produces a table of phonetic-graphemic resolutions for LQA, alike to that of Abu Elhij'a but within which Bou Tanios more readily accepts variation, thus marking another improvement on the original table of Yaghan (2008). We trace in such works a growing interest in the particular study of more specific varieties of QA as written in the Roman script which, in turn, allows for more descriptive and detailed linguistic analysis, and indeed, as a by-product, a growth in dialectological work alongside sociolinguistic analysis of the very kind Horesh and Cotter (2016) have called for (see 1.3.3). These do, however, mostly continue to follow the trend of prescriptivist, SLC-rooted perceptions as set out by Abu Elhij'a's work. There is to the best of my knowledge, no study of the CMC writing of LQA (nor other QA dialects) which takes as its central approach the question of grassroots conventionalisation, despite much of the variation in various QA dialects showing promising potential signs of just such a phenomenon. Our work is unique not only in the clear focus on a single, geographically and culturally emplaced dialectal variant within Lebanese QA that allows for specific analysis, but also in examining conventionalisation in CMC through the otherwise rich field of QA CMC studies, being not only an extension of Deuber and Hinrichs' (2007) *grassroots conventionalisation* in the context of a new language, but also in the context of a

new *kind* of non-standard language given that conventionalisation has not been studied in a Type 2/NSR context such as that of LQA CMCR. Not only do we build upon the aggregational database-based approach of Hinrichs (2004), Deuber and Hinrichs (2007) and the latter half of Rajah-Carrim (2008), but we are also able to fully understand the role of phonetic realisation through our experimental interviews with Tripolitan speakers of LQA which we combine with records of their CMC writing, allowing us to analyse the relationship between the written and spoken realisations of this non-standard language on an individual basis, something which all previous studies have had to approach in a limited and abstract manner, and which Abu Elhij'a (2012, 2014) does with limited participants and without understanding her work through the lens of conventionalisation. Our work is thus finally also unique in its theoretical grounding and the understanding of standardisation that we have developed, allowing us to accurately discuss *conventionalisation* (and by extension, the role it plays within the broader paradigm of standardisation), and thus avoiding the conflation of the two processes. In this way, we are not limited either in our ability to discuss the sociolinguistic phenomena underpinning the use of LQA CMCR, nor our ability to determine what role conventionalisation plays in light of them.

Chapter 6: Preliminary Analysis

We begin this chapter with a summary of the historical and linguistic context of the city of Tripoli (6.1), including a background of spoken Tripolitan LQA and the sociolinguistic realities underpinning its modern-day use. From there follows our preliminary analysis in 6.2, where we discuss the features specific to the non-standard writing system of LQA CMCR and develop an understanding of the primary points of variation, which we then use in our preliminary conclusions (6.3) to construct five research questions to guide the rest of our analysis in the chapters to follow.

6.1 Sociocultural & Sociolinguistic Background

Some discussion in this section will partly consist of anecdotal observation simply due to the fact that topics pertinent to the specific locale of Tripoli have not been covered in any academic literature; nevertheless, references are given wherever they are available.

6.1.1 The City of Tripoli: Recent Historical Context

I. Background

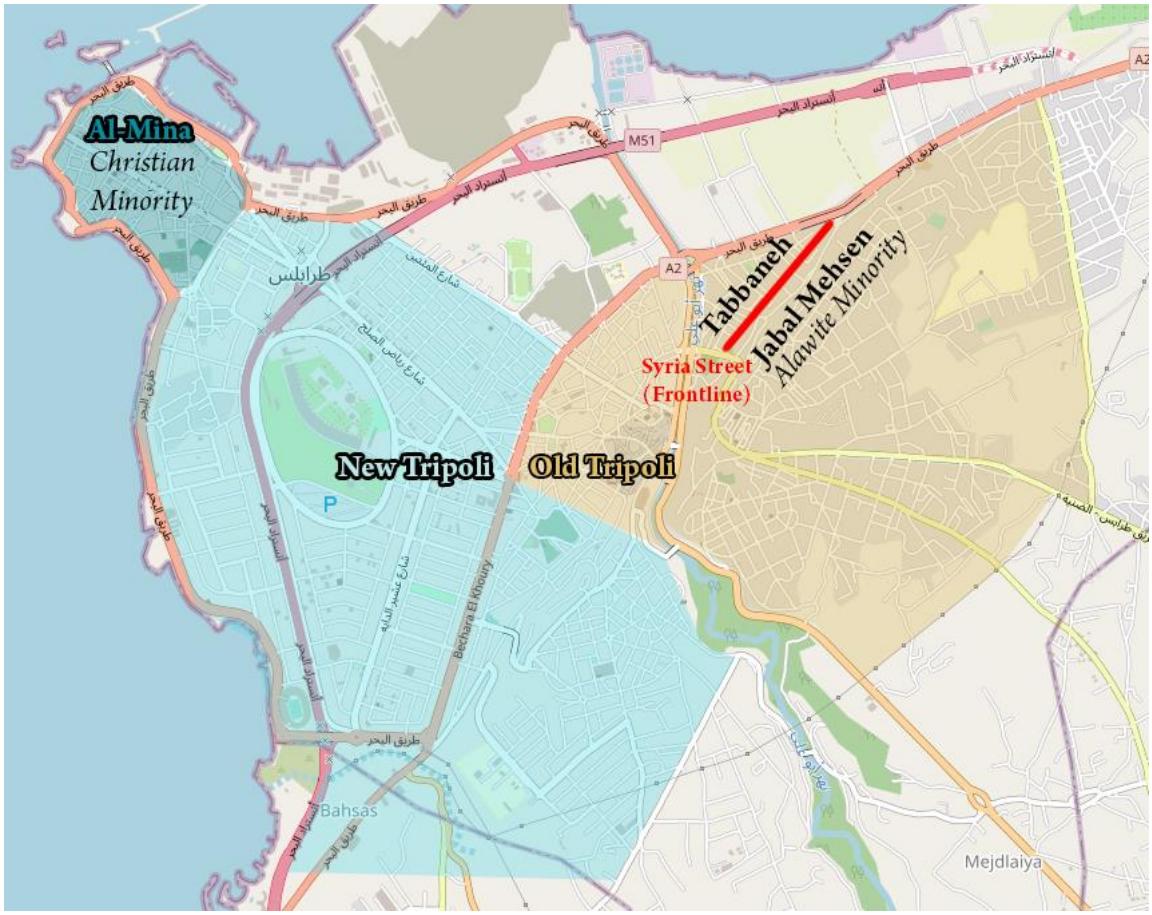
Tripoli is the second largest city in Lebanon, with a population of around 500,000 (Official Website of the Municipality of Tripoli, n.d.) that is predominantly Sunni Muslim (reported at about 80%), with Christian and Alawite minorities (Gade, 2015). Despite being the second largest city in Lebanon, Tripoli has a history of neglect by the centralised state, something it has in common with the entire predominantly Sunni region of the Northern Governate (Volk, 2009). The predominantly Shi'ite Muslim Southern Governate and Beqaa Valley regions have seen similar (and worse) neglect historically (ibid.), though recent geo-political changes have introduced increasing prosperity to those regions, something not replicated for Tripoli and the north of Lebanon. This history has established an underlying sense of inferiority, particularly in comparison to the capital Beirut, as well as manifesting a historical media bias, where coverage of news pertaining to Tripoli has been historically limited only to occasional mention of the gruelling conflicts that the city endured, as well as various Islamist infestations in the older (and poorer) neighbourhoods of the city, resulting in Tripoli being perceived as wild and dangerous by people from other regions of Lebanon. Sociologically, the city can be broadly split into two regions, Old Tripoli (more poverty-

stricken but also more traditional and communal) and New Tripoli (more 'modern' and affluent but also relatively more culturally westernised). The region that comprises Old Tripoli is, by and large, the site of the historical city of Tripoli, with New Tripoli comprising the areas that the city has expanded into, including the district of Al-Mina, which hosts the most part of the Christian minority in the city-proper, whereas the Alawite minority live primarily in a region of Old Tripoli known as Jabal Mehsen. There is a lower level of social contact between residents of Old Tripoli and New Tripoli, with large portions of the two communities living essentially parallel lives on either end of the same city (Seurat, 1985).

II. 2008-2014 Conflict

Spanning roughly from 2008 to 2014, a long series of armed clashes between two neighbourhoods of Old Tripoli occurred, over the course of which the city itself came to be associated with armed strife and instability. The clashes took place between the Sunni Muslim neighbourhood of Bab Al-Tabbaneh (referred to simply as Tabbaneh) and its predominantly Alawite neighbour of Jabal Mehsen (Gade, 2015; Knudsen, 2017). The conflict was more political than ethnic, however, being an extension of the nation-wide political tension between two primary political alliances, known as March 14th (broadly pro-USA and anti-Iran, to which the Sunni Tabbaneh residents subscribe) and March 8th (anti-USA and pro-Iran, to which the Alawite Jabal Mehsen residents subscribe). This was later greatly exacerbated with the beginning of the Syrian Civil War in 2011, which lent the conflict a new dimension as the people of Tabbaneh took an anti-Syrian government stance and those of Jabal Mehsen took a pro-Syrian government stance, both consistent with their respective ethno-political allegiance. The Syrian war was also a boon for fundamentalist Islamist militancy in the Sunni Tabbaneh neighbourhood, including foreign fighters from Syria taking advantage of the then-unregulated borders and crossing into Tripoli. It was during this phase that the conflict reached its zenith, and rounds of fighting would frequently break out, only to die down again until the next bout. The frontline between the two neighbourhoods happened to be a street by the name of Syria Street, an aptly named line of division both metaphorically and physically (Gade, 2015).

Figure 6.1



Base map and data from OpenStreetMap and OpenStreetMap Foundation, available under the Open Database License [https://www.openstreetmap.org/copyright]. *Figure 6.1 above is modified with my own highlighting and labels, accessed from: https://www.openstreetmap.org/#map=14/34.4382/35.8447*

Figure 6.1 shows the broad outlines of the areas of New and Old Tripoli as well as Al-Mina, in addition to the once-warring neighbourhoods of Tabbaneh and Jabal Mehsen within Old Tripoli and towards the outer bounds of the city. The fact that the conflict took place near the outskirts of the old city meant that, for the most part, life went on as usual in New Tripoli, though sounds of gunfire rounds and shelling (particularly when the Lebanese Army intervened in the fighting) could be heard throughout the city as a whole, and as such was experienced– one way or another– by all residents of the city.

III. Lacking Media Coverage & Alternative News Sources

The majority of the conflict described above was not covered by Lebanese media in any detail beyond the basic news that conflict had broken out again in Tripoli. More recently, we can trace a maturation of Lebanese media in the time since the conflict died down in 2014

(after a security initiative finally held and the Lebanese Army was given full authority to enforce it; Knudsen, 2017), where now stories from around the country are more widely reported and in more detail. At the time of the conflict, however, and at the time most of the tokens in Dataset 1 were produced (with the exception of the later data dating to 2015), the conflict was ongoing and remained critically underreported in official Lebanese media. For this reason, a number of Facebook groups became widely popular, dedicated to reporting the details of the conflict in Tripoli on a live basis, including breaking news and details about what was happening and in what parts of the city. These groups, in turn, would become not only sources of information for the residents of Tripoli, but also hubs of discussion, initially centred around the conflict and its news, but in time would also become general spaces of discussion for even otherwise unacquainted Tripolitans to discuss a wide range of local issues in an online, CMC environment. Being Facebook groups, these were not instances of synchronous communication (as they were not live chats), but their subscription-only nature also meant that they were separate from users' primary Facebook profile, as communication that takes place within them would only be seen by other subscribers to the groups, and not the entirety of an individual subscriber's *friends* network. For this reason, the usual expectations of performativity (such as Androutsopoulos, 2015 and Hillewaert, 2015; see 5.1.2 III) in the context of the Facebook *wall* as a public personal space accessible to all Facebook friends is not directly applicable in these instances. These groups form the basis for a significant part of our study, from which we derive our data for Dataset 1.

6.1.2 The Sociolinguistics of Tripoli

The residents of Old Tripoli are more likely to retain a strong historical Tripolitan accent, which has come to signal ill-education and general low prestige among the population of New Tripoli, who for the most part tend to avoid using particularly the very distinctive features of Old Tripolitan, such as the clear change from SA and LQA /a:/ to Old Tripolitan LQA /o:/, as in the pronunciation of the city's name changing from LQA /tˤr**a:**blos/ to Old Tripolitan LQA /tˤr**o:**blos/, which has come to be a by-word for reference to the Old Tripolitan dialect. Citizens of New Tripoli generally speak a dialect closer to the prestige dialect of the capital Beirut than those of Old Tripoli do, though as will become apparent in

our analysis, speakers from New Tripoli sometimes use more Old Tripolitan speech features than they are aware of (see 10.3.4). In recent years, a more clearly gendered split is traceable within the community of New Tripoli, whereby for young male speakers masculinity is expressed through Old Tripolitan LQA, perceived as gruff and macho, and perhaps influenced also by the general perception of the men of Old Tripoli being more masculine, as they retain more traditionally masculine gender roles within Old Tripolitan society. This is consistent with our discussion in 1.3.3 and work by Ibrahim (1986) and Albirini (2016, 197-198) on gender-arranged indexing of identity and prestige, just as it is also consistent with the phenomenon of anti-modernist masculinity in the region, which Zaatari (2015) demonstrates through her study of the historical Syrian TV drama *Bab Al-Hara* (a hugely popular programme followed across the Arab world and certainly in Lebanon) and its nostalgic lionisation of traditionally masculine male figures, often portrayed in anti-colonial roles where their resistance to the French occupation is also an ideological resistance to the cultural concept of western modernity. In the context of Tripoli, this performative masculinity through the use of Old Tripolitan LQA is also likely in part influenced by a lionisation of those who participated in the small-scale civil war that Tripoli underwent between 2008 and 2014, the majority of whom were speakers of Old Tripolitan LQA. On the other hand, young female speakers from New Tripoli tend to emphasise the Beiruti LQA features of their speech, indexing femininity through its perception as being softer and more delicate. This is are consistent with Abu Elhij'a's (2012) own observations of a gendered split in speakers of Palestinian Arabic where Bedouin features are perceived as macho and index masculinity for men, while urban features are perceived as gentle and index femininity for women (Abu Elhij'a, 2012, 78-80, see 5.3.3). I myself have observed in one instance a gathering of a dozen or so young New Tripolitan teenagers at one of Tripoli's many resto- cafés (the primary non-domestic space for socialisation for New Tripolitans), in which the male members of the group were quoting, from memory, lines from a popular viral video in which the tale of Little Red Riding-Hood is retold in an exaggeratedly Old Tripolitan accent; upon recognising the quotations, it became quickly clear that such material was being mimicked, and re-utilised as a source of learning *how to speak Old Tripolitan* in order to employ its masculinity-indexing prestige in a performative manner- this being necessary because the previously taboo nature of Old Tripolitan among prior generations meant that these teenagers did not have access to the accent from their

parents' generation, nor are they likely to be in direct contact with those who do retain the speech aspects of Old Tripolitan in the neighbourhoods of Old Tripoli. In contrast, the female members of the same group spoke in a register markedly alike to that of Beiruti LQA, noticeably more so than one would expect of general New Tripolitan LQA. We might therefore posit a shift in prestige perception among the newer generation, who do not retain the same negative perceptions towards the low-prestige associations of Old Tripolitan that the previous generation does (particularly socially upwardly-mobile lower middle class families seeking to escape a lower class background), but instead the speech of Old Tripoli might be inducted into a gendered duality to represent masculinity for male speakers. Our data derives from too early a time-period to be able to observe this shift in prestige, though we will find at least one instance of performative Old Tripolitan in our later phonetic analysis (10.3.4).

We discussed in Chapter 1 a complex understanding of disglossia, whether it is re-termed 'polyglossia' or otherwise modified without relabelling. Within such a model, H and L functions vary between SA and QA depending on context, and an in-between space is acknowledged, even if it cannot be definitively defined into discrete intermediate stages. The ascription of prestige is therefore more complicated than in the original mode of disglossia, given for example that the use of highly formal SA in spoken contexts is frequently seen as risible. Replicated below are the series of generalised prestige forms for Arabic-speaking communities we developed in Chapter 1 (1.3.3) on the basis of Ibrahim (1986), Vesteegh (2001) and others:

- **Capital:** highest-prestige QA variant per nation; Beiruti LQA, Damascene Syrian QA.
- **Urban:** below *capital,* but above other variants.
- **Rural:** less prestigious, speakers assimilate to urban variants when moving to cities.
  Bedouin: despite low apparent prestige, a notion persists in the Arab world that Bedouin preserves archaic and thus *true* features of Arabic tongue, thus granting it its own prestige.

We are now able to apply this to the specific situation of the city of Tripoli and the various registers of Tripolitan LQA, where the generalised outline is necessarily made more complex:

- **Capital:** the capital dialect of Beiruti LQA, traditionally highest-prestige form overall; for present teenage generation of New Tripolitan speakers, potentially becoming gendered as a feminine register.
- **Urban (New):** New Tripolitan LQA, which is modified with some Beiruti LQA elements, though retains some (non-exaggerated) Tripolitan pronunciation.
- **Urban (Traditional):** Old Tripolitan LQA is traditionally low-prestige among speakers of New Tripolitan, but for Old Tripolitans, the reverse holds true: Old Tripolitan LQA is a prestigious in-group marker for Old Tripolitans to whom New Tripolitan is regarded as an outsider register and negatively associated with Beirut rather than Tripoli. Recently, Old Tripolitan LQA may be in the process of becoming an index for masculinity among young speakers from New Tripoli. Typically Old Tripolitan pronunciations include /o:/ instead of /a:/ (as in /tˤro:blos/, *'Tripoli'*) and the use of /a/ in the stressed word final position where generalised LQA would use a schwa (such as in /baˈħar/ instead of /baˈħər/, *'sea'*).
- **Minority Dialects:** the Alawite population of Tripoli have their own register, and while the Christian minority generally does not, speakers from Al-Mina are perceived to speak with a distinct dialect which is sometimes loosely associated with the Christian minority as they live predominantly in Al-Mina, though the register is generally spoken by residents of Al-Mina irrespective of confessional background, and thus is a geographical rather than ethno-religious register.
- **Rural:** in the case of Tripoli, this includes the dialects of the northern countryside, in which various rural registers exist with a close relation to Old Tripolitan, though they have no real presence within Tripoli itself, where these accents tend to become quickly assimilated with the urban dialects, usually that of Old Tripoli as the economic status of rural-to-urban migrants places them (both geographically and socially) within the Old Tripolitan part of the city.
- **Bedouin:** in the course of the 20th century the traditional Bedouin populations of coastal Lebanon have migrated elsewhere along the coast (mostly into Syria), and

> generally speaking, Bedouin Arabic is no longer a feature of the Lebanese linguistic landscape.

As we see, the Tripolitan context cannot be accurately described with a simplistic scale of prestige such as that of the old disglossic model distinguishing only SA and QA, and moreover, prestige is indexed differently for different communities within the same relatively small city. Foreign languages can also considered part of the same disglossic scale, and in our case both standard French (StF) and standard English (StE) play important roles within Lebanon and LQA, as StF is the historical colonial language of the previous generation and StE, as the second-language of the newer generation, is replacing StF as the second language of Lebanon (Shaaban & Ghaith, 2002). We therefore also understand disglossia as a series of resources available to users in the form of registers, including the various forms of LQA (Old Tripolitan, New Tripolitan and Beiruti), as well as the SA that is still the official language of the government, news channels and newspapers and which is a highly-valued resource available to most speakers, being regularly worked into LQA conversations, particularly ones pertaining to topics such as politics and history, as well as in the use of religious phrases. These resources are drawn upon along with (often limited) repertoires of StF and StE in acts of *translanguaging* (Tagg, 2015, 204) that see sentences constructed out of a series of different available resources, ranging from SA to colloquial forms to foreign languages such as StF and StE. Though we expect our data to come primarily from New Tripolitan speakers, or at least those acquainted with the New Tripolitan social milieu and capable of adhering to its social and linguistic expectations, we can still expect Old Tripolitan features to be used in CMC in a limited manner to give a flavour of Old Tripolitan in the appropriate context, even if most representations remain closer to New Tripolitan or indeed Beiruti LQA, echoing Hinrichs' (2004) description of Jamaican Creole speakers using a limited number of JC words to indicate that the passage in its entirety is to be read as JC and not as StE (Hinrichs, 2004, 94; 5.2.3), though in a Type 2/NSR context, even when Old Tripolitan LQA is not performed orthographically, some degree of LQA (whether New Tripolitan or Beiruti) nevertheless forms the transcriptional basis of CMCR writing.

6.1.3 An Emerging Methodology

I. Research Programme: Overview

Tripoli is a highly suitable city for our study, allowing us to focus our attention on a singular (but complex) sub-type of not only general QA but of LQA too, which enables us to describe the use of the Roman script in a localised, confined sphere where we can also examine interplay with other LQA variants such as urban capital Beiruti LQA. The strong, localised identity of the city along with its active sphere of online communication, in part a result of the underreported troubles in its recent history, makes it ideal for such a study, particularly in the availability of high-frequency online communication between unacquainted members of its population, a great many of whom use the Roman script for writing their native Tripolitan LQA. Through the Facebook groups in question, we are able to examine a specific form of Lebanese colloquial Arabic through a specifically defined community within the population of the city of Tripoli and, moreover, the digital community that this physical population echoes in the CMC sphere. In this way, we can conduct a study of grassroots conventionalisation within a highly specific form of QA and, thus, conduct a novel analysis of a Type 2/NSR orthography. Our primary aim is clear: ascertaining whether there is indeed detectable conventionalisation observable within the non-standard orthography of LQA CMCR and, if there is, what parameters guide it and determine the forms it takes. We begin, therefore, by describing the non-standard aspects of the orthography, including the transcriptional expressivity that we expect of Type 2/NSR writing, and which we can understand in opposition to the dynamic emergence of conventions through the process of grassroots conventionalisation, building upon the work of Hinrichs (2004) and Deuber and Hinrichs (2007), and ultimately developing a still more sophisticated understanding of how CMC conventionalisation takes place through the microcosmic example of the Tripolitan CMC language community.

The study will consist of two primary approaches. The first constitutes a number of user-made comments from the Facebook groups based in Tripoli, written by and for Tripolitans, which we use as a basis for understanding and analysing the origins of many of the points of variation within the writing of LQA CMCR, as well as identifying prevailing patterns within this variation. The two primary datasets for this analysis will be Dataset 0 and Dataset 1, the

first consisting of hand-picked comments from the Facebook groups, allowing us to produce basic orthographical building blocks which will then inform our approach to the richer data of Dataset 1, consisting of comments automatically collected from Facebook via a Python script (and thus consisting of a far greater number of comments), which we use in Chapters 7 and 8 to understand variation, convergence and conventionalisation within LQA CMCR in greater analytical depth. The data of our second mode of analysis (using Dataset 2) will consist of a series of interviews conducted in Tripoli, wherein participants were prompted with phrases and required to re-write them (via smartphone) and read them aloud into a recording microphone, providing us with the ability to better understand the links between the orthographical and phonetic realisations of speakers of Tripolitan LQA. This approach is delineated in further detail in the relevant chapters (Chapter 9 for solely its written data, and Chapter 10 for both the written and spoken data derived from it). Ultimately, we combine the findings from both modes of analysis in order to build our final conclusions regarding the phenomenon of grassroots conventionalisation within the Lebanese dialect of Arabic as used online in Tripoli.

II. Dataset 0 & Dataset 1

The data of Datasets 0 and 1 comes from comments retrieved in 2015 from three of the most popular Facebook groups at the time, which had between 45,000 and 81,000 subscribers respectively at the time of data collection. These groups, by their very nature, see much comment and discussion, and there is frequent (though not exclusive) use of the Roman script for Lebanese QA communication. Other groups existed at the time, not focused on news but Tripoli in general, many of them just as popular in terms of subscribers, but which saw very little comment and discussion because they, unlike the groups utilised, did not update as frequently nor was their content as provocative as to induce individuals to react or interact. The groups we use are the ones that had the most raw data available at the time for the study of the LQA CMCR of Tripoli, serving the entire city and being (even to this day) frequented by many of its citizens, and thus containing long sentences and even paragraphs written in LQA CMCR (alongside Arabic-script LQA CMCA). At the time of data collection, other social networking sites such as Twitter were largely unused by citizens of Tripoli, and to this day generally produce lower quantities of LQA CMCR writing (where users there instead prefer Arabic script LQA CMCA). The anonymity of all users whose

comments we analyse is fully maintained, with no reference throughout the thesis to any identifying features for any commenter or individual. Additionally, permission has been sought and received from the owners of each of the three groups in question for the use of data from their publicly accessible Facebook groups exclusively within this thesis. The preliminary analysis to follow comes from the initial Dataset 0, where data collection was performed manually by copying text directly from the Facebook groups and compiling it in word-processor documents, keeping note of the comment's date, the date it was collected and the group it was collected from; all data gathered in this manner comes from 2014.

6.2 Preliminary Analysis

6.2.1 A Basis for Conventionalisation

We can anticipate the potential function of grassroots conventionalisation within LQA CMCR in a number of ways. Following Hinrichs (2004) and particularly Deuber and Hinrichs (2007), we expect to see certain written forms converging on a limited number of spellings, and in some cases, a single conventionalised written form may emerge with a clear majority. In the case of Jamaican Creole (Hinrichs, 2004) this occurs primarily on the basis of the desire to avoid semantic confusion between related forms, though in this case it is because JC CMC writing is a Type 1/SR orthography, wherein words deriving directly from its English lexifier are more easily confused with JC-exclusive words that developed separate meanings from their standard English etymons. Even in the case of our Type 2/NSR orthography, such conventions could still serve to clarify semantically problematic instances that arise by other causes. Additionally, in Deuber and Hinrichs (2007) we also see that certain orthographical items can converge on a single spelling simply through high frequency of usage, such as the form <mi> in JC for English "*me*", a case clearly not motivated by semantic clarity. The conventions of a non-standard orthography in CMC are also likely to derive from other orthographies, in keeping with Sebba's (2007) view of almost all new orthographies emerging from pre-existing ones (see 4.2). In Type 1/SR instances, especially those of creole and pidgin languages like JC and NP, there is a clear orthographical connection with the lexifier language, though we also see the influence of standard orthographies of other standard languages, such as Yoruba in the case of NP leading to the emergence of conventionalised written forms such as the near-invariant forms <abi> and <sebi>. In our

case, the nearest equivalent would be the influence of SA writing, even if this is complicated by the fact that it uses an entirely different writing system (whereas Yoruba uses the Roman script just as NP CMC does), still we can anticipate certain standard Arabic orthographical conventions to play a role in emerging orthographical convergences in our LQA CMCR.

Similarly, analogous to the role of the lexifier language in Type 1/SR, we expect a similar (though not identical) role to be played by the more distant language or languages that form the basis of the introduction of the new writing system. This is complicated by the fact that there is not necessarily a clearly inherent basis for which particularly Roman script orthography influences the adoption of the Roman script for the CMC writing of an originally non-Roman script language, given that the Roman script is often adopted out of necessity due to it being originally the only script initially available for CMC (see 5.3.1). Nevertheless, in the case of LQA, and given the sociolinguistic situation of Lebanon generally, we can expect the standard orthographies of French (the old colonial language of Lebanon) and English (the new 'global' colonial language of the globalised world) to play a role in furnishing the orthographical sound-symbol correspondences that are put to use by users of LQA CMCR. The example of <dey> and <wey> in NP (Deuber and Hinrichs, 2007) is especially pertinent as a means of understanding how this orthographical basis might function, in that those NP conventions do not take the expected English etymological <there> and <where> forms, but instead derive from an indirect and more broadly conventionalised <ey> form that generally comes to represents the NP sound /e:/ instead. In this way, we too might expect basic sound-symbol correspondences to be borrowed and used as the basis for certain conventionalised spellings, particularly given the lack of any direct linguistic link between StE, StF and our LQA dialect.

Finally, through the work of Abu Elhij'a (2012) we predict another approach to how conventionalisation might be observed in LQA CMCR. Building on her example of <2> (signalling the glottal stop) being used in places where the individuals producing the writing vocalise /g/, we posit another signal of conventionalisation to be the use of written forms that are no longer directly transcriptional representations of an individuals' speech, usually as a result of prestige motivations. This is especially pertinent given that LQA CMCR, being

Type 2/NSR, is untethered to any standard orthography and its writing therefore is expressively transcriptionally; when this type of writing is replaced by forms no longer signifying individuals' own speech, we understand this as a move from transcriptional towards conventional writing. In order to fully explore this particular means of conventionalisation, we require an understanding of individuals' vocalisation of the same words they are writing, and as Dataset 0 (and later, Dataset 1) comprise only of written data, this particular approach to conventionalisation will be returned to in Chapter 10 when we examine the written and spoken data of Dataset 2. On the basis of the expectations for conventionalisation that we have sketched out, therefore, we now form five research questions for the means through which we can anticipate the emergence of written conventions in LQA CMCR:

1. **(How) does high-frequency usage of specific words lead to conventionalised spellings?**
2. **(How) does the need to maintain semantic clarity affect conventionalisation?**
3. **(How) does conventionalisation take place on the basis of the sound-symbol correspondences of the standard English and French orthographies?**
4. **(How) does standard Arabic writing affect the writing of LQA CMCR?**
5. **(How) can we observe conventionalisation on a basis of phoneme-grapheme divergence?**

We begin our investigation by seeking to understand the basis on which variations occurs, examining the grapheme-phoneme correspondences that underpin the writing of LQA CMCR. Our initial analysis of Dataset 0 will be used to determine the building blocks of LQA CMCR as it is used by individuals in the Facebook groups, to serve in turn as the basis for our future analysis using our larger datasets. The rest of this section (6.2) is dedicated to outlining, highlighting and analysing the building blocks of LQA CMCR.

6.2.2 Arabic Orthographic Basis

We see in Dataset 0 some indication of how orthographic conventions can be adopted from SA writing and utilised in non-standard LQA CMCR, despite the two orthographies utilising different scripts. Chief among these is the omission of vowels in writing, leaving it to the reader to determine which vowels are intended from the context of the word or sentence. This is a trait largely unseen in the modern writing of the Roman script (certainly not in StE

or StF), but is a staple of the Arabic semi-Abjadic writing system, wherein long vowels are written but short vowels are unmarked except when diacritics are used (see 1.2.3). In LQA CMCR we thus see the Roman script adapted to the needs and experiences of those utilising it, where familiar conventions impact the use of this new writing system. Some examples of this from Dataset 0 are as follows:

Extract 1[1]	Ya estz 3azzam, lw mank 5nzir mtlo maknt defa3t 3ano eh kel whd bynch2 3n l jech bkoun 5nzir w abou eroun asln tb3oun tbeneh keloun 5nezir mbhbo jech
Extract 1 *missing vowels added*	Ya estez 3azzam, law manak 5anzir metlo makent defa3t 3ano eh kel wahad [*or* wehid] byincha2 3an l jech bkoun 5anzir w abou eroun aslan tab3oun tibeneh keloun 5nezir mabihibo jech

Extract 1 contains words missing up to three vowels, which appear as a cluster of consonants, such as <mbhbo> ("*they don't like*"), realised as something like /ma: biħəb.bu/ in speech. It is nevertheless possible to infer the intended vowels from context, even in initially complex clusters such as this one. On the other hand, for words that appear in forms like <whd>, while the intended meaning is clear (in this case, "*one*"), a distinct phonetic realisation is impossible to distinguish, with two possible variants, /wa:ħad/ or /we:ħɪd/ both being plausible. In this case the potential indeterminate nature of vowel omission leads not to uncertainty of the intended meaning, but rather the specific vernacular variant being used[2]. Vowel omission occurs in the SA orthography following firm, codified rules for where vowels are not written, based on vowel length. The question therefore arises, as to whether vowels in LQA CMCR are omitted in accordance to where they would be unwritten in the SA orthography, therefore indicating the possibility of LQA CMCR being transliteration of the Arabic script, or else whether it is merely the *convention* of vowel omission that has been borrowed, but is used more freely and without strict guidelines.

[1] ***Extract 1***: "*If you weren't a pig like him you wouldn't have defended him, yeah everyone who splits from the army would be a pig and a traitor, and anyway the Tabbeneh guys are all pigs who don't like the army*"

[2] This recalls Hillewaert's (2015) discussion of *indeterminacy* as a strategy to circumvent societal and prestige pressures pertaining to certain vernaculars while still maintaining their (coded) use; though it is outside the scope of this thesis, such an investigation of vowel omission in LQA CMCR (or the CMC of other QA variants) has potential to be of much interest.

Table 6.1

LQA CMCR Text	Arabic Script	SA Transliteration	IPA	Translation
<5nz**i**r>	خنزير	kh-n-z-**i**-r	**/xanzi:r/**	*"pig"*
<bk**ou**n>	بكون يكون	b-k-**u**-n y-k-**u**-n	LQA: **/biku:n/** SA: **/jaku:n/**	*"would be"*

Using Extract 1 from above, we see in Table 6.1 cases where omission does indeed occur in the same place that we expect it to in SA writing. In both the LQA CMCR of Extract 1 and the SA written form, the initial (short) vowel in both words is omitted and the final (long) vowel is depicted (highlighted red). But other words from Extract 1 do not follow the standard convention at all. In Table 6.2 below we see <whd>, where the final short vowel /a/ (or /ə/) is unmarked in SA and in the LQA CMCR of Extract 1, but the initial /a:/ (usually read /e:/ in Tripolitan LQA) *is* omitted in the LQA CMCR text despite being a long vowel and therefore being distinctly written in the SA form of the word. In the case of <mbhbo>, only the final vowel /u/ is marked in the LQA CMCR example (which is marked in SA too), but the initial long vowel /a:/ is omitted in the LQA CMCR example despite being marked in a SA spelling.

Table 6.2

LQA CMCR Text	Arabic Script	SA Transliteration	IPA	Translation
<whd>	واحد	w-**a**-ħ-d	Beiruti LQA: **/wa:ħad/** Tripolitan LQA: **/we:ħɪd/**	*"one"*
<mbhb**o**>	ما بحبوا	m-**a** / b-ħ-b-**u**	**/ma: biħəb.bu:/**	*"they don't like"*

This is not limited to Extract 1, but rather in Tripolitan LQA CMCR more generally, vowels omitted in the Roman script writing do not directly correlate with where they would have been omitted in SA orthography. While the orthographic act of vowel omission is borrowed, the same strict guidelines for when this orthographic omission takes place are not; instead, it is better understood as speakers of LQA being comfortable not writing vowels as a result of their familiarity with unwritten vowels in the SA orthography. As a result, comments with high vowel omission are in fact read similarly to how the Arabic script is, where a reader must substitute in the missing vowels just as they do when reading Arabic (and which is

almost never the case when reading standard Roman script orthographies). In this way, LQA CMCR is to some degree closer to a semi-Abjadic system than most alphabetic Roman script writing is. While the process of reading words with missing vowels is done naturally by readers of Arabic, the fact that LQA CMCR does not retain any system for where it is acceptable (or expected) to omit vowels, there can still be cases that are ambiguous, or at least, that take a longer time to process, requiring sounding out rather than scanning and thus putting readers in the 'decoding-mode' that Jaffe describes for non-standard orthographies. It is also likely that it is not only the non-standard nature of LQA and its non-standard CMCR writing that informs this (sometimes exaggerated) omission, but also the contribution of CMC, many genres of which, as we have seen (in 5.1.2 II) incline users towards non-standard and abbreviated styles of writing. While it is not feasible in this study to determine to which degree omission is a result of the CMC genre or the specific orthography itself, nevertheless it is pertinent to recall the CMC context in which our study necessarily takes place, while also acknowledging that vowel omission of this nature generally does not take place regularly in other CMC writing outside of Arabic, meaning that vowel omission is certainly not solely a feature introduced by the use of CMC. In terms of conventionalisation, we expect this feature to introduce a great deal of variation given that there exist no *rules* for its usage in the CMCR writing of LQA. We examine vowel omission further in 7.3 in the next chapter, where we determine what patterns underpin this phenomenon in LQA CMCR. For now, we address Research Question 4 initially by concluding that vowel omission is the primary SA orthographical borrowing into the Roman script as used for writing LQA CMCR online, and results in indeterminacy rather than either expressivity or conventionalisation.

6.2.3 English & French Orthographic Bases

We find in Dataset 0 a series of sound-symbol correspondences that can be largely distinguished as deriving either from StF or StE orthographical conventions, though the full orthography of LQA CMCR cannot be said to derive directly from the orthography of either StF or StE in the same way that Type 1/SR orthographies do from their lexifier. While Type 1/SR orthographies are to varying extents divergences away from the standard of the lexifier language, the relationship of our Type 2/NSR orthography with the orthographies of

StF and StE is one of borrowed sound-symbol correspondences. We see clearly in Dataset 0 the influence of both, most prominently in the choice of grapheme for the representation of the voiceless palatal-alveolar fricative /ʃ/, between the diagraph <ch> (from StF) or the diagraph <sh> (from StE).

Extract 2[3]	Sara7 ne7na balad kelou ta5alof abl ma ta3rf**ou chou** fi aw **chou** sayer bt7tot**ou** t3li2at bala **ta3me** w kl wa7ad 3am yseb ya3ref enou haydi sifet**ou** abl ma y7ki 3ala 8ayr**ou** ba3den hayda l **groupe** esm**ou c[h]**abaket [*group name*] me**ch ch**abaket a5bar l a3de2 kl wa7ad bired la teni

In Extract 2, the features that appear predominantly French are highlighted in red, where in addition to the use of <ch> or <sh>, other indicators of orthographic origin include the choice between <ou> and <u>. Extract 2, with its abundance of French features, also serves as a preliminary example of how some users use one convention throughout their online speech. We see several instances of <ch> to mark /ʃ/ (and none of <sh>): <chou>, <chabaket> and <mech>, as well as an exclusive use of <ou> to mark /u/ (with no instances of <u>), such as <sifetou>, <8ayrou> and <bt7totou>. We even see an instance of the word <groupe>, used in reference to the Facebook group that the user is posting on, with the word-final <e> an indicator that it is StF that is being drawn upon here for non-Arabic words. Finally, words like <ta3me> might also be considered indicators of StF orthographical rules, where by StE convention it is more usual to use an <eh> instead of <e> in a word-final position to indicate that it is the final letter's own vowel-sound that is being signalled, and not the modification of the quality of the medial written vowel (as in StE *line*, *wine,* etc). In a similar vein, we anticipate the representation of word-final construction /i:n/ to diverge between <in> or <een> using English conventions and <ine> using French conventions. This all becomes clearer still when we look at examples of comments that instead draw primarily on StE conventions in Extract 3 below (with English-derived features highlighted in red), where the exclusive use of <sh> over <ch> demonstrates that the user is utilising a primarily StE-based register: the <shou> used here is the same word as the <chou> used in Extract 2 (meaning "*what*"). Here too we see a case of word-final <eh> for <senneh>, which for the

[3] ***Extract 2:*** "*To be honest we are a country rife with ignorance, before you even know what's happening you write out pointless commentaries, and everyone swearing at others should know that his words describe him, before he says them about others, anyway this group is called [group name redacted], not the 'Enemies News Group', each one hitting back at the other"*

purposes of the final sound is syntactically identical to <ta3me> from Extract 2 for which no final <h> was deemed necessary by the (presumably) StF-informed writer of Extract 2.

Extract 3[4]	whede l (**sh**ahid) wbi**sh**ahedet a**sh**khas bya3rfou **sh**akhsyan mannou lebneny aslan whouwe nxayri men 3ayle 7arbet l senn**eh** bsourya , fik tfasserli kif ken 3am y2atel betrablos?????????? wba2yet (**sh**ouhada l jei**sh**) kellon men l nabatyeh wmen l b2a3, **sh**ou tafsirak???????

In summary, we are functionally able to isolate a number of orthographic variables on the basis of which of the two primary standard orthographical conventions are being utilised, and as a result we become able to understand a large portion of the variation that occurs within the LQA CMCR of Tripoli as variation in which set of sound-symbol correspondences are made use of. Given the relatively weak link between non-standard LQA and the standard languages from which the orthographical associations of LQA CMCR derive, we cannot expect these to appear on a strict or regular basis, and indeed even in the extracts examined in this section (chosen because they demonstrate strong preference for one set of conventions or the other), we still see the occasional use of mixed conventions, such as the use of <shou> in Extract 3, a single word that combines a feature we have defined as English (<sh>) with a feature we have defined as French (<ou>). Nevertheless, understanding the variation that appears in our data in terms of the binary derivation of sound-symbol correspondences forms an important facet of our understanding of the LQA CMCR of Tripoli. We further develop our understanding of the binary nature of the sound-symbol correspondences of LQA CMCR in 7.2 in the following chapter, and so continue to address Research Question 3.

6.2.4 Novel Orthographical Distinctions

In the case of SA, StF and StE orthographical features being utilised (one way or another) within LQA CMCR, the variation we discuss is primarily orthographical in nature, with neither lexical nor phonetic variation aside from the indeterminacy effected by the use of

[4] ***Extract 3:*** "*And this "martyr", by testimony of people who know him personally isn't even Lebanese, he is Nusayri* [slur] *from a family who warred with the Sunnah in Syria, can you explain to me how he was fighting in Tripoli???????? and the rest of the "army martyrs" are all from Nabatiyah and the Beqa'a region, what is your explanation for this?"*

vowel omission. Our understanding of this kind of variation is, therefore, useful for understanding of the overall orthographical structure of LQA CMCR. However, as is often the case with non-standard writing, there are also ways in which users of LQA CMCR are able to utilise the resources at their disposal for a higher expressivity than they might have previously had access to using a standardised orthography such as that of SA. This recalls, for example, the users of non-standard Alsatian writing in Germany as described by Sebba (2007, discussed in 4.3.2), or Siebenhaar's (2006) CMC community of Swiss German users (see 5.1.2 II), who were able to express their local dialects using a modified (Type 1/SR) non-standard orthography, as well as the unregulated and often difficult-to-read nature of such transcriptional writing being the cost at which such expressivity often comes (discussed in the context of Jaffe, 2000, also in 4.3.2). In the case of LQA CMCR, we expect still more expressivity to arise due to its Type 2/NSR nature, in contrast to the non-standard writing of both Swiss German and Alsatian German, both of which remain tethered to the same standard German orthography and within which orthographic deviation is optional. For users of LQA CMCR, however, there exists only broad orthographical bases that are formed variously by StE, StF and SA, but is no single standard writing from which their writing *can* stray. This can theoretically allow for greater expressivity, particularly in cases where the LQA varies from SA in ways that are not possible to represent using the SA orthography. One such example (discussed in 4.1.2) is the sounds /e/ and /e:/[5] for which sound there is no Arabic script representation, and which are thus usually represented even in the non-standard Arabic script writing of LQA CMCA using the same character <ﻲ> that also represents /j/, /i/ and /i:/, or else with the same <ا> that represents /a/ and /a:/. In LQA CMCR, however, this sound becomes possible to distinguish in writing through the use of grapheme <e>. This phoneme is particularly important due to being one of the primary phonetic distinctions between LQA on the one hand and SA and even other QA dialects on the other (such as closely-related Syrian QA, which retains SA /a:/ in many positions where it becomes /e:/ in LQA). Thus, the ability to specify /e/ and /e:/ apart from /a/ and /a:/ in LQA CMCR makes the distinction between Syrian and Lebanese QA in writing possible in many

[5] These are perhaps realised phonetically closer to /ɛ/ and /ɛ:/ in many cases, though we will use /e/ for simplicity as the exact acoustic quality of the sound is of no real consequence to our work.

cases where otherwise the actual dialect specified would not have been possible using the Arabic script, particularly given the closeness of Syrian QA and LQA.

Extract 4[6]	**Sou2 lkhedra ta7et jeme3 l3alli 3ened farouj tallel**

Table 6.3

Lebanese Colloquial Arabic		***Standard Arabic***		
Original (LQA CMCR)	**LQA IPA**	**SA IPA**	**Transliterated SA**	**Arabic Script (SA/CMCA)**
<sou**2**>	/sˤu:**ʔ**/	/su:**k**ˤ/	suu**q**	سوق
<ta**7et**>	/ta**ħət**/	/ta**ħta**/ /ta**ħtu**/	ta**7t**(a) ta**7t**(u)	تحت
<j**e**me3>	/ʒ**e:**mɪʕ/	/ʒ**a:**miʕ/	j**aa**mi3	جامع
<3**ene**d>	/ʕ**ənə**d/	/ʕ**inda**/	3**ind**(a)	عند
<tall**e**l>	/tˤal**e:**l/	/tˤal**a:**l/	tal**aa**l	طلال

Extract 4 above makes abundant use of grapheme <e> to distinguish the LQA sound /e:/, which we break down In Table 6.3, marking the divergent features between LQA and SA in red. The specific manner of colloquial pronunciation is outlined in the Roman script by the writer of the comment, something that near-impossible to depict (or to be inferred by the reader) had the comment been written in the Arabic script rather than the Roman, even in non-standard LQA CMCA. This is in part due to the partially logographic nature of SA writing (see the discussions in 4.1.2 and 4.3.3) as well as the semi-Abjadic nature of the Arabic script where forms like <عند> ([ʕ-N-D]) are pronounced /ʕinda/ in SA but /ʕənd/ or /ʕənəd/ in LQA with no orthographical distinction. It is, however, primarily the availability of grapheme <e> that has the most pronounced effect on dialectal expression, which we also see in Extract 4. The form <j**e**me3> for example cannot be unambiguously expressed in SA writing as distinct from SA form <j**a**me3> (/ʒa:miʕ/), given that both must be written <جامع> (transcribing to <J-A-M-ʕ>, where the Arabic script <A> must signify either /e:/ and /a:/). The same goes for /tˤale:l/, which can only be written identically to SA (and Syrian QA) /tˤala:l/ even in LQA CMCA <طلال>, but can be represented with higher phonetic detail as <tallel> using LQA CMCR. Even <sou2>, the pronunciation of which diverges from SA only in the final

[6] ***Extract 4:*** "*The vegetable market under the 'Alli Mosque, by Farrouj Talal* [Talal's roast chicken shop]"

consonant being realised as a glottal stop instead of an emphatic /kˁ/, is nevertheless conventionally written as <سوق> with the emphatic final consonant retained even in LQA CMCA. In this case, the Arabic writing system *is* equipped to represent a glottal stop final consonant (and does so in other words), but nevertheless even the use of non-standard CMCA writing of LQA is generally conservative wherever possible (primarily for prestige reasons, but also for reasons of readability like those of other Type 1/SR orthographies where communicability is often preferred to transcription that requires decoding). In such cases CMCR is not unique by providing the sound-symbol correspondence in the first place (as is the case for <e>), but rather, for providing a new space within the newly-adopted script where there is no additional perception of lost prestige for using the <2> over <q>, and where the readability of <sou2> and <souq> do not differ significantly in the Roman script.

Extract 5[7]	**asalan** howe seffe7 ma bas **mojrem** w teni chi iza kl li **3emlo** binazaro jihed fa l awla **ya3mol** jihad l akbar lama ykoun 3endo 3ayle w wled

Finally, though it will not be possible to fully discuss phonetic realisation before utilising the voice recordings of Dataset 2, we are nevertheless able to draw basic links between different LQA written forms and the equivalent LQA phonetic realisations of these, all within the context of the unregulated nature of the non-standard CMCR writing system. In the short sentence of Extract 5, we see a few examples of words written with an apparent phonetic realisation in mind, possibly reflecting how the individual might realise the words vocally, given that alternative LQA pronunciations exist for the same words:

Table 6.4

Original Spelling	Indicated Pronunciation	Other Possible Pronunciations
<mojrem>	**/moʒrim/**	/məʒrim / /muʒrim/
<ya3mol>	**/jaʕmol/**	/jaʕməl/ /jəʕmal/

[7] ***Extract 5:*** "*Anyway he's a butcher, not just a murderer, and the second thing is, if everything he did, in his view, was Jihad, why doesn't he first do the Greater Jihad and get a family and children?"*

For the purposes of our preliminary analysis, if we assume the orthographical choices made by this individual transcriptionally reflect their phonetic realisation, it means these orthographical forms not only differentiate this individual's LQA from SA and other QA forms, but in fact also from alternatively possible LQA pronunciations. To which degree this assumption holds true is at the heart of our fifth research question: if such orthographical choices directly reflect pronunciation, they are examples of transcriptional and therefore non-standard and non-conventionalised writing (with subsequently high expressivity). On the other hand, where we find common orthographical forms that diverge from the LQA pronunciations of those producing them, we observe a weakening of transcriptional writing and therefore the potential emergence of conventional forms. This will be the primary focus of Chapter 10, in which we will use the experimental data of Dataset 2 to ask (and answer) this question, addressing our fifth research question.

6.2.5 Lost Orthographical Distinctions

I. Distinctions Retained with Novel Solutions

We close this section with a discussion of the orthographical challenges that arise in the switch to the Roman script, and the solutions that users of LQA CMCR reach to resolve them, forming the rest of the underlying structure of the orthography of LQA CMCR. LQA sounds with no unambiguous graphemic representation in the Roman script are most often written using numerals based on broad similarities between the shapes of the numbers and the original corresponding Arabic letters (as occurs in the Roman script writing of most QA dialects; see 5.3.1), and which have largely come to be accepted and immediately recognisable as stand-ins for those missing letters. Most common among these are the numerical grapheme <3> for writing /ʕ/, <2> for writing /ʔ/ and <7> for writing /ħ/. The velar fricatives /x/ and /ɣ/ are typically represented either with digraphs (<kh> and <gh>) or numbers (<5> and <8>) respectively, leading to a purely orthographical point of variation depending on which form is preferred. There is high variation in the representation of the voiceless pharyngeal fricative /ħ/ as a result of the additional use of <h> for the same sound, meaning that grapheme <7> represents only /ħ/, and phoneme /h/ is represented only with the grapheme <h>, but the grapheme <h> itself can indicate either /ħ/ or the voiceless glottal fricative /h/, introducing a new point of ambiguity. In below extract below, we see

<h> used by the same individual to first signify /ħ/ in the word <yehmiyon> (/jəħmij.jon/), and then to signify /h/ in the word <bhal> (/bhal/):

Extract 6[8]	Alla ye**h**miyon b**h**al ta2os w y2awiyon

Further still, we see in Extract 7 below examples of all three variations in a single sentence. In the first instance, the user uses the <7> to signify /ħ/ in the word **<7a2>** (/ħaʔ/), then immediately uses <h> instead to signify the same /ħ/ sound in the word **<nihna>** (/nəħna/), and finally goes on to use <h> to signify /h/ instead of /ħ/ at the beginning of the word **<hal>** (/hal/).

Extract 7[9]	ma3ak **7**a2 ni**h**na.kilna.joubna law mana joubna.mawslet mwaselna.**h**al seni.lahoun

The variable representation of the voiceless pharyngeal fricative and its overlap with the voiceless glottal fricative is a major source of orthographical variation within LQA CMCR. A key question will be whether we are able to discern patterns behind the choice of <h> or <7>, and within that possible conventionalised forms (or conventionalised positions for the use of one or the other). We examine this feature further in Chapter 7, using our first two research questions (high-frequency usage and an inclination to maintaining semantic clarity) to probe for emerging conventions within the writing of this specific sound. For the time being, we understand the use of <7> as another point of variation within LQA CMCR, and alongside the other means of representing sounds not catered to by the Roman script we further understand the make-up of this non-standard orthography, which is comprised on the one hand of adopted orthographical features from SA, StE and StF (some of which allow for distinctions specific to LQA to be made, such as the availability of <e>), and on the other hand made up of novel representations of sounds, with the case of the voiceless pharyngeal

[8] ***Extract 6:*** *"May God protect them in this weather and strengthen them"*

[9] ***Extract 7:*** "*You're right, we are all cowards, if we weren't cowards we wouldn't have reached the state we did this year"*

fricative (and to a lesser degree, the binary representations of the velar fricatives) leading to further variation within its writing.

II. Distinctions No Longer Made

Finally, not all phonetic differences distinguished in spoken LQA are represented in the writing of LQA CMCR. Most notably missing is the distinction of emphatic consonants from their non-emphatic forms, which are not specified in the writing of LQA CMCR. Abu Elhij'a (2012) records minor instances of the use in online LQA writing of <S> (capitalised) for emphatic /$s^{ʿ}$/ and <T> (capitalised) for emphatic /$t^{ʿ}$/. These appear minimally in Abu Elhij'a's data, however (<S> showing a single token and <T> two; Abu Elhij'a, 2012, 84), consistent with our own data which in fact shows zero instances of the emphatic consonants being represented at all. Abu Elhij'a states that the emphatic forms are generally not used even in the speech of urban speakers of LQA, and while it may be true that the distinction has weakened in the case of the Beiruti LQA dialect of the capital, we certainly expect speakers of the LQA spoken in Tripoli to maintain the distinction of emphatics phonetically, even if they do not do so orthographically when using LQA CMCR, which we demonstrate using our recorded data in 10.3.2.

6.3 Preliminary Conclusions

We have built in our preliminary analysis a framework for the analysis to follow in the rest of the thesis. We have developed an initial understanding of the building blocks that form the primary points of variation within LQA CMCR, and have seen that a large quantity of variation occurs on an orthographic basis, based on the various sound-symbol correspondence resources available for users of LQA CMCR, and additionally where there are competing novel solutions for sounds with problematic or ambiguous representation. Our primary approach to conventionalisation will be focus therefore on a phonemic-graphemic rather than a fully lexical basis, though we will also examine the lexical results of this graphemic variation and, potentially, the reduction of this variation, where conventionalisation is to be understood in effect through the potential resolution of these points of variation. We recall our research questions as a framework for understanding this potential conventionalisation:

1. **(How) does high-frequency usage of specific words lead to conventionalised spellings?**
2. **(How) does the need to maintain semantic clarity affect conventionalisation?**
3. **(How) does conventionalisation take place on the basis of the sound-symbol correspondences of the standard English and French orthographies?**
4. **(How) does standard Arabic writing affect the writing of LQA CMCR?**
5. **(How) can we observe conventionalisation on a basis of phoneme-grapheme divergence?**

The chapter to follow will focus on Research Questions 3 and 4, and the sound-symbol correspondences adopted from various standard orthographies. At first sight, the borrowings from Roman-script orthographies (RQ3) lead not to conventionalisation, but rather the opposite: further variation. This is the result of the fact that there are two discrete orthographies whose conventions are borrowed, and though we might be able to discern conventions emerging within the context of one orthography or the other, the fact that both feature within LQA CMCR complicates our understanding of conventionalisation within the non-standard orthography. To examine this relationship further, we must better understand how these conventions function, and ascertain whether individuals tend to use one convention or the other, or else to what extent these conventions are mixed. One possible path to conventionalisation would therefore be to envision sub-orthographies, one based on StE orthography and the other on StF, which though unorthodox, would make for a novel and possibly unique division of variation within a non-standard system, and this discussion will form the basis of the Chapter 7 to follow. Similarly, and unlike for example the influence of Yoruba on NP, where standard written forms are adopted from standard Yoruba and developed into conventionalised usages by users of NP, the borrowing of SA orthographic conventions (RQ4) results in the opposite: increased ambiguity, in a large part because of nature of the convention that is borrowed, though it is also the change in script from Arabic to Roman that means that this convention is borrowed as a general resource (that is, as the *capacity* for individuals to simply not display vowels), without the strict orthographical rules that this convention follows in its SA usage. The change in script also means that, unlike NP using StE orthographical forms, it is not possible for users of Type 2/NSR LQA CMCR to borrow full forms as they are spelt in the Arabic script, but must transliterate them, which process itself leads to further variation. To further examine any

conventionalisation that could take place on this basis, we further examine the use of vowel omission in LQA CMCR and determine whether there are any emergent rules governing its use by Tripolitans online. This too will be done in Chapter 7 to follow, examined in the same context of sub-orthographical variation at the heart of that chapter.

We examine thereafter in Chapter 8 the novel orthographic solutions for the LQA sounds with no Roman script representation, and in particular the variation caused by the variable representations of the voiceless pharyngeal fricative, using Research Questions 1 and 2 (word-frequency and semantic differentiation) as guidelines. This approach will also allow us to devise a new means for understanding orthographical variation, both that which results from these particular variant features as well as a general means of understanding all variation within LQA CMCR, as well as how conventionalisation might take place among users of the orthography. Finally, the Roman script allows for certain colloquial features of LQA to be represented for the first time in writing using LQA CMCR, particularly the vowels /e/ and /e:/ which have no representation in the Arabic script, meaning that LQA CMCR allows for an expressivity not possible using even the non-standard CMCA form written in the Arabic script. In this, too, however, variation is introduced through the question of whether- and where- these vowels are represented, a question dependent both on phonetic as well as orthographical motivations. This is the crux of our final Research Question (RQ5): we have seen that at least some of the orthographic variation potentially reflects *phonetic* variation too, and clarifying that distinction will form an important avenue for both understanding the composition of the orthography as well as determining how conventionalisation functions within it. The question of phonetic realisation and its role within the non-standard orthography (and RQ5 more generally) will be examined in Chapter 10, using both the audio recordings as well as textual data of Dataset 2, though only after we first use the written data of Dataset 2 in Chapter 9 to review and further build upon our work in Chapters 7 and 8.

Chapter 7: The Sub-Orthographical Model

We now begin our further analysis of the major points of variation that we have identified in our preliminary analysis, with an aim to determining the specific patterns underpinning them within the broader orthography of LQA CMCR. This includes determining the rates at which each variant occurs, as well as the underlying factors influencing the choices of individuals for any one orthographic form over others, which will afford us a deeper understanding of the variational structure underpinning users' choices within the orthography, and will allow us to formulate a framework for observing consistency within on the basis of the workings of these features.

3. **(How) does conventionalisation take place on the basis of the sound-symbol correspondences of the standard English and French orthographies?**
4. **(How) does standard Arabic writing affect the writing of LQA CMCR?**

In this this chapter we address specifically our third and fourth research questions (reproduced above), focusing on the variation introduced through the borrowing of features from standard orthographies (Roman-script StE and StF, and Arabic-script SA), on the basis of which we formulate a sub-orthographical model aimed at explaining some degree of the variation in LQA CMCR on the basis of two discrete orthographical veins that run through it. To this end, we focus on features derived from the StE and StF orthographies in 7.2, and on vowel omission as adapted from SA writing in 7.3, finally concluding with a review of our sub-orthographical model in 7.4. In this way, we develop at once an understanding of these features and variations and how they function while simultaneously positing framework for how we might understand conventionalisation within this variation.

7.1 Methodology for Dataset 1

In this chapter we use Dataset 1, which consists of some 8,000 comments (comprising about 25,000 tokens) collected from the same Facebook groups used for Dataset 0. The data collection for Dataset 1 was automated, using a script written in Python to gather data from the source code of the Facebook pages in question, extracting whatever comments are visible on the page, which Facebook limits in number until the page is scrolled down, at which point more (older) posts and comments are loaded. This process was automated

using a web driver (Selenium), controlling the web browser and simulating the scrolling down (in this case by simulating the pressing of the Page Down key), causing the page to scroll down and load more comments. The code utilises XPath to select nodes from the source code of the page, looking for elements which start with the relevant ID for comments, and then prints the comment itself and the date and time that accompanies it. The program collected data in this way from the three Facebook pages discussed at the beginning of Chapter 6 (6.1.3) and stored the comments along with the relevant metadata (anonymised usernames, time, date, and Facebook group name). This comment corpus comes from a period between 2012 and 2015 (during which time the internecine Tripolitan conflict discussed in 6.1.1 II of Chapter 6 was at its height). Dataset 1 will be used for our analysis in this and the next chapter, as well as being called back upon in Chapters 9 and 10.

7.2 The English and French Orthographical Modes

7.2.1 Defining Two Convention Groups

The distinction between the orthographical correspondences in LQA CMCR derived variously from StE or StF can be feasibly seen to split– to a certain extent– the writing of LQA CMCR into two halves, each comprised of conventions adopted from each respective orthography. In our preliminary analysis (6.2.3) we observed a tendency for some users to utilise conventions deriving from one orthography or the other, which we now examine further. As determined previously, we consider <sh> and <u> spellings to be the English-derived counterparts of French-derived <ch> and <ou>. The LQA word realised as /ʃu:/ (meaning "*what?*"), is comprised solely of these two sounds, and so provides an ideal marker for determining how these sounds couple in the orthographical realisations of users of LQA CMCR.

Table 7.1 - *"What"*

		Harmonic	
Mixed	**French**	**English**	Mixed
<shou>	**<chou>**	**<shu>**	**<chu>**
21	49	45	25
15%	**35%**	**32%**	18%

The four major variant spellings of the word *"what"* are composed of different arrangements of the four orthographical variants available (<sh> and <ch> for /ʃ/, <u> and <ou> for /u:/). The harmonic forms highlighted are those in which both conventions used derive from the same standard orthography, for which we immediately see evidence of a coupling effect between French <ch> and <ou> spellings, and likewise for English <sh> and <u> spellings, with StF-based <chou> appearing at 35% and StE-based <shu> at 32% total frequency, each roughly twice as frequently as the mixed forms, which appear at 15% and 18%. These mixed forms (comprising of one StE and one StF grapheme), though less common than harmonic forms, still appear frequently enough to indicate that a sizable proportion of users do not draw directly from conventions deriving from a distinct language and for whom features from both exist as potentially interchangeable resources. These resources, therefore, all exist within the orthography of LQA CMCR itself, and, for some users at least, are no longer necessarily accessed directly from a prior knowledge of or familiarity with StF or StE orthography. In this way, we understand the split between harmonic and mixed forms to reflect what repertoire users of LQA CMCR are drawing upon for their conventions, allowing us to reach the preliminary conclusion that those using harmonic forms consistently are likelier to be drawing directly from the conventions of the standard orthography in question than those using mixed conventions, who in turn are likelier to be accessing them from the resource pool of LQA CMCR itself (containing the totality of all four forms as available resources). As we see in Table 7.1, the two harmonic forms are roughly twice as popular than the two mixed forms, and yet with a largely even split between most popular harmonic forms themselves. In this way, we might better understand conventionalisation by categorising each of the different conventions separately as sub-orthographies– one based on the sound-symbol correspondences of the StF and the other of StE. To investigate this connection further, we examine how these features match together across all words, and not just the marker word <chou>/<shu>. To do this, we use the representation of the phoneme /u/ as a basis for selecting comments which utilise only <u> for one set, and comments that only utilise <ou> for the other set, within which we then examine the frequency of the graphemic representations of /ʃ/ across all words that appear within either of these.

Table 7.2 – Representation of **/ʃ/** in comments using **French <ou>** exclusively

/ʃ/	Tokens	%	
<sh>	**291**	47%	*English (Mixed)*
<ch>	**328**	**53%**	***French (Harmonic)***

Table 7.3 – Representation of **/ʃ/** in comments using **English <u>** exclusively

/ʃ/	Tokens	%	
<sh>	**172**	**63%**	***English (Harmonic)***
<ch>	**101**	37%	*French (Mixed)*

We see in Table 7.2 that the orthographic convention defining each set (French <ou> or English <u>) predicts which representation of /ʃ/ (French <ch> or English <sh>) will appear most frequently, meaning that the tendency for harmonic forms to align does not only apply for the word <shu>/<chou> alone but is also reflected across all words that appear within our two sub-sets. We also see, however, that this effect is more pronounced in the English <u> group (where 63% of comments also used StE <sh>), as opposed to the French <ou> group, within which only a bare majority of 53% comments also consisted of harmonic French <ch>. This discrepancy, however, is better understood in the context of the overall popularity across all our data of the form <sh> as compared with that of <ch>:

Table 7.4 – Representation of **/ʃ/** across full Dataset 1

	Tokens	%
<sh>	782	**55%**
<ch>	635	**45%**

As <sh> is the more popular representation of /ʃ/ in the entirety of Dataset 1 (at 10% higher frequency than <ch>), the fact that <ch> shows any majority at all (even at 53%) in the French harmonic group in Table 7.2 above is meaningful, as the impact of selecting comments for our <ou>-only set actually reverses the general trend of the use of this grapheme. In fact, selecting for comments on the basis of whether they utilise <u> or <ou> demonstrates an identical effect on the ratio of both representations of /ʃ/: <ch> rises by 8% from its 45% overall frequency to 53%, and so too does <sh> rises by 8% from its overall frequency of 55% to 63%.

Table 7.5 – Change in Representation of /ʃ/ in Sub-Sets vs. Overall Data

	Overall Data	Exclusive Comments	Change
<sh>	**55%**	**63%**	**8%**
<ch>	**45%**	**53%**	**8%**

This is at once indication of the harmonic effect, as well as a demonstration of the limited extent of this harmonic effect, given that there is still a sizable minority of non-harmonic usage: 47% of /ʃ/ tokens are written with non-harmonic <sh> even in comments exclusively using French-derived <ou>, and 37% of /ʃ/ tokens are written with non-harmonic <ch> even in comments exclusively using English-derived <u>. We now further filter our comments to include only those which adhere strictly to one set of conventions across *both* key forms /ʃ/ and /u/, creating two fully-harmonic sub-groups of comments. Within these we find a total of 209 comments that use both <ch> and <ou> (and never <sh> nor <u>), and a total of 331 comments that use both <sh> and <u> (and never <ch> nor <ou>). These groups, which we will refer to by the short-hand notation <*shu> (for the English-harmonic group) and <*chou> (for the French-harmonic group) are composed of the following totals for each specific feature:

Table 7.6 – Harmonic Sub-Group Grapheme Frequencies

***French – Harmonic* <*chou>**

<ch>	316
<ou>	426
Total	**744**

***English – Harmonic* <*shu>**

<sh>	459
<u>	177
Total	**638**

The lower number of total tokens in the English group is unsurprising considering that <u> is a less popular form compared to <ou> across all data, in the same way that we observed <sh> to appear at a rate of 55:45 compared to <ch> (Table 7.4). Moreover, <u> is sufficiently less common than <ou> across all of Dataset 1 that it overcomes the higher <sh> to <ch> ratio in favour of English <sh>, and thus means that the English-harmonic sub-has a total of 638 tokens compared to the 744 of the French-harmonic sub-group – even while the English comment group consist of 122 more comments than the French one. The fact that these groups consist of a total of only 540 comments means that about 7% of our comments

feature harmonic forms exclusively, thus meaning that the primary tendency of users of LQA CMCR is to mix together these forms, in addition to which we must also consider that some of the shorter comments within these groups are likely to have a representation of only /ʃ/ or /u/ and not necessarily both. The limited extent of fully-harmonic usage is also evident in the overall tendency we have determined for /ʃ/ and /u/ across the full data, whereby the highest-popularity graphemic representation for each phoneme derives from a separate standard orthography, as we summarise below:

Table 7.7 – Frequencies for /ʃ/ and /u/ across full Dataset 1

		Tokens	%			Tokens	%
English	**<sh>**	**782**	**55%**	*English*	**<u>**	915	38%
French	**<ch>**	635	45%	*French*	**<ou>**	**1,466**	**62%**

That <ch> is only marginally less popular than <sh>, compared to the far greater prevalence of <ou> compared to <u> is what leads to the English-harmonic sub-group to return less total tokens than the French equivalent did. In this way, we understand there to be two primary effects at work: the first is the individual, non-harmonic popularity of respective forms <sh> and <ou> as seen in Table 7.7 above, while the second is the clear clustering effect of harmonic forms that is sufficient to overturn the overall popularity of representing /ʃ/ when we select for comments using either <ch> or <sh>, as we saw in Table 7.2. This is the same effect we also saw in Table 7.1, where the word /ʃu:/ also shows a very clear harmonic clustering effect in spite of the overall individual popularity of <sh> and <ou>. In this way, we note two separate, almost entirely opposed effects underpinning the orthographic choices of users of LQA CMCR: one of admixture (which we hypothesise to indicate the integration of these sound-symbol correspondences within the LQA CMCR repertoire), and another of harmonic exclusivity (whereby some users continue to draw directly from StE or StF). The low percentage of comments which show fully harmonic usage, however, indicates that admixture is generally more widespread than harmonic exclusivity. We return to the discussion of these two distinct currents underlying the choice of features at the end of the chapter; before that, we end this section by using our harmonic sub-groups to consider how other features of LQA CMCR correspond with the StE and StF split, before moving on to examine our next primary feature (vowel omission) in 7.3, which

we will then also consider within the basis of the harmonic split between StF and StE forms, giving us a strong basis with which to finalise our overall sub-orthographical approach in 7.4

7.2.2 Cross-Feature Analysis across the Convention Groups

I. Velar Fricatives as Numerals vs. Digraphs

We hypothesise that digraphs <kh> and <gh> will be less popular representations of the velar fricatives for those utilising StF-based orthographic conventions, given that these digraphs are rarely used in StF writing, and so meaning that numerical representations <5> and <8> will be preferred (for /x/ and /ɣ/). Testing for representations <8> and <gh> for /ɣ/ within the two fully-harmonic sub-groups we have devised, we see at first glance that the data does not support the hypothesis for the voiced velar fricative; rather, the reverse is true, as <8> appears at a higher frequency to <gh> (showing 45%) in the English-based <*shu> group, while it shows only 35% frequency in the French-based <*chou> group. Here we note that the total number of tokens (45) is low.

Table 7.8 - Harmonic Sub-Groups

	<*chou> group		**<*shu> group**	
<8>	8	**35%**	10	**45%**
<gh>	15	**65%**	12	**55%**

For the voiceless velar fricative /x/, however, we find much higher total of 201 tokens, and so are able to observe a significant preference (at 78%) for numeric <5> over digraphic <kh> in the French-harmonic <*chou> sub-group, while both resolutions are evenly split in the English-harmonic <*shu> group, confirming that numeric representations of the velar fricatives couple with the French sub-group (though the reverse is not true for digraphic forms within the English sub-group, which are evenly split with numerical ones).

Table 7.9 – Harmonic Sub-Groups

	<*chou> group		**<*shu> group**	
<5>	91	**78%**	44	**52%**
<kh>	25	**22%**	41	**48%**

In effect, this indicates a further layer of distinction even within the highly selective groups we have devised: within the English-harmonic <*shu> group, there are variable representations of /x/, drawn largely equally from the repertoire of LQA CMCR, while in the French <*chou> group, a proportion of individuals also draw from the general features available but there is a higher preference for forms which are more in line with the conventions of StF writing. We hypothesise that we would find the same effect for the voiced velar fricative given enough tokens, and review this connection using Dataset 2 in Chapter 9 (9.2.3).

II. The Voiceless Pharyngeal Fricative

The voiceless pharyngeal fricative /ħ/ is represented in LQA CMCR variously with either <h> (ambiguous) or <7> (specific). This feature and its representation will be the focus of Chapter 8 to follow, but for the time being it is most relevant for us to consider the overall split between the use of <7> and <h> to represent /ħ/ shows a consistent ratio of 71:29 (<7>:<h>) across the entirety of Dataset 1. With this in mind, we can test for how the ratio varies within our sub-groups:

Table 7.10 – Harmonic Sub-Groups

	<*chou> group		**<*shu> group**	
<h>	27	**13%**	53	**21%**
<7>	181	**87%**	200	**79%**

The French-harmonic <*chou> sub-group sees a noticeably higher percentage of <7> to <h> (87:13 compared to the overall 71:29 expected), but the <*shu> group also shows a higher percentage of <7> than is predicted (79:21 compared to overall 71:29). This may appear at first glance to reflect the same preference for numerical representation within French <*chou> group that we observed for the velar fricatives, though the alternative representation of <h> is not digraphic (and as per our hypothesis, not problematic within StF writing), and moreover as we will examine in depth in the next chapter, the representation of the voiceless pharyngeal fricative is impacted by a wide array of additional factors. We instead posit here that there exists a general preference for numerical resolutions among French <*chou> users, while noting also that that both groups have higher percentages of <7> compared to <h> than is predicted by the overall ratio for the full data, potentially

explained by the fact that the sub-groups themselves consist of comments filtered for their specificity in usage, a by-product of which might be higher specificity, such as specifying the voiceless pharyngeal fricative with <7> rather than an ambiguous <h>.

III. <ine> vs. <in>/<een>

Finally we consider another orthographical featured hypothesised (in 6.2.3) to be potentially linked with the orthographical associations of StF and StF, that being the representation of words ending with the sound /i:n/. A StF-derived convention is thus hypothesised to be <ine> (as in *tartine, vitrine,* etc), which orthographical form appears in StE more usually signifying an /aɪ/ sound (as in *wine, alpine, supine*, etc), making it less congruent for writing /i:n/, for which we would expect either <in> or <een> instead. As we see in Tables 7.11 and 7.12 below, though the ending <ine> does not appear with any great frequency, it is still slightly more common in the French <*chou> group (where it also appears once for schwa instead of /i:/) than it is in the English <*shu> group, particularly taking into account the fact that the English-harmonic group <*shu> has some 150 more comments than that of French group <*chou>.

Table 7.11 – /i:n/ in English-Harmonic Sub-Group <*shu>

Token	IPA	*Tokens*	
<mshawbine>	/mʃawb**i:n**/	*1*	*"We are hot"*
<mine>	/m**i:n**/	*1*	*"Who"*

Table 7.12 – /i:n/ in French-Harmonic Sub-Group <*chou>

Token	IPA	*Tokens*	
<alemeddine>	/ʕalaməd.d**i:n**/	*1*	*"Alameddine* [family name]*"*
<mazloumine>	/mazˤlum.m**i:n**/	*1*	*"They are wronged, hard done by"*
<ibine>	/əbən/	*1*	*"Son of"*
<earfine>	/ʕa:rf**i:n**/	*1*	*"They know, they knew"*

IV. Summary

We have developed an understanding of how the sound-symbol correspondences derived from the StF and StE orthographies inform the orthographical choices of users of LQA CMCR, and through our examination of the specifically filtered harmonic groups we have also understood how other features of LQA CMCR interface with the potential sub-

orthographical strands running throughout the orthography, even if we note that there is even within these sub-groups a significant degree of variation, which is in addition to the fact that these precise and specific harmonic groupings are strictly adhered to by a relatively small sub-section of individuals. Nevertheless, the links remain strong enough to overturn the overall popularity of English <u> and French <ch> individually to the degree where <chou> and <shu> are twice as popular as the mixed forms are, despite the fact that the overall most popular representations of /ʃ/ and /u/ (StE <sh> and StF <ou>) are non-harmonic. Now we turn to the phenomenon of vowel omission, first examining the general function of the phenomenon, before considering how it fits within the sub-orthographical division of our convention groups, after which we will return to the sub-orthographical split overall and consider to what degree we can understand variation and conventionalisation using this model.

7.3 An Arabic Convention: Vowel Omission

7.3.1 Analysing the Convention

I. Vowel Length & Word-Frequency

We hypothesised the tendency among users of LQA CMCR to omit vowels to derive from the orthographic rules of SA (in which short vowels are not marked except by optional diacritics). We also noted in our preliminary analysis (6.2.2) that in the case of LQA CMCR, it is not only short vowels that would be unwritten in SA that are omitted, which raises the question of whether omission is better understood as a short-hand form that emerges due to the genre of synchronous (or semi-synchronous) CMC, rather than being directly derived from SA writing as we hypothesise. The length of the vowels of LQA can be broadly split into three groups, in the context of which we will examine vowel omission within Dataset 1. LQA retains most of the distinctions between long and short vowels made in SA and SA writing, such as the distinction between LQA words /ħal/ (meaning *"solution"*) and /ħa:l/ (meaning *"situation"*). The third vowel-length grouping derives from certain SA vowels dropping to schwa in LQA, such as SA /kul/ ("*every, each"*) becoming LQA /kəl/, or SA /min/ becoming LQA /mən/. The nature of Dataset 1 means that we have a limited number of forms showing vowel omission with a requisite number of repeated tokens for analysis, and the words with

the highest omission rates as well as plentiful tokens are single-vowel high-frequency forms that we collect in the following tables:

Table 7.13 – Vowel Omission

A. Schwa

	<e>	<i>	<ø>	<ø>%	Tokens
/wəl/	20	8	11	***28%***	39
/kəl/	46	8	21	***28%***	75
/bəl/	56	16	45	***38%***	117
/mən/	83	47	79	***38%***	209
Total	**205**	**79**	**156**		
	47%	18%	**35%**		

B. Short Vowels

	<a>	<ø>	<ø>%	Tokens
/law/	15	1	***6%***	16
/ʕal/	58	12	***17%***	70
/ʕam/	100	27	***21%***	127
/bas/	95	39	***29%***	134
/hal/	159	10	***6%***	169
Total	**427**	**89**		
	83%	**17%**		

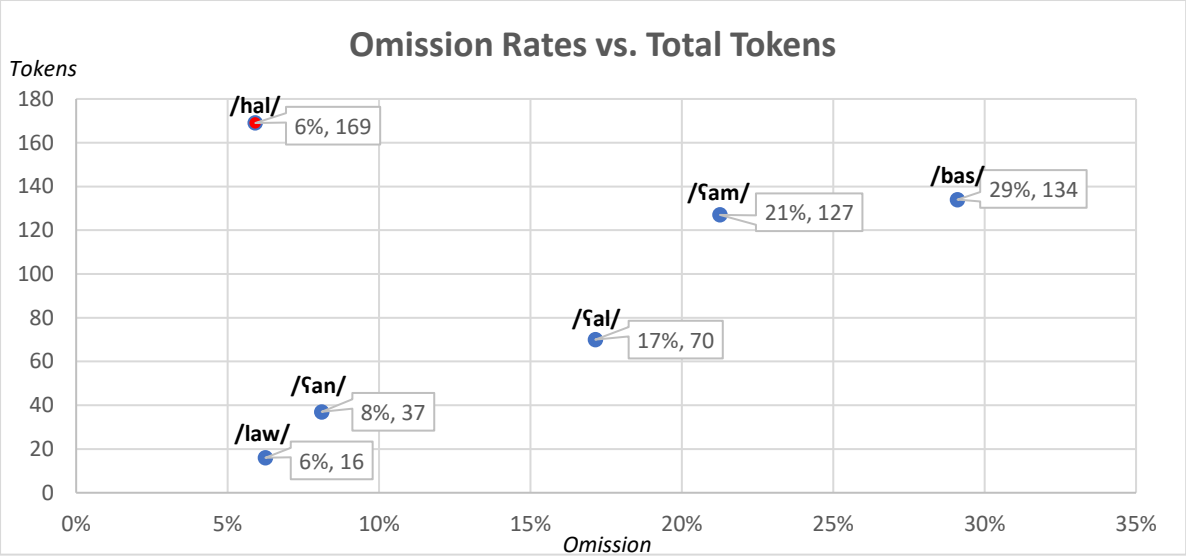
C. Long Vowels

	<V>	<ø>	<ø>%	Tokens
/he:k/	70	6	***8%***	*76*
/kti:r/	70	3	***4%***	*73*
Total	**140**	**9**		
	94%	**6%**		

In the schwa group (7.13A), omission ranges from 28% (for *"and the"* /wəl/ and *"every"* /kəl/) to 38% (for *"from"* /mən/ and *"in the"* /bəl/). The high similarity in internal structure for these words indicates that rate of omission does not vary on morphological grounds; instead, the token totals for each word give a better indication for rate of omission, as the 28% omission words show 39 and 75 tokens respectively, while the 38% omission words show 117 and 209 tokens respectively, indicating instead a frequency effect, whereby the words likely to be used more frequently overall within LQA CMCR (and not necessarily only in Dataset 1) see higher omission due to the familiarity of the words (recalling the frequency effect of RQ1) and thus the lower risk of omission causing ambiguity (recalling the semantic clarity effect of RQ2, both effects being ones we also address in the context of the voiceless pharyngeal fricative in the next chapter). In the case of short vowels (7.13B), it is striking that all such words with significant omission rates consist of the short vowel /a/, and as in

7.13A, all these are also short, simple, single-vowel utility words, most of which see high-frequency usage. With the exception of /hal/, which appears to be an outlier, we otherwise again find a near-perfect scaling between word-frequency and omission rates, ranging from /law/ ("*if, if only*") at 16 tokens and 6%, rising proportionally all the way through to /bas/ (*"but, only"*) at 134 tokens and 29% omission, again demonstrating that omission rates are higher for words that are more familiar through high-frequency use, which we also see charted below:

Figure 7.13D



The clear outlier here is /hal/, showing only 6% omission despite consisting of the highest number of tokens at 169. Here it is most likely the potential semantic ambiguity (RQ2) that the omitted form <hl> results in, which can be misread as /ħəl/ (*"solve",* or more commonly used in various colloquial forms essentially meaning *"buzz off",* as in /ħəl ʕan.ni/), or even as /ho:l/ ("*those, them* [m.]"). As such, while high-frequency use generally leads to higher omission rates, this effect can be overturned by the risk of semantic ambiguity as in the case of /hal/ (which same word we revisit using Dataset 2 in 9.3.2 IV, Table 9.27, and where we find the very same effect). The short vowels of Table 7.13 see an average omission rate of 17%, though this appears at 21% if we remove the data for /hal/; either figure of omission for short vowels (17% or 21%), compared to the 35% for the schwa set, turn demonstrates clearly that omission rates overall also vary on the basis of our vowel length groupings. This is further demonstrated in the very minimal omission we find for long vowels, which are

omitted more than once in only two words, for a total of 6% omission across all tokens of these words (/he:k/ *"like this, like so"* and /kti:r/ "*much, many*"), in addition to which other forms show only one-off omission, such as the single token <nkhod> for /ne:xod/ (*"we take"*). As such, we conclude that word-omission- at least in the case of these specific, high-frequency, high-omission words- can be seen as an emerging convention which in the first instance is predicted by the length of the vowel in question, and in cases of high-frequency use across LQA CMCR, and thus high overall familiarity, higher omission can be expected to take place, except where omission leads to semantic ambiguity such as in the case of /hal/, which overturns both effects and as a result actually shows omission consistent with that for the long-vowel set rather than the short-vowel set to which it belongs. To what degree this effect holds up in lower frequency words and longer, more morphologically complex words, and indeed whether (and how) it affects vowels other than schwa and /a/ will be addressed in Chapter 9 (9.3.3) using the richer data of Dataset 2.

II. An Adapted Convention or a CMC Short-Hand?

These generalised rules for when we can expect omission to take place (and at what rates) bear little resemblance to the strict orthographical rules determining where vowels are or are not written in the SA orthography; rather, as hypothesised conventions within a non-standard orthography, they are not *rules* but rather indications of the frequency at which we can expect omission to take place for different types of vowels. We have hypothesised that this convention itself as used in LQA CMCR derives from the semi-Abjadic writing of the standard Arabic script (see 1.2.3), but has been adapted loosely within the Roman script writing of LQA CMCR, likely also encouraged by the synchronous or semi-synchronous genres of the CMC it is usually utilised within, but which does not arise solely as a form of CMC-induced short-hand convention (such as English <thx> or <k>). To confirm this hypothesis, we look for correlation between the omission that occurs in our common words (in Tables 7.13A and 7.13B above) and generalised omission in the further writing of individuals who utilise these common words. The following extracts are split into two groups, the first consisting of high schwa omission and the second of high short-vowel omission, with the common forms shown in bold and red, while other omitted words are only emboldened.

Table 7.14A – Schwa-Omission Extracts

Extract 1[10]	Fi 3alam ma **by3jba** l 3ajab la en sar shi mni7 wala en sar shi **bsh3** 7terna ya ar3a **mn** wen badna nbosik
Extract 2[11]	shi ktir 7elou **wmhm bl** 7ayet lzm kilna n3aml aya insen **bhl** tari2a in **kna** mna3rfou aw la2 la2n ra7 yiji yom w **nkhod** dawron **bhl** 7ayet w yiji min **ys3dna**
Extract 3[12]	Aslan tripoli dayi3 7a2a,**mtl kl** manati2 lbnenlk **lw** 3anjad fi dawla ken **lbnen** 3anjad jana 3al ard! **bs** l7a2i2a houwa jana **bs** bnazarna **cz mn7b** nshoufo **hk**!!

Table 7.14B – Short Vowel-Omission Extracts

Extract 4[13]	fiyi **a3rf** lesh samm l badannnnnnn **3l** sobohhhh ya e5tiii alla uhani sa3id b sa3ideee trekini **a3rf** etrawa2 **hl** ka3ke 3a rawa2 la7awla wala kuwata ela bellahhhhh
Extract 5[14]	tfeh ! Shu hal araf la2 **wl** shabeb lmu7taramin **3m ytfrju** w allu trekon ntfaraj ya 3aybshummm
Extract 6[15]	**L72oni 3l mstshfa**

We see in Table 7.14A a high concentration of our common schwa words from Table 7.13A such as <mn> and <bl>, along with a high quantity of omission across other words also, including longer words such as <wmhm> (/wə məhəm/, "*and is important*"), <by3jba> (/bjəʕʒəba/, *"is pleasing to her"*) and <ys3dna> (/js**e:**ʕədna/, *"help us, aid us"*), showing omission both for schwa and non-schwa sounds, including some long vowels such as /e:/. In

[10] ***Extract 1:*** *"Some people are never pleased, whether something good or something bad happens, I really don't know what you all want"*

[11] ***Extract 2:*** *"That's something very good and important in life, for all of us to deal with other people in this way, whether we know them or not, because a day will come when we'll be in their place in life, and someone will come to our aid"*

[12] ***Extract 3:*** *"Anyway, there is no justice for Tripoli, like all the other regions in Lebanon, if there really was a government then Lebanon would honestly be heaven on earth! But the truth is, it's only a heaven in our eyes, because we like to see it that way!!"*

[13] ***Extract 4:*** *"Can someone tell me what's the point of this negativity so early in the morning, man, if they're happy then let them be and let me eat this ka'akeh* [pastry] *in peace; there is no might nor power except in Allah!"*

[14] ***Extract 5:*** *"Disgusting! What is this vileness, and what's more these 'respectable' youth are just watching, he even said 'leave them be, let's just watch', how shameful!"*

[15] ***Extract 6:*** *"Follow me to the hospital"*

Table 7.14B, we see instances of common short vowel forms from Table 7.13B such as <3n> and <3l>, alongside which we also find omission of both schwa and non-schwa vowels, as in <ytfrju> (/jətf**a**r.r**a**ʒu/,*"they watch, spectate"*) and <mstshfa> (/məst**a**ʃfa/, *"hospital*"). In this way we see that even while the common forms of Table 7.13 are the most common and highest-omission forms, vowel omission is a phenomenon that takes place across the writing of LQA CMCR, and we posit for the time being that the generalised rules for omission rates we have devised hold true throughout our data, though we will take this discussion up again in 9.3.3 using the data of Dataset 2 to examine vowel omission more precisely for various vowels and vowel-positions outside of only the common forms.

7.3.2 Vowel Omission in the Convention Groups

We are now in a position to examine what role vowel omission plays within the two convention groups we devised in the first half of this chapter (in 7.2), and whether these omission tendencies differ meaningfully between those utilising StF-based and StE-based sound-symbol correspondences. In this way, we intersect the two orthographical divisions we have made thus far, combining the two harmonic sub-groups with the three vowel-type categories we have now developed.

Table 7.15A – Schwa Vowels in the Sub-Groups

French Harmonic

<Ø>	<e>	<i>
25	32	19
33%	42%	25%

English Harmonic

<Ø>	<e>	<i>
31	27	13
44%	38%	18%

Table 7.15B – Short Vowels in the Sub-Groups

French Harmonic

<Ø>	<a>
10	80
11%	89%

English Harmonic

<Ø>	<a>
22	57
28%	72%

Table 7.15C – Long Vowels in the Sub-Groups

French Harmonic

<Ø>	<V>
0	8
0%	100%

English Harmonic

<Ø>	<V>
2	8
20%	80%

Recalling that our sub-groups consist of only a small sub-section of the overall comments of Dataset 1 (and thus the total forms for each sub-group will only be a sub-section of the total forms shown in Table 7.13), we see a clear pattern emerge wherein comments from the English-harmonic sub-group show higher rates of omission for all three omission categories: 44% for schwa (compared to 33% in the French-harmonic group), and 28% for short vowels (compared to 11%). In the case of long vowels, the percentage is not significant considering the low number of overall tokens for the long-vowel words within the comments, though it is still notable that two instances of long-vowel omission take place in the English-harmonic group versus zero in the French-harmonic group. Though the convention of vowel omission takes place within both sub-groups, it is more pronounced among users that draw exclusively on StE orthographical conventions. As neither StE nor StF feature vowel omission in their standard orthographies, this is best understood in the context of those utilising StE-derived written conventions likely being younger than those who use StF-derived ones (see 6.1.2), and as such are more likely to make use of what is (in the Roman script at least) an unconventional convention. Ultimately, we understand higher omission to be a feature of our proposed StE-derived sub-orthography, with lower omission more characteristic of the StF-derived sub-orthography, even if omission is present in both, as well as across our entire data.

7.4 The Uses and Limitations of Sub-Orthographies

On the basis of the analysis that we have conducted in this chapter, we now posit two sub-orthographical convention groups, which though are not always fully adhered to individually, can be considered two primary distinct orthographical strands running through the writing of LQA CMCR. In the following conceptual sentence, written once with the characteristics of each harmonic groups, we can see the extent of the differences between these two convention groups:

Table 7.16

	1	2	3	4	5	6	7	8
French	**Chou**	**hal**	**cha8leh**	**ma**	**3arfine**	**chi**	**ya**	**a5i**
English	**Shu**	**hl**	**shaghleh**	**ma**	**3arfeen**	**shi**	**ya**	**akhi**

This sentence (translating roughly to *"what's the deal with this, we have no idea what's going on man!"*) showcases the primary points of difference that we have determined, diverging on the basis of whether an individual uses orthographical conventions derived strictly from StF or StE, along with the other choices that we have found to correlate with this sub-division. Position 1 consists of the core of the split, encapsulated in <shu> (English <sh> and <u>) and <chou> (French <ch> and <ou>). In position 2, the higher tendency to omit vowels (including short vowels) by those using exclusively StE features is reflected by the use of <hl> in the English sub-orthography and <hal> in the French sub-orthography. In positions 3 and 8, we see the higher indications of numerical graphemes <8> and <5> in the StF-derived writing, as opposed to digraphs <gh> and <kh> which users of StE-derived conventions utilise more frequently. Finally, we see in position 5 alternation between the form <ine> of the French convention and its English <in> for the word-final /i:n/ sound. In this way, we understand a certain degree of the variation we have examined so far to not be random, but instead occurring as a result of conventions anchored in standard orthographies external to that of LQA CMCR, broadly divisible into these two sub-conventions, which we envisage to be the two primary strands along which we are able to arrange a certain degree of the orthographical variation we have examined thus far, even if the vast majority of users of LQA CMCR will mix conventions from both strands rather than adhere strictly to one or the other (as users of a standard orthography might be expected to do).

Ultimately, however, the fact that the majority of users utilise features from both sub-orthographies means that our proposed sub-orthographies are better imagined as an etymological backdrop to the non-standard orthography of LQA CMCR, where these strands are the roots of the variants available in the repertoires of LQA CMCR users, and within which only a minority still draw exclusively from conventions of one or the other. In this way, we instead consider those belonging to the strictly harmonic groups to be the LQA CMCR users who retain a closer connection to the external standard orthographies of StF and StE and continue draw from their sound-symbol correspondences, whereas the greater majority, who use features of both, derive their orthographical forms from within the rich but variable pool of orthographical resources adapted into the repertoire of LQA CMCR through a longer-term grassroots process, rooted in the etymological basis of StE and StF

writing without retaining the strict boundaries between the two as discrete and unrelated systems. This means that, rather than demonstrating any newly-arising conventions, our strictly devised sub-groups are instead representative of the initial conventions with which the non-standard orthography of LQA CMCR began. Conventionalisation here is thus understood not through the dissection of the orthography into sub-orthographical strands that might be expected to show higher uniformity at any meaningful level, but rather the opposite: conventionalisation is instead the very grassroots process of the adoption of these conventions into a single repertoire. This is best encapsulated in the fact that we find StF <ou> but StE <sh> to be the two graphemic resolutions with the highest respective frequency. Conventionalisation is thus the admixture of features and the newly-emerging popularity of individual conventions from within both, and the emergence of preferred graphemic resolutions irrespective of their etymological orthographical source. This will be the focus of the next chapter, where we will take an approach based on overall frequency of appearance on a phonemic-graphemic level, but our work in this chapter has nevertheless been crucial for understanding the structure of much of the variation that takes place within LQA CMCR, and our sub-orthographical division provides a useful way of understanding the roots of much of the variation we find, through the loose reconstruction of hypothetical past boundaries that are gradually eroded through the process of conventionalisation. We have also developed a keen understanding of some of the core underlying orthographical points of variation within the writing of LQA CMCR, which arise as direct result of LQA CMCR being what we have defined as a Type 2/NSR orthography with no single orthographical anchor to draw upon, instead deriving conventions from the standard orthographies of unrelated languages as well as drawing on SA writing in an asymmetrical, trans-scriptural manner through features such as vowel omission introduced to the Roman script writing of LQA CMCR. It is the Type 2/NSR nature of LQA CMCR that explains why our work so far looks very different to that of Hinrichs (2004), Deuber and Hinrichs (2007) and even Rajah-Carrim (2008), all of whom examined the conventionalisation of Type 1/SR non-standard writing, closely tethered to (or, depending on context, untethering from) standard English.

We have thus far addressed (to various extents) the first four of our Research Questions, most emphatically RQ3 concerning the effect of the StF and StE orthographies on the writing of LQA CMCR, as well as the effect of SA writing (RQ4) through the adapted

convention of vowel omission, though which we also began to address RQ1 in the high-frequency (schwa or short) single-vowel words that showed greater vowel omission than their lower-frequency counterparts, while the motivation to maintain semantic clarity (RQ2) also coloured our understanding of why the word <hal>, despite its high frequency, nevertheless saw a very low rate of vowel omission. In the next chapter we extensively analyse the occurrence of the voiceless pharyngeal fricative and further address RQ1 and RQ2. This will be followed by In Chapter 9 in which we will use Dataset 2 to re-examine the points of variance that we have discussed to that point and to further develop our new model for conventionalisation, while in Chapter 10 we use the full extent of the new Dataset 2 (including audio recordings) in order to answer RQ5 and so devise yet another approach to our understanding of how conventionalisation functions.

Chapter 8: The Lexemic-Aggregational Model

8.1 Revisiting Research Questions & A New Outlook

The voiceless pharyngeal fricative /ħ/ in LQA CMCR is either rendered with the numeral <7>, which is used exclusively for this sound, or the letter <h>, which is also used to represent the voiceless glottal fricative /h/. Given that <7> maps onto /ħ/ alone, while <h> maps onto two distinct sounds /ħ/ and /h/, this feature is a point of much interest and variation, and will be the central point of discussion of this chapter, in the course of which we use this variation to examine the various means by which we can anticipate conventionalisation to occur, as well as developing a second framework for approaching and understanding grassroots conventionalisation more generally across all features on the basis of word-frequency. In this way we primarily address RQ1 (the effect of high-frequency usage) and RQ2 (the role of semantic clarity), in addition to other factors that arise in the context of the examination of the variation induced by the variable and overlapping representations of the voiceless pharyngeal and glottal fricatives. We hypothesise on the basis of our understanding of conventionalisation thus far four primary factors likely to influence users' choice between <7> and <h> to represent the voiceless pharyngeal fricative, which follow below.

I. Use in Other Languages, Place Names & Proper Nouns

A widely-used and readily recognisable spelling is likely to impact the orthographical choices of users of LQA CMCR, including words that appear as place names or proper nouns with a relatively high frequency. This recalls what Palfreyman and Khalil (2007) call "Common Latinized Arabic" (CLA), being the conventionalised Roman script spelling of Arabic in street signs and other official usages (5.3.2), though CLA is highly variable across different Arabic-speaking countries, and in Lebanon, this CLA is based on StF rather than StE orthographical conventions. Personal and brand names are also frequently written in contexts where the use of numerical symbols is either not prestigious or else simply not possible, and thus such words might be expected to tend more towards <h> over <7>. Examples include the word Helweh (/ħəlwe/, "*nice, pretty* [*fem*.]") which also happens to be a surname; the city of Haifa (/ħajfa/, in Occupied Palestine), and Hallab (colloquially pronounced /ħəlle:b/ in LQA), the oriental dessert-makers hugely popular in Lebanon and the region. The same goes for

words that occur with high frequency in other languages, recalling the standard Yoruba spellings that are adopted by speakers of non-standard Nigerian Pidgin. In our case, it is not the SA (that is analogous to standard Yoruba) which is likeliest to have this effect, but potentially StE, given its use of the same Roman script as LQA CMCR, which could potentially lend conventionalised forms in cases where SA (or even LQA) forms see use in StE writing. Examples of this would include words like <Allah> and <Mohammad>, or even colloquial forms like <habibi>, which is relatively frequently written in colloquial English, even if it is not standardised, and for which we hypothesise a higher usage of <h> over <7> partly because of this. Abu Elhij'a (2014) determined (in Lebanon and elsewhere) that <h> is exclusively used to represent proper nouns, and <7> exclusively used in all other contexts (see 5.3.3), though we will find a great deal more variation than her in our own work to follow.

II. Semantic Clarity

Where there is no distinct semantic meaning for the use the voiceless glottal /h/ (instead of pharyngeal /ħ/) fricative in the same position, such as again in /habi:bi/, the pressure to distinguish the sound with <7> rather than using <h> will likely be diminished, recalling in particular the use of orthographical differentiation for distinguishing homographic semantic variants by users of Jamaican Creole (Hinrichs, 2004; see 5.2.2 I). Conversely, we anticipate higher frequency use of <7> in cases where potential ambiguity is introduced by equally valid variant readings of words where either /h/ or /ħ/ can be realised in the same position (e.g./ħar/ "*spicy*" and /har/ "*crumbled*"). This is the basis for our second research question, meaning that the analysis of the realisation of the voiceless pharyngeal fricative in this way allows us to address RQ2 directly. We note here also that for words like <habibi>, we already find two potential factors influencing users' orthographical realisation: it is both a word frequently used outside of LQA CMCR, as well as having no semantically ambiguous reading if the sound is read as /h/ (and a third, given that it is a word likely to show high frequency of appearance). We therefore anticipate this manner of convergence of factors to provide us with the clearest examples of these pressures at play. We examine semantic clarity in LQA CMCR in 8.2.3, noting also that while motivations of semantic clarity held for Jamaican Creole in Hinrichs (2004) and Deuber and Hinrichs (2007), they did not for Nigerian Pidgin in Deuber and Hinrichs (2007).

III. Frequency of Use

Words that see high-frequency use can be expected to lower the tendency of users of LQA CMCR to write a specific <7> over ambiguous <h>. The more commonly a form appears, the more easily recognisable it becomes for readers and the less need there potentially is to specify it through marking the sound with <7>. This is an effect we have already encountered in the context of vowel omission (7.3.1), where high-frequency words saw higher rates of potentially ambiguous vowel omission. In this sense, <7> becomes a special marker only strictly necessary to indicate a particular form when it is not familiar, with <h> otherwise used. This, by extension, also follows the general rules of conventionalisation discussed by Deuber and Hinrichs (2007), in which NP words are observed to develop conventionalised forms simply through high-frequency usage, even where other orthographical pressures such as semantic clarity were not necessarily present, and it is on this basis that we formed our RQ1, which we shall address in this chapter and this context, starting in 8.2.1.

IV. Word Position

Finally, a purely linguistic factor we must consider is whether the choice of <h> or <7> is influenced by the position within the word at which the sound occurs. Considering the use of <h> at the end of a word is sometimes used to mark that the final vowel is sounded, we expect this position to show the greatest impact on this sound. This discussion follows in 8.2.2.

8.2 Analysing the Voiceless Pharyngeal Fricative

8.2.1 Frequency of Appearance

Across Dataset 1 we find a total of **1,071** tokens in which the phoneme /ħ/ is indicated, out of which **764** represent the sound with <7>, and **307** represent the sound with <h>. This comes to a ratio of **71:29** for <7> to <h>, and, as will become apparent in the course of this analysis, this is the golden ratio for understanding the bi-graphemic variation for this phoneme. We can therefore use this ratio as an anchor, significant variation from which can be understood within the context of the effect of one of our hypothesised factors. In terms of word-frequency analysis, we hypothesise that deviation in the ratio away from <7> and in

favour of ambiguous grapheme <h> is expected in the case of the words that appear most frequently in our data, as users of LQA CMCR are expected to perceive a lowered need for specification in these words.

Table 8.1 – Ratio of <7> to <h> in Dataset 1

<7>	**764**	**71%**
<h>	**307**	**29%**
Total Tokens	*1,071*	

I. Word-Set Examination

Given the high number of total tokens and the fact that many tokens for the same words appear in variant forms (varying in gender, person, plural, and so on), we can group together all tokens that represent variant forms of the same word. Where there is a low number of tokens for each individual variant, we are thus able to examine a more meaningful quantity of tokens in this way. We then tally the total <7> forms versus the total <h> forms for each word-set, giving us an indication of the variation between the two graphemes. Below is a simple example, showing the full lexical data that one of these word-sets is composed of:

Table 8.2 - Word-Set Breakdown - *"Good"* /mni:ħ/

Total	**<7> Form**	**Tokens**	**%**	**%**	**Tokens**	**<h> Form**	
15	mni7	10	67%	33%	5	mnih	*"Good* [m.]"
3	mni7a	2	67%	33%	1	mniha	*"Good* [f.]"
1	mne7	1	100%	0%	0	mneh	*"Good* [pl.]"
19	***mni7**	**13**	**68%**	**32%**	**6**	***mnih**	

The <h> and <7> that we are testing for are highlighted in red, a blue background is used for all <h> forms, and a light orange background for all <7> forms. In the total tally (and our general labelling of the word-set), <*mni7> and <*mnih> are marked with an asterisk to indicate that this is a generalised form representing the overall word-set, and not necessarily a representation of every token that appears within it. We see for each form both a token count as well as a percentage showing the frequency of a particular form relative to its direct equivalent as written with the other grapheme (so the 10 tokens of precise form <mni7> show a frequency of 67% relative to the 5 of <mnih>, which in turn shows a percentage of 33%). We also see the same at the bottom for the tallied totals of the

full word-set under <*mni7> and <*mnih>. In this case, in addition to the most popular forms <mni7>/<mnih> after which we named the generalised word-set, we also see <mni7a>/<mniha> (the feminine form /mni:ħa/), and <mne7> (plural form /mne:ħ/). Gender variation (as in this example) is just one of the ways in which these forms can vary, and the most popular form in this case is masculine <*mni7>/<*mnih>, though this is again not always the case, depending on context. We can summarise the above table as follows:

Table 8.3 – Word-Set Summary - *"Good" /mni:ħ/*

Total	<7> Form	Tokens	%	%	Tokens	<h> Form
19	<*mni7>	13	68%	32%	6	<*mnih>

These summarised forms show a simplified view of the totality of our data and the variation within the ratio across it, allowing us to compare different word-sets more manageably, such as in Table 8.4 below comprising the 14 word-sets (tallied in the same way as we did for *"Good"* in Tables 8.2 and 8.3) in which at least 25 or more individual tokens appear per word-set:

Table 8.4 – All Word-Sets with **25≤ Total Tokens**

Total	<7> Form	Tokens	%	%	Tokens	<h> Form	
33	*we7ed	21	64%	36%	12	*wehed	*"One, a person"*
39	*ne7na	26	67%	33%	13	*nehna	*"We"*
36	*7elwe	24	67%	33%	12	*helwe	*"Nice, pretty"*
85	*7aram	57	67%	33%	28	*haram	*"Poor [thing]"*
29	*ra7	20	69%	31%	9	*rah	*"[He] went"*
39	*a7la	28	72%	28%	11	*ahla	*"Nicer than"*
100	*ye7mi	72	72%	28%	28	*yehmi	*"Protect"*
55	*7ob	40	73%	27%	15	*hob	*"Love"*
26	*de7ek	19	73%	27%	7	*dehek	*"Laughter"*
30	*7a2	22	73%	27%	8	*ha2	*"Justice"*
31	*rou7	23	74%	26%	8	*rouh	*"Go"*
70	*7ada	52	74%	26%	18	*hada	*"Someone"*
42	*7aki	32	76%	24%	10	*haki	*"Talk [noun]"*
30	*7et	25	83%	17%	5	*het	*"Put"*
645	**<7>**	**461**	**71%**	**29%**	**184**	**<h>**	

Not only is the ratio of the tallied total of all these word-sets **71:29**- exactly the ratio that appears when examining the entirety of the 1,071 individual tokens of our data- but we also see that the ratio for each individual word-group only ranges between 64:36 and 76:24, on

either end of the 71:29 ratio (with an outlier of 83:17 for <*7et>/<*het>). In this case, it is not a question of whether the high frequency appearance of a word influences individual users' choice of representation for the sound, as we hypothesised, but rather a matter of statistical likelihood that this expected ratio is reached, in relation to how much data is available. Doing the same for all the word-sets with less than 25 total tokens respectively produces a similar table (which can be found in the appendix under Table 8.5X), the full data of which we further summarise in Table 8.5 below by showing only the combined frequencies of <7> and <h> for the tallied word-sets of this table:

Table 8.5 – Summary of All Word-Sets with **25> Total Tokens**

[*Expanded table available in appendix, under Table 8.5X*]

Total	**<7> Form**	**Tokens**	**%**	**%**	**Tokens**	**<h> Form**
392[16]	**<7>**	**273**	**70%**	**30%**	**119**	**<h>**

There is much greater variation within these word-sets, where the word-set ratios vary between 29:71 to 91:9, further reinforcing the conclusion that much of the deviation present in the ratio derive simply from the fact that there are not enough tokens to reach the expected 71:29. When these high-variation word-sets are tallied together, however, as in Table 8.5 above, we see a ratio of 70:30 (only 1% off from our golden ratio). Though individually these categories do not have sufficient data to display the 71:29 ratio, the ratio appears emphatically when they are combined. The ratio of 71:29 is therefore a very good predictor for the variation of <7> and <h>, statistically becoming likelier to appear the more tokens any set consists of. Our hypothesis that the more commonly used a word is within the writing of LQA CMCR, the lower motivation there is to use a specific <7> to mark the sound (and therefore the more we expect <h> to appear in frequencies higher than its predicted 29%) is complicated by this statistical effect whereby the more data we analyse, the more likely a benchmark ratio is to appear. To disentangle these two effects, we must consider another approach.

[16] The 645 tokens of Table 8.4 and 392 tokens of Table 8.5 come to a total of 1,037, with the remaining 34 tokens showing the voiceless pharyngeal fricative (to form our grand total of 1,071) appearing in words that appear only once each and therefore do not feature in either of our tables.

II. Token-Convergence Examination

We see below the full token data for the word-set <*ye7mi>/<*yehmi> (meaning *"protect"* or *"protect him"*, most frequently used in the construction <Allah ye7mi> meaning *"May God protect him"*):

Table 8.6 - Word-Set Breakdown - *"Protect"* /jəħmi/

[*A translation for each individual form is available in the appendix under Table 8.6X*]

Total	<7> Form	Tokens	%	%	Tokens	<h> Form
48	**ye7mi**	**32**	**67%**	**33%**	**16**	**yehmi**
29	y7mi	27	93%	7%	2	yhmi
3	ye7me	2	67%	33%	1	yehme
2	y7mik	2	100%	0%	0	yhmik
2	yi7mi	0	0%	100%	2	yihmi
2	y7mikon	1	50%	50%	1	yhmikon
1	ye7meh	0	0%	100%	1	yehmeh
1	y7me	0	0%	100%	1	yhme
1	i7miyon	1	100%	0%	0	ihmiyon
1	y7miyun	0	0%	100%	1	yhmiyun
1	ye7mekkk	0	0%	100%	1	yehmekkk
1	y7miha	0	0%	100%	1	yhmiha
1	y7miki	1	100%	0%	0	yhmiki
1	ye7meyoun	1	100%	0%	0	yehmeyoun
1	wyi7miyon	1	100%	0%	0	wyihmiyon
1	wye7mikon	1	100%	0%	0	wyehmikon
1	by7mo	1	100%	0%	0	byhmo
1	yen7amou	1	100%	0%	0	yenhamou
1	7amina	1	100%	0%	0	hamina
1	y7mikkkkkkk	0	0%	100%	1	yhmikkkkkkk
100	***ye7mi**	**72**	***72%***	***28%***	**28**	***yehmi**

The tallied tokens of the full word-set show an expected ratio of **72:28**, represented under the generalised form <*ye7mi>/<*yehmi>. Examining the specific tokens, however, we find that <ye7mi> and <yehmi> in that particular form show a ratio of **67:33**, as we see in the first row. This leading couplet accounts for 48 of the total 100 tokens, in addition to demonstrating a 4% shift in favour of the less specific grapheme <h>: precisely the effect we predicted for high-frequency forms. The overall tallied ratio of 72:28 (1% more in favour of <7>) appears as a result of the great variation we find within the other tokens of the word-set (many of them uncommon or unconventional, thus requiring transcriptional

specification in their writing), but <ye7mi> and <yehmi> appear to be a convergent couplet, both in their popularity (accounting for nearly half the word-set) as well as the detectable move away from the expected ratio in favour of ambiguous grapheme <h>, despite its high number of tokens. We see this more clearly still when we compare the high-popularity convergent couplet <ye7mi>/<yehmi> to the tallied sum of all other less frequent, unconventional forms that make-up its word-set:

Table 8.7 – Convergent Form vs. Other Forms – *"Protect"* /jəħmi/

Total	**<7> Form**	**Tokens**	**%**	**<h>%**	**%**	**Tokens**	**<h> Form**
48	**ye7mi**	32	***67%***	**+4%**	***33%***	16	**yehmi**
52	**<7>**-variants	40	***77%***	**-6%**	***23%***	12	**<h>** -variants

We observe for the high-frequency form <ye7mi>/<yehmi> a rise (by 4%, from 29%) of the use of indistinct but conventional grapheme <h> over distinctive, unconventional grapheme <7>, and for the rest of the non-convergent forms the opposite: a rise (by 6%, from 71%) of the use of the distinctive and descriptive grapheme <7> (represented in Table 8.7 above as a 6% fall in <h>-frequency). Though both categories have near-identical token counts, comparing the difference between them (67:33 and 77:23) paints a still clearer image of the effect at play: specific forms <ye7mi> and <yehmi> see convergence on a single orthographic realisation and is thus potentially an emerging convention through high-frequency use, though it necessarily takes the form of a couplet due to the variant representation of the voiceless pharyngeal fricative. In this way, we can now further divide the rest of our word-sets from Table 8.4 between a high-frequency pairing (the convergent form), for which we expect higher frequency of <h>, and a pairing comprising the sum of all other forms, for which we expect a tendency towards higher use of <7>, now understanding that it is the sum of these two pairings that show statistical convergence on the 71:29 ratio that globally describes the graphemic choices for this sound in our data. The first row in each sub-table of Table 8.8 below shows the convergent pairing (not a generalised couplet like <*mni7>/<*mnih> of Table 8.2 but tokens that appear specifically with this exact orthographic form), and the second row shows the rest of the word-form's variants grouped together, along with the ratio and the degree to which each tends towards <h>.

Table 8.8 – Convergent vs. Non-Convergent Forms per Word-Set

Total	<7> Form	Tokens	%	<h>%	%	Tokens	<h> Form
48	**ye7mi**	32	***67%***	**+4%**	***33%***	16	**yehmi**
52	**<7>**-variants	40	***77%***	**-6%**	***23%***	12	**<h>** -variants
80	**7aram**	52	***65%***	**+6%**	***35%***	28	**haram**
5	**<7>**-variants	5	***100%***	**-29%**	***0%***	0	**<h>** -variants
26	**a7la**	17	***65%***	**+6%**	***35%***	9	**ahla**
8	**<7>**-variants	6	***75%***	**-4%**	***25%***	2	**<h>** -variants
50	**7ada**	37	***74%***	**-3%**	***26%***	13	**hada**
20	**<7>**-variants	15	***75%***	**-4%**	***25%***	5	**<h>** -variants
24	**ra7**	16	***67%***	**+4%**	***33%***	8	**rah**
5	**<7>**-variants	4	***80%***	**-9%**	***20%***	1	**<h>** -variants
19	**ne7na**	14	***74%***	**-3%**	***26%***	5	**nehna**
19	**<7>**-variants	11	***58%***	**+13%**	***42%***	8	**<h>** -variants
18	**7elwe**	11	***61%***	**+10%**	***39%***	7	**helwe**
18	**<7>**-variants	13	***72%***	**-1%**	***28%***	5	**<h>** -variants
17	**7a2**	12	***71%***	**0%**	***29%***	5	**ha2**
13	**<7>**-variants	10	***77%***	**-6%**	***23%***	3	**<h>** -variants
17	**wa7ad**	9	***53%***	**+18%**	***47%***	8	**wahad**
16	**<7>**-variants	12	***75%***	**-4%**	***25%***	4	**<h>** -variants
15	**mni7**	10	***67%***	**+4%**	***33%***	5	**mnih**
4	**<7>**-variants	3	***75%***	**-4%**	***25%***	1	**<h>** -variants
14	**a7san**	8	***57%***	**+14%**	***43%***	6	**ahsan**
2	**<7>**-variants	1	***50%***	**+21%**	***50%***	1	**<h>** -variants

13	**7aki**	8	***62%***	**+9%**	***38%***	5	**haki**
29	**<7>**-variants	24	***83%***	**-12%**	***17%***	5	**<h>** -variants

A very clear pattern emerges across the 12 word-sets above (which have the highest token counts within our data): each set shows a preferred convergent form, the vast majority of which demonstrate a clearer tendency towards <h> than predicted by the general ratio, whereas the remaining, variable, non-conventionalised forms tallied together very often show a tendency towards <7>. There are only three notable exceptions to this pattern: the convergent couplet <7a2>/<ha2> remains precisely at the expect ratio, while <ne7na>/<nehna> reverses the effect, showing higher <h> for the non-convergent forms; in both these cases, this is likely a result of the low popularity of the convergent forms compared to the other forms (which we discuss in more detail in 8.3). Finally <7ada>/<hada> sees higher <7> in both rows, and is clearly explained by the semantic overlap with the word for *"someone"* /ha:da/, which we discuss in 8.2.3. For now, we note that other factors work in tandem with the frequency effect, and must be considered alongside it, but otherwise also note that all nine other cases conform to our expectation that convergent forms will see lower specification, reaffirming not only our word-frequency hypothesis but also the very fact that there are words with a significantly observable conventionalisational effect, whose convergent orthographical forms have gained not only popularity but also other effects of conventionalisation, such as a lessened need for specificity in the higher tendency to write <h> and not <7>. This is especially notable in <wa7ad>/<wahad>, which has the greatest tendency to <h> of the forms above with a ratio of 53:47. That this convergent spelling reflects the Beiruti LQA pronunciation of /wa:ħad/ rather than Tripolitan equivalent /we:ħɪd/ is significant, indicating that the use of this form is highly conventionalised not only by frequency-convergence but also on the basis of the prestige Beiruti form, the highest-popularity orthographical form not reflecting the spoken Tripolitan LQA form (represented here only a total of 13 times) but the Beiruti LQA pronunciation (which appears with 20 tokens).

Table 8.9 – *Convergent Forms - "One, a person"* /we:ħɪd/ or /wa:ħad/

Total	<7> Form	Tokens	%	<h>%	%	Tokens	<h> Form
20	wa7ad	12	***60%***	+11%	***40%***	8	wahad
13	we7ed	9	***69%***	+2%	***31%***	4	wehed

Splitting the representation of this word in Table 8.9 above on the basis of which pronunciation is indicated, we see that while the Tripolitan LQA form <we7ed>/<wehed> does see an additional 2% <h>, it is the Beiruti LQA form which has a remarkably high ratio of <h> (overrepresented by 11% compared to the ratio). The combination in this form of both high-frequency graphemic conventionalisation (as per RQ1) as well as conventionalisation on the basis of an emergent prestige form (as predicted in RQ5) has the greatest visible effect, observable here within the choice of <h>. Both conventionalisational effects reduce the freely-transcriptional writing of this word, whether it is through lower graphemic variation (in frequency-convergence on particular graphemic choices) or else through a loss of transcriptional phonetic detail, with both effects leading to more conventional writing, and in cases such as this where both take place, highly conventional forms such as <wa7ad>/<wahad> appear, even if the variation in the writing of /ħ/ prevents a singular conventional form emerging. We further examine this phenomenon in 10.2.

In summary, we have observed that the higher frequency of usage of convergent forms encourages users to prefer <h> over <7> more frequently than they usually do, bearing in mind that <7> is almost always the preferred choice. In light of our analysis of prestige forms like <wa7ad>/<wahad>, we also understand this effect not only in terms of conventionalisation, but also transcriptionality: those writing words in a transcriptional and individual manner tend to rely on the specificity of <7> to mark their writing, whereas those using words with emerging, converging conventions are less likely to require the specific marking of phonemes such as /ħ/ and can instead resort to a less-defined but conventionalised <h>. We have noted that this effect is most highly visible when a word appears frequently, where there is no effect from other factors such as a loss of semantic clarity, and finally, where there are few points of variation other than this phoneme. We recall here the similar effect we saw for vowel omission in 7.3.1, whereby higher-frequency words were more likely to see higher rates of vowel omission, introducing an ambiguity

analogous to that of the use of <h> and which in turn sees greater use in the more familiar forms that appear with greater frequency. There too, as we have glimpsed in couplet <7ada>/<hada> (and will explore further in 8.2.3), we saw that semantic clarity can be a sufficient motivation to reverse this effect, as was the case with the non-omission of the vowel in /hal/ due to the ambiguity of <hl> in spite of its high frequency. We therefore now discuss the other hypothesised factors affecting the choice of <7> or <h>, including word-position (8.2.2), semantic clarity (8.2.3) and the effect of proper nouns (8.2.4), before returning to use our frequency-convergence analysis to develop our new model of understanding conventionalisation as a whole in 8.3.

8.2.2 Word Position

If the word-position of the /ħ/ sound has a tangible impact on which graphemic form is chosen by users of LQA CMCR, we should expect the ratio of 71:29 for <7> to <h> to differ meaningfully in different word-positions. Thus far, we have grouped words semantically rather than phonemically, thus forms like <7ayet> (*"life"*) and <bi7ayito> (*"in his life"*) comprise the <*7ayet>/<*hayet> (*"life"*) word-set, irrespective of the position of the sound in the word. Here, however, to understand the effect of word-position, we collect the individual raw tokens from our data anew and group them instead into three new categories: word-initial (like <7ayet>), word-medial (like <bi7ayito>) and word-final (like <rte7>, *"he rested"*). This has the additional advantage of giving us groupings with a high number of tokens, given that the entirety of our 1,071 words containing the voiceless pharyngeal fricative are thus present within these three groups.

Table 8.10 – Token Occurrence by Word-Position

	<7>			<h>	
	Tokens	**%**	**<h>%**	**%**	**Tokens**
Word Initial	275	**69%**	**+2%**	**31%**	122
Word Medial	396	**72%**	**-1%**	**28%**	154
Word Final	95	**77%**	**-6%**	**23%**	29

For both the word initial and word medial positions, the ratio across all words differs from the golden ratio of 71:29 by only 2% and 1% respectively. In the word-final position,

however, the use of <7> appears notably higher than expected, with <h> appearing a full 6% less than it does elsewhere, the ratio showing as 77:23 for <7> to <h>. In spite of the word-final position being less represented at a total of 124 tokens (versus 397 for word-initial and 550 for word-medial), it nevertheless has a richer representation than any single semantic word-group such as those we analysed in Table 8.5 (in which <*ye7mi>/<*yehmi>, the most represented, consisted of 100 tokens). This overrepresentation of <7> in the word-final position is potentially explained by the rules of StE orthography. Given that SA /a/ drops to /e/ in LQA in the word-final position for feminine nouns (SA f. noun /safi:na/ for *"ship"* becoming /safi:ne/ in LQA), using a word-final <e> to represent this common sound can be misunderstood (using StE orthographical conventions) to be modifying the preceding vowel rather than being sounded (turning to /aɪ/, as in StE *mine, twine, brine*), hence the preference in some cases for the form <eh> to indicate the sounding of the word-final /e/. This is potentially also influenced by the use in SA writing of what is called a *taa marbuta* <ۃ> at the end of these same feminine nouns, marking a silent <h> (in the orthographically related form <ہ>) unless the word is followed by certain grammatical forms in which case it is sounded as /t/. The use of <h> in the word-final position following word-final vowels (and particularly in the case of feminine noun endings) is therefore intuitive in both SA and StE writing, leading to forms such <alileh> (*"little, few* [f.]*"*, from Dataset 1) being used to ensure the marking of intended form /ali:le/ rather than a misreading of something like /alaɪl/, a risk accepted by the writer of the one token of <alile> we find in our data. Though we note that this <eh> ending is not widely used in all cases, we nevertheless find for example that 17 of the total 35 tokens in our word-set <*7elwe>/<*helwe> are produced with a silent grapheme <h> in the word-final position (showing as <7elweh>, <helweh>, <7lweh> and <hlweh>). This prevalence of the use of word-final, unpronounced <h> explains why <7> is used more frequently than <h> for the voiceless pharyngeal fricative in this position in order to avoid misreading, and thus likely why we see a 6% increase of <7> in this position compared to word-initial and word-medial, where the general ratio is largely retained. Nevertheless, it is significant to see that this effect is cancelled out by the high-frequency conventionalisation effect described in 8.2.1 above, which we see below in a table consisting of the six word-final /ħ/ token-forms (specific spellings, not generalised word-sets) with the highest frequency:

Table 8.11 – Word-Final /ħ/ - Specific Token-Forms

Total	<7> Form	Tokens	%	<h>%	%	Tokens	<h> Form
24	ra7	16	67%	+4%	33%	8	rah
15	mni7	10	67%	+4%	33%	5	mnih
10	sa7i7	9	90%	-19%	10%	1	sahih
6	sa7	6	100%	-29%	0%	0	sah
6	rou7	5	83%	-12%	17%	1	rouh
5	la7	4	80%	-9%	20%	1	lah
66	**<7>**	50	***76%***	**-5%**	***24%***	16	**<h>**

Four of the six forms show the heightened use of <7> we have associated with the word-final position with the notable exception of the two most common forms, <ra7>/<rah> and <mni7>/<mnih>, both of which appear with ratios of 67:33, thus 4% more in favour of <h>. Both of these forms show a convergence effect, which overcomes the word-final effect that favours <7>, while the rest of the forms in Table 8.11 show a lower number of tokens and thus retain the expected word-final effect more in favour of <7>. We must therefore consider the overall sum of factors affecting the choice of /ħ/-representation: were it not for the word-final pharyngeal for the two convergent words, the frequency effect might be expected to be more pronounced than 4%; conversely, were it not for their high frequency of appearance, these tokens would have appeared with the higher tendency to <7> exhibited by the rest of the lower-frequency word-final forms. In summary, the divergence from the expected ratio is a sum of the relevant factors for any given token.

8.2.3 Semantic Overlap with /h/

Based on prior work on grassroots conventionalisation (particularly Hinrichs, 2004), we have hypothesised that the need for semantic clarity will influence the orthographic choices made by individuals. In the case of the voiceless pharyngeal fricative, this is best seen in cases where the grapheme <h> has semantically valid readings either as /h/ or /ħ/, which leads to higher use of unambiguous <7>. To analyse this effect, we add an additional column to the tables we have used thus far, in which we indicate the alternate /h/ meaning and the number of tokens which (as clarified by the context they appear in) are to be read with an /h/[17].

[17] For clarity, the words in which the use of <h> has been identified (by context) to be representing glottal /h/ and not pharyngeal /ħ/ have not been included in any of the totals we have used elsewhere; they only feature

Table 8.12 - Semantic Comparison: /aħla/ vs. /ahla/

/ħ/ *Tokens*	<7> Form	Tokens	%	<h>%	%	Tokens	<h> Form	Tokens
					"Nicer" /aħla/			*"Her family"* /ahla/
26	a7la	17	**65%**	**+6%**	**35%**	9	ahla	15
4	2a7la	3	75%	**-4%**	25%	1	2ahla	1
1	2a7le	1	100%	-29%	0%	0	2ahle	0
2	27la	2	100%	-29%	0%	0	2hla	0
3	a7le	3	100%	-29%	0%	0	ahle	0
1	a7lehon	1	100%	-29%	0%	0	ahlehon	0
1	7la	1	100%	-29%	0%	0	hla	0
1	a7laha	0	0%	+71%	100%	1	ahlaha	0
1	a7lahe	0	0%	+71%	100%	1	ahlahe	0
40	***a7la**	**28**	***70%***	**+1%**	***30%***	**12**	***ahla**	**16**

In the /ħ/ reading, /aħla/ means "*nicer, nicer than*", while the /h/ reading as /ahla/ indicates the feminine genitive form "*her family*". Out of a total 24 appearances of the tokens <ahla> and <2ahla> (the latter overtly marking the word-initial glottal stop with <2>), taking into account semantic context, a total of 16 tokens indicate the /h/ form "*her family*", and only 10 indicate the /ħ/ form *"nicer than"*. The total for all 40 tallied /ħ/ tokens comes to an expected 70:30, diverging only by 1% in favour of <h>, though we also note that couplet <a7la>/<ahla> (with 26 tokens) is convergent via the high-frequency effect, and thus shows a ratio of **65:35**, 6% more in favour of <h> than expected. Conversely, as we saw in Table 8.8, the non-convergent pharyngeal forms see a 4% rise in <7>, indicating transcriptional writing, but for convergent <a7la>/<ahla>, the favouring of <h> overcomes the semantic risk of glottal misreading /ahla/, even though /ahla/ is not a fringe form but is itself relatively popular with 15 tokens of <ahla> intended to represent it. It is certainly possible that this convergent form might have produced a still-higher ratio in favour of <h> were it not for the potential for semantic confusion with the glottal form, though we conclude that overall, in this case at least, the frequency-convergence effect that favours <h> overrides the semantic

in this section in the additional column as indications of the non-pharyngeal readings in potentially ambiguous contexts, and here are also naturally are not counted in the totals and ratio calculations for the pharyngeal readings.

confusion effect that would favour <7>. There are two additional forms with significant variable readings as /ħ/ or /h/.

Table 8.13 - Semantic Comparison: /ħada/ vs. /ha:da/

					"Someone" /ħada/			*"This one"* /ha:da/
/ħ/ *Tokens*	**<7> Form**	**Tokens**	**%**	**<h>%**	**%**	**Tokens**	**<h> Form**	**Tokens**
50	7ada	37	74%	**-3%**	26%	13	hada	5
5	7adan	1	20%	+51%	80%	4	hadan	0
8	ma7ada	7	88%	-16%	13%	1	mahada	0
7	ma7adah	7	100%	-29%	0%	0	mahadah	0
70	***7ada**	**52**	***74%***	**-3%**	***26%***	**18**	***hada**	**5**

The word-set <*7ada>/<*hada> has a high-frequency convergent couplet in <7ada> and <hada>, with an emphatic total of 50 tokens (compared to only 20 remaining tokens that are not part of the convergent couplet), and yet does not show the expected higher tendency to <h>, in fact showing 3% higher <7> instead. This is most likely precisely because of the effect of semantic confusion, which in this case cancels out the convergence effect given that <h>-form <hada> is more likely to be avoided for fear of confusion with the word /ha:da/ (meaning "*this*"), which 5 of the total 17 <hada> tokens do indeed represent. In contrast, only 13 tokens of the same form <hada> indicate the pharyngeal pronunciation /ħada/ (meaning "*someone*"), versus 37 tokens as <7ada>, and therefore the convergent couplet <7ada>/<hada> has a ratio of **74:26**, with a preference for <7> (as predicted by the semantic clarity effect) rather than for <h> (as the high-frequency convergent form predicts). These factors can therefore have different effects, and with our present understanding, we cannot predict which is likely to prevail, as summarised in the table below showing the opposite effects for the two words discussed in this section, where frequency-convergence increases the use of <h> for writing /aħla/ by 6%, but semantic clarity increases the use of <7> by 3% for writing /ħada/.

Table 8.14 – Comparison of Convergent Tokens – "*Someone*" and "*Nicer than*"

					/ħ/		/h/
/ħ/ *Tokens*	<7> Form	Tokens	%	<h>%	%	Tokens	<h> Form	Tokens
50	7ada	37	74%	-3%	26%	13	hada	5
26	a7la	17	65%	+6%	35%	9	ahla	15

8.2.4 Place Names & Proper Nouns

The final possibility we hypothesised to affect choice of <7> versus <h> is the frequent appearance of certain orthographic forms as names or common proper nouns, whether in use in Lebanon itself or else in standard orthographies such as that of StE. In both cases, it is more likely that <h> will be used (except in specifically performative purposes, where applicable), such as in names on Facebook profiles, or place names as they appear on street signs, maps, addresses and so on. On this basis, we expect such forms to show a higher use of <h> within LQA CMCR too, given their familiarity in that form. This is unlike the high-frequency effect we discussed in 8.2.1 because these are not necessarily likely to appear with high frequency either in the data we are examining or necessarily even across the entirety of LQA CMCR writing, and may not even be commonly produced by users of LQA CMCR at all, but nevertheless appear and are read in the <h>-form frequently outside the immediate context of LQA CMCR. In Table 8.15 below we see two clear examples of this phenomenon in the case of two very common personal names:

Table 8.15A – *Mohammad* /mħam.mad/

Total	<7> Form	Tokens	%	<h>%	%	Tokens	<h> Form
4	mo7amad	0	0%	+71%	100%	4	mohamad
4	m7amad	3	75%	-4%	25%	1	mhamad
2	mou7amad	0	0%	+71%	100%	2	mouhamad
1	mo7amed	0	0%	+71%	100%	1	mohamed
1	mo7ammed	0	0%	+71%	100%	1	mohammed
1	ma7amad	0	0%	+71%	100%	1	mahamad
1	mou7amed	0	0%	+71%	100%	1	mouhamed
1	m7ammad	1	100%	-29%	0%	0	mhammad
1	mou7ammad	1	100%	-29%	0%	0	mouhammad
16	***mo7amad**	**5**	**31%**	**+40%**	**69%**	**11**	***mohamad**

Table 8.15B – *Ahmad* /aħmad/

Total	<7> Form	Tokens	%	<h>%	%	Tokens	<h> Form
6	a7mad	0	*0%*	+71%	*100%*	6	ahmad

The effect is very clear in both cases. For the name *Ahmad*, there is not a single instance across all of Dataset 1 in which it appears with <7> as something like <a7mad>, but it does appear six separate times as in the precise orthographical form <ahmad>. Though the total token count in this case is low, the lack of a single rendering of it with the grapheme <7> is indicative of a major difference from the usual 71:29 rate. We see something similar for *Mohammad*, which appears more frequently with a total of 16 tokens and sees some <7>-forms too, though again the fact that these are very clearly not the majority is indicative of the same effect. The ratio for *Mohammad* is almost completely reversed, appearing at 31:69 in favour of <h> for the entire word-set, and thus with 40% more frequency for <h> than the ratio predicts. There is therefore a clear effect on the basis of the frequent appearance of these names as written with an <h> outside of LQA CMCR. In the case of *Mohammad,* this effect is in fact not accompanied with the emergence of a convergent orthographical form (as we find in <ahmad> for *Ahmad*), but the familiarity of the word in its <h>-form is enough to reduce the need to specify the phoneme as pharyngeal (and not glottal), even while the individual forms (as in Table 8.15A) remain largely transcriptional in nature, therefore further distinguishing the effect of this factor from the one of the high-frequency factor, for which rates of <h> rise only in tandem with convergence towards a single orthographic form. That we see in the name *Ahmad* six identical tokens of <ahmad> does not necessarily indicate a different effect, but rather we hypothesise that this occurs because the composition of the word consists of phonemes that happen to have straight-forward and largely invariable orthographical resolutions within the repertoire of LQA CMCR, with the sole exception of the voiceless pharyngeal fricative- thus the resolution of that singular variant results in the resolution of the entirety of the word and its convergence on a single form by that means, a process we discuss further in 8.3 to follow. It is also clear (both from this section and our analysis overall) that the clean split between exclusive use of <h> for names and proper nouns and <7> for everything else proposed by Abu Elhij'a (2014) does not hold up in our data.

Table 8.16 – "*My love*" /ħabi:bi/

Total	<7> Form	Tokens	%	<h>%	%	Tokens	<h> Form	
6	7abibi	2	33%	+38%	67%	4	habibi	"*My love* [m.]"
1	7abib	1	100%	-29%	0%	0	habib	"*My love* [m.]"
1	7abibtna	0	0%	+71%	100%	1	habibtna	"*Our love* [f.]"
1	7bibti	0	0%	+71%	100%	1	hbibti	"*My love* [f.]"
1	7abibti	0	0%	+71%	100%	1	habibiti	"*My love* [f.]"
10	***7abibi**	**3**	**30%**	**+41%**	**70%**	**7**	***habibi**	

While we find that no single place name comprising the voiceless pharyngeal fricative features prominently enough in our data for meaningful analysis, we examine instead the word <7abibi>/<habibi>, often rendered in English as <habibi> even if it is not formally part of StE (for example, very commonly used as the name of Lebanese restaurants, shisha bars, and the like). We see in Table 8.16 that <*7abibi>/<*habibi> does indeed show the expected effect, with a full reversal of the ratio in favour of <h>, going from **71:29** to **30:70**, and as with *Mohammad,* this effect occurs not only for the aggregated totals, but actually across nearly all variants of the word, including feminine form <7abibit>, (/ħabi:bti/, "*(my) love* [f.]" and first person plural <7abibtna> (/ħabi:bətna/, "*(our) love* [f.]"). In this case, even less common, more transcriptional forms show a decreased pressure for marking the voiceless pharyngeal fricative as a result of external familiarity with the word-form, which we do not see in the case of frequency-convergence. That the couplet <7abibi>/<habibi> has 6 of the 10 total tokens does allow us to predict that there is likely also convergence in effect, which would become more apparent with enough tokens. This word is likely to be much more common within the use of LQA CMCR in other CMC sub-genres, including fully synchronous one-on-one conversation, and features prominently in Dataset 2, where we will revisit this analysis in 9.2.2 IV.

8.3 Formulating the Lexemic-Aggregational Model

8.3.1 Conclusions: Convergence and Conventionalisation

We have defined the most important variables affecting the choice between <7> and <h>, and understood that these factors often work in tandem (or opposition), and so must be considered together in order to understand the divergence from the overall ratio of 71:29, whether for a word or word-set. We have also specifically addressed the predicted

frequency (RQ1) and semantic clarity (RQ2) effects that we hypothesised on the basis of work in this field so far, and discussed further some of the impact of Roman script standard orthographies (RQ3) on these variables (outside of the immediate sub-orthographical context of the previous chapter), as well as further effects of the SA orthography (RQ4) in the same way. Starting with our overall ratio of 71:29 for <7> to <h>, we have observed how the need to overtly delineate this phoneme sees users opt for <7> more frequently than predicted by the ratio, whether in the case of the risk of semantic overlap with /h/ readings of the sound in the same position (8.2.3), the case of risking that the <h> is not read at all in the word-final position (8.2.2), and more generally in the writing of less frequently used and observed forms, where the pressure to distinguish this sound is greater because of the lack of conventional readings of the word such as those we observed to arise for commonly-used words in the case of the convergent forms identified in 8.2.1. Conversely, we expect the frequency at which the grapheme <h> is utilised by users of LQA CMCR to rise in cases where the risk of misreading is lower, such as the absence of the risks described above, and more generally, where there is a perceived expectation that the form being produced is readily recognisable, something that can also occur as a result of the abundance of a particular spelling outside of LQA CMCR such as in personal names, place-names or in the use of other standard orthographies (8.2.4), but most tellingly, in cases where the high-frequency appearance of specific orthographical representations within our data (and LQA CMCR more generally) results in the emergence of convergent forms, where the use of <h> therefore becomes more frequent as a result of the reduced pressure to overtly mark the sound in question: in this way, we can observe the workings of grassroots conventionalisation. We recall our discussion in Chapter 4 of the nature of transcriptional writing in non-standard orthographies (discussed at length in 4.3.2), such as Sebba's (2007) discussion of non-standard Alsatian writing which results in readers being required to 'sound out' the text being read, and which Jaffe (2000) calls *a decoding mode* that she contrasts with the 'scanning' mode of standard writing, and which in turn we determined is made possible not exclusively by the use of standard orthographies, but also through relatively more defined conventions even within non-standard orthographies, whose orthographical forms can exist in variable positions along the scale between fully transcriptional and fully standardised, depending on the degree to which any given form is conventionalised. We thus interpret the choice between <7> and <h> in the context of this

duality of transcriptional versus conventionalised writing: the delineation and clarity afforded by <7> is generally favoured in writing that more closely resembles the transcriptional, whereas the orthographically more conventional <h> is preferred in cases where the production of a form is to some degree conventionalised. We now broaden our interpretation of the frequency-convergence effect beyond only this single sound and explore the wider orthographical workings of this process.

8.3.2 Exploring the Phonemic-Graphemic Space

The emergence of convergent forms that occur as a result of high-frequency usage must be understood in terms of the existence of a highest-popularity graphemic choice in each position within a word, meaning that when any given word is frequently utilised, the convergent form that emerges will be the form consisting of the highest-popularity grapheme in every position within the word. This in turn means that words with lower variability in their individual phonemic-graphemic positions will converge more easily on such conventionalised forms, an effect we have already seen in 8.2.4, where the 6 tokens of the name *Ahmad* appeared identically as <ahmad> within the dataset, whereas the 16 tokens of the name *Mohammed* showed much greater variation and no convergence, despite both these words being subject to the familiarity effect and both showing higher instances of <h> as a result. This difference is a direct result of the relatively simple graphemic choices for *Ahmad* and the relatively variable graphemic choices for *Mohammad*. To illustrate this better, we extrapolate the phonemic space of the word *Mohammad* on the basis of the orthographic representations of it present in our data in Table 8.17 below. We see high graphemic variation in a number of positions: whether the initial /u/ vowel is omitted or written, and if so, whether it is as <o>, <ou> or even <a>; whether the geminated consonant /m.m/ is written with a reduplicated <mm> or a single <m>; and whether the second /a/ vowel is written as an <a> or <e>. The only constant positions are 1 (<m>), 4 (<a>) and 7 (<d>).

Table 8.17 – Phonemic-Graphemic Distribution – *"Mohammad"* /mħam.mad/

Pos.	1	2	3	4	5	6	7	**Tokens**
IPA	**/m/**	**/u/**	**/ħ/**	**/a/**	**/m.m/**	**/a/**	**/d/**	
1	m	-	**7**	a	mm	a	d	1
2	m	-	**7**	a	m	a	d	3
3	m	ou	**7**	a	mm	a	d	1
4	m	o	**h**	a	m	a	d	**4**
5	m	ou	**h**	a	m	a	d	2
6	m	o	**h**	a	m	e	d	1
7	m	o	**h**	a	mm	e	d	1
8	m	a	**h**	a	m	a	d	**1**
9	m	ou	**h**	a	m	e	d	1
10	m	-	**h**	a	m	a	d	1

This same information can be still more usefully represented by showing the graphemic frequency for each phoneme (rather than showing every representation individually), which we see in Table 8.18 below, allowing us to observe both the range of graphemic variants as well as the frequency of their appearance within the given word. We see that in the case of *Mohammad,* even should the choice between <h> and <7> resolve in favour of one single form or the other, there remains nevertheless a great number of variations in other positions in the word that makes the emergence of a convergent form less likely, though we also see that for each position, there is a variant with majority popularity, which in turn would predict, given enough tokens, the emergence of a convergent word <mohamad> (taking the most popular variant for each position), or instead a variable convergent form of <mohamad>/mo7amad>. As we see in Table 8.17 above, <mohamad> is in fact precisely the form with the highest number of tokens (4), even if it not by any great majority.

Table 8.18 – Phonemic-Graphemic Breakdown – *"Mohammad"* /mħam.mad/

	IPA	Variant	Variant%
1	**/m/**	**m**	**100%**
2	**/u/**	**o**	**38%**
		ø	**31%**
		ou	25%
		a	6%
		u	0%
3	**/ħ/**	**h**	**69%**
		7	31%
4	**/a/**	**a**	**100%**
5	**/m.m/**	**m**	**81%**
		mm	19%
6	**/a/**	a	**81%**
		e	19%
7	**/d/**	d	**100%**

The variable point with the smallest majority is the choice of <o> to represent initial vowel /u/ at 38%, with an omitted vowel close behind at 31%, which in turn results in the second most frequent form <m7ammad> (with 3 tokens in Table 8.17). The likely reason this second-most popular form returns to <7> rather than <h> (which in the case of this word is the less-preferred variant) is due to the undesirable ambiguity introduced by the use of <h> in dense consonant clusters (which we return to in our discussion of /jəħmi/ below). In higher-frequency words, therefore, variant points without a clear resolution in a single grapheme (like the 38% <o> vs. 31% omission in this case) can in fact lead to the emergence of more than a single convergent form or pairing (and in turn, lessening the degree to which the most-convergent form or pairing is dominant). We have already observed this in passing in Table 8.6 (section 8.2.1) in the case of the word-set <*ye7mi>/<*yehmi> ("*may [he] protect*"). In order to focus in on the specific form /jəħmi/, we now remove the additional grammatical variants (such as "*protect them", "protect you [pl.]*"- see Table 8.6X in the

appendix for the full forms with translations), retaining only variants of the specific grammatical and phonemic form /jəħmi/ in the third person singular.

Table 8.19 – Word-Form Breakdown – "*Protect*" /jəħmi/

Total	<7> Form	Tokens		<h> Form
48	**ye7mi**	**32**	**16**	**yehmi**
29	**y7mi**	**27**	**2**	**yhmi**
3	ye7me	2	1	yehme
2	yi7mi	0	2	yihmi
1	ye7meh	0	1	yehmeh
1	y7me	0	1	yhme

Here we observe two separate emergent forms: <ye7mi>/<yehmi> is the most popular at 48 total tokens (varying only in the choice between <7> and <h>), followed by a second, relatively less popular <y7mi>/<yhmi> form that nevertheless still stands out at 29 tokens; beyond these two forms (which together comprise 77 tokens), only 7 further alternative tokens exist in our data for this specific grammatical form. Again, here it is the omission of the first /e/ vowel that is the important point of difference, in addition to the variation in the voiceless pharyngeal which splits the 48 tokens of the most-popular form into 32 for <ye7mi> and 16 for <yehmi>, and the 29 tokens of the second-most popular form into 27 for <y7mi> and 2 for <yhmi>. This is particularly significant given that the second-convergent form as written with a <7> (<y7mi>, with 27 tokens) is in fact nearly twice as popular as what we have labelled the most-convergent form in its <h> manifestation as <yehmi> (which has 16 tokens).

Table 8.20 – Competing Convergent Forms – "*Protect*" /jəħmi/

	<e> **48**	**<ø>** **29**
<7> **59**	<ye7mi> **32**	<y7mi> **27**
<h> **18**	<yehmi> **16**	<yhmi> **2**

The data in Table 8.20 challenges our labels of most-convergent and second-convergent which we have applied on the basis of using variation between <7> and <h> as the primary fulcrum for variation, where here perhaps it is more accurate to say that <ye7mi>/<y7mi> is the convergent form (varying in representation of the first vowel, and not of the voiceless pharyngeal fricative, showing 59 tokens when tallied), followed by <yehmi>/<yhmi> (the <h> form of the same omitted/non-omitted set, which shows 29 tokens when tallied). It may well be the case that for this word (and so for other words with multiple convergent forms), such an approach is preferable, though we must also bear in mind that there is some degree of anomaly in the fact that the omitted set <y7mi>/<yhmi> shows a staggering 93:7 ratio in favour of <7>, which is anomalous not only statistically (given that it is a relatively high-token word) but also in conventionalisational terms, given that vowel-omitted <y7mi>/<yhmi> is a convergent form itself, which we then expect to tend towards <h>. This is explained simply by the semantic ambiguity of the clustering of <yhmi>, which is particularly difficult to decipher and overlaps a word like /jhəm.mi/ (as in the construction <ma byhmi> /ma: bəjhəm.mi/ "*I don't care"* literally, *"it does not concern me"*). As such, the factor of semantic confusion intrudes distinctly in the case of the couplet <y7mi>/<yhmi> and leads to a highly anomalous ratio of 93:7 for <7>, whereas the first convergent form <ye7mi>/<yehmi> that shows the ratio of 67:33 that tends slightly (by 4%) towards <h> as we expect. In this case, semantic ambiguity arises not necessarily as a result of an alternate /h/-reading (though one is possible in our example of /jhəm.mi/), but rather as a result of the difficulty of parsing the consonant cluster <yhm> in relation to where the omitted vowels are to be replaced (for which <y7m> is much clearer). Nevertheless, this example draws attention firstly to the assumption that the axis of <7> and <h> is always an appropriate approach for paired words involving the sound /ħ/, which is not always the case, as well as the assumption that a single convergent form can always unproblematically be identified, which again is not always true.

Table 8.21 – Phonemic-Graphemic Breakdown – "*Protect*" /jəħmi/

IPA	Variant	Variant%
/j/	**y**	**100%**
/ə/ or **/e/**	**e**	**62%**
	ø	**36%**
	i	2%
/ħ/	**7**	**73%**
	h	27%
/m/	**m**	**100%**
/i/	**i**	**94%**
	e	5%
	eh	1%

We see in the variable breakdown for specific orthographical form <ye7mi>/<yehmi> the points of variation clearly represented: in addition to invariable graphemes <y> and <m>, the most popular variant is <i> for the word-final vowel /i/ with 94% frequency (being the grapheme used in both convergent forms). On the basis of Table 8.18 (the variable breakdown for *Mohammad*), the less clear choice between <e> (at 62%) and omission (at 36%) for initial vowel /ə/ (or /e/) demonstrates why there is an additional split between <e> and omission, as it is less clear-cut than the split between 73% <7> and 27% <h> for the representation of the voiceless pharyngeal (which would have otherwise provided the primary point of variation). It is because <yhmi> is avoided (due to semantic motivation) that variation in this word is split in the unconventional manner that we have seen.

Table 8.22 - Phonemic-Graphemic Breakdown – "*Ahmad*" /aħmad/

IPA	Variant	Variant%
/a/	**a**	**100%**
/ħ/	**h**	**100%**
	7	0%
/m/	**m**	**100%**
/a/	**a**	**100%**
	ø	0%
/d/	**d**	**100%**

On the other extreme, we see that in the case of the name *Ahmad*, the only real viable variant grapheme (other than the voiceless pharyngeal fricative) is potentially word-omission for the second /a/ (which does not occur for this word in this dataset). Here there are only two feasibly variable positions and these alternatives are not used in any of the six tokens. In this case, by resolving the variation between <7> and <h> (which resolution here occurs due to the use of the name outside of LQA CMCR, as discussed in 8.2.4), a uniform convergent form <ahmad> becomes the form of choice among users, even with a small number of tokens. When such readily-resolved words occur with much greater frequency, this effect is only further emphasised, as we see in Table 8.23:

Table 8.23 – "*Poor thing, woe!*" /ħara:m/

Total	**<7> Form**	**Tokens**		**<h> Form**
80	**7aram**	**52**	**28**	**haram**
1	7arem	1	0	harem
1	7arram	1	0	haram
1	7ram	1	0	hram
1	ya7aram	1	0	yaharam
1	7aramekk	1	0	haramekk
85	***7aram**	**57**	**28**	***haram**

Quite remarkably, with a total of 85 tokens, there are only three instances of genuinely alternative constructions of the word <7aram>/<haram>, considering that <ya7aram> is a combined form of the construction <ya 7aram> (/ja: ħara:m/ being an emphatic form of the same word), consisting of the same basic construction <*7aram> at its core, as does <7aramekk> (which takes the second person feminine, meaning "*woe is you! [f.]*"). The only true variation outside of the voiceless pharyngeal fricative is one instance of omission for the first /a/ vowel (highly unconventional in this position, particularly considering that <7ram> is easily misread as /ħra:m/ meaning *"blanket"*), one instance of reduplication in <rr>, which even in cases of geminate consonants is infrequent (as we will see in 9.2.4) while the /r/ here is not geminated, and finally an <e> for the second /a/ vowel (also anomalous, as we will see in 9.3.2 IV., unless otherwise indicative of a transcriptional production of a particular accent). That the variant, transcriptional forms all utilise specific <7> and not <h> serves also to reinforce our conclusion that <7> is preferred in transcriptional writing that

reproduces the intended sounds individually, whereas <h> is more preferred when calling upon a more conventional, frequent and familiar form.

Table 8.24 – Phonemic-Graphemic Breakdown – "*Poor thing, woe!*" /ħara:m/

IPA	Variant	Variant%
/ħ/	**7**	**67%**
	h	**33%**
/a/	**a**	**99%**
	ø	1%
/r/	**r**	**99%**
	rr	1%
/a:/	**a**	**99%**
	e	1%
/m/	**m**	**100%**

In the phonemic-graphemic breakdown for this word, we see again a clear lack of any meaningful graphemic choice for users of LQA CMCR writing this word; instead, here it is the lack of resolution between <7> and <h> (the only truly variable position) that, despite an almost-perfectly convergent couplet of <7aram>/<haram>, prevents the emergence of a near-unanimously preferred word by conventionalisation through grapheme-preference frequency. Though we certainly see the ratio shift (by 6%) in favour of <h> due to the frequency of the word and its highly convergent spelling, showing 65:35 instead of 71:29, this is still not enough to lead to a form we could call fully conventional even for this high-frequency, high-token word. In fact, that the generalised ratio itself begins in favour of <7> at 71:29, and shifts towards <h> in cases of convergence and lack of ambiguity means that, somewhat ironically, the emergence of conventional convergent forms leads to *more* variation rather than less, increasing the frequency of infrequent grapheme <h> and moving the ratio in the direction of 50:50: that is, towards greater variability for this sound, whether for an individual word or word-set. Cases like the name *Ahmad,* which appears at a 0:100 ratio in favour of <h>, are quite rare, and *Ahmad* shows this kind of ratio primarily because of its use as a personal name, as discussed, but also due to the low number of tokens, where had it had more than 6 tokens we would have expected at least some instances of specific

delineation with <7>). Finally, we also recognise that not all forms can be expected to reach convergence, no matter their total number of tokens. This is the case when there are too many variable points without clear majority but instead closely-competing graphemic alternatives, resulting in the emergence of a handful of forms which may be more popular than others, but do not produce one or two forms with overall popularity. Though we do not have good examples of this in our data in the context of the voiceless pharyngeal fricative, the word "*If God wills [it]*", pronounced /ənʃa:l.la/ shows such an effect. We split this word-set by which representation of /ʃ/ is used (as we did for this split in Chapter 7 prior, and in the same way we have done in this chapter for the representation of /ħ/), for the sake of familiarity and simplicity.

Table 8.25 – Word-Set Breakdown <ch>/<sh> – "*If God wills*" /(ə)nʃa:l.la/

Total	**<ch> Form**	**Tokens**		**<sh> Form**
19	nchalla	11	8	nshalla
13	nchallah	6	7	nshallah
16	nchala	7	9	nshala
5	nchalah	2	3	nshalah
1	nchallh	1	0	nshallh
3	inchala	2	1	inshala
5	inchallah	4	1	inshallah
3	inchalla	0	3	inshalla
1	enchalah	0	1	enshalah
3	enchallah	1	2	enshallah
69	**<ch>**	**34**	**35**	**<sh>**

We see clearly that no form (nor even a paired-form) takes real precedence over the rest, with each specific variation ranging between 1 and 11 tokens for a total of 69. The reasons for this become apparent in our phonemic-graphemic breakdown for this word:

Table 8.26 – Phonemic-Graphemic Breakdown – "*If God wills*" /(ə)nʃa:l.la/

Position	**IPA**	Variant	Variant%
1	**/ə/**	**ø**	78%
		i	16%
		e	6%
2	**/n/**	**n**	**100%**
3	**/ʃ/**	**ch**	49%
		sh	51%
4	**/a:/**	**a**	**100%**
5	**/l.l/**	**ll**	64%
		l	36%
6	**/a/**	**a**	**99%**
		ø	1%
7	**/h/**	**a**	59%
		ah	41%

Though this word sees much greater variation in its output by users of LQA CMCR than *Mohammad* in Table 8.18, it is not composed of more variable positions. We defined 4 variable positions in *Mohammad* (positions 2, 3, 5 and 6 in Table 8.18), and here again we see 4 truly variable positions: 1, 3, 5 and 7 (the omission of position 6 is minimal, appearing in a single token). Both words, too, are composed of seven positions, and both are commonly-used and commonly-seen words. The real difference, however, is the rate of variation in each of the four positions: for *Mohammad,* the phonemes in positions 3, 5 and 6 show clear preference of a single grapheme (at 69%, 81% and 81% respectively), and the only choice of grapheme that varied significantly was for /u/ in position 2 (38%, 31%, 25% and 6%). For *"If God wills"*, however, all of positions 3, 5 and 7 show the same kind of variation that appeared in a single position for *Mohammad,* with almost evenly-matched variation between two graphemic alternatives for each position, that are more or less equally viable- meaning no convergent form can emerge. The variation in position 1

between three resolutions for the initial vowel does favour omission, but because of the three highly variable positions, rather than an emergence of a word set differing on the use of omission versus <i> for example, this position merely adds to the multiplicity of forms that emerge. We now move on to using Dataset 2 in the next chapter in order to review our overall conclusions so far in the light of the new written data of the new dataset, as well as simultaneously further developing our Lexemic-Aggregational model by constructing a more widely definitive set of frequencies for each of the significant graphemic choices made by users of LQA CMCR. This will allow us to address in the final Chapter 11, alongside our other pertinent questions, whether and to what extent this model allows us to observe grassroots conventionalisation within LQA CMCR (within which we also examine the word *"God-willing"* again in light of our new data in 11.1.6), alongside the prestige-based reduction of transcriptionality that we address in Chapter 10.

Chapter 9: Dataset 2 – Experimental Data

9.1 Dataset 2

9.1.1 Methodology & Analytical Approach

Dataset 2 consists of a series of interviews conducted and recorded in Tripoli in October of 2016, with a selection of 49 locals of varying ages and genders and of different educational, linguistic, and socio-economic backgrounds. The interviews consisted of two stages: in Part 1, participants were presented with six sentences of LQA written in what we have called CMCA, (non-standard Arabic script writing; see 4.1.2). They were asked to first read these sentences orally into the recording microphone, then to write the sentences out on a smartphone using the Roman script of LQA CMCR, as if they were texting them to a friend. In Part 2, I read out another six sentences to the participants myself in my own voice (in a broadly New Tripolitan accent), and they were then tasked with first repeating these same sentences orally, and then writing them out on the smartphone in the same manner as in stage one, using LQA CMCR. Both mine and the participant's readings were recorded. This means that, across these two stages, we have for each of the 49 participants a total of 12 written CMCR sentences (with a total of 127 tokens per participant), as well as 6 voice recordings produced by the participants of the same six sentences from Part 1. Though in reality there are occasional differences even in content (with a small number of words changed, omitted or misheard), this also means that each token is (mostly) replicated at least 49 times, across the participants (with some words appearing more, on account of appearing more than once within the 12 sentences; <inshallah> and variants shows up a full 5 times within the 12 sentences, for a total of about 245 written tokens across all participants). Further detail on the methodology of these interviews can be found in the introduction to the next chapter (10.1.2), where we move on to explore the possibilities afforded by the ability to compare written LQA CMCR tokens with the equivalent spoken LQA realisations of the same individual. In this chapter, however, we use the new written data of Dataset 2 to map out the graphemic word-space of Tripolitan LQA CMCR, focusing in particular on the most commonly used highly-variable phonemic features, and in doing so we further develop the Lexemic-Aggregational model developed in the course of the previous chapter. We will also be able to re-examine and further develop our conclusions from Chapters 7 and 8 on the various means of understanding both the variational structure and conventionalisational potential of LQA CMCR, and the factors and effects that we have demonstrated to affect this process.

9.1.2 Points of Difference: Dataset 1 vs. Dataset 2

I. Dates of Data

There are important points of difference to consider between our two datasets. Dataset 1 consists of comments collected digitally from three publicly-accessible Facebook groups, produced between 2012 and 2015, while Dataset 2 consists of data gathered from 49 individual participants in October of 2016. Though the newest data from Dataset 1 is at most a year apart from the data of Dataset 2, there is nevertheless a gap between Dataset 2 and the older data of Dataset 1, in which changes in conventions, trends and tendencies are very much possible, particularly given the dynamic, non-standard nature of LQA CMCR.

II. Age of Participants

Though it is not possible to determine the ages of the users who produced the comments on the Facebook groups of Dataset 1, the ages of the participants in my experimental interviews of Dataset 2 were collected as part of the experimental procedure, and can be seen below:

Table 9A – *Age of Dataset 2 Participants*

18-21	22-25	26-30	31-35	36-40	41+
15	13	13	4	1	3

The data of Dataset 2 is therefore somewhat biased towards the younger age-groups, whereas the Facebook groups are likely to have had a wider age-span, if not one biased in the opposite direction and away from the younger age-groups, particularly considering the time-span in which Dataset 1 was collected, by which time Facebook was no longer primarily used by younger age-groups as it had been at its inception.

Table 9B – *Percent of Internet users who use Facebook, by age, 2012*

18-29	30-49	50-64	65+
86%	73%	57%	35%

Reproduced from Duggan and Brenner, 2013

The above table (from Duggan and Brenner, 2013) demonstrates that there is likelier to be a higher spread of ages on Facebook than appears in our Dataset 2 participants. This data

does not provide absolute numbers, given its presentation of the percentage of Facebook users per age-group for all internet users, nor does it take into account geo-cultural factors (such as whether these numbers are also applicable for Lebanon), nor more recent developments (such as likely changes in these ratios between 2012 and 2015). Finally, we would also expect participants on Facebook groups dedicated to sharing news in Tripoli to likely to be older than average. Ultimately, the likely age difference across the two datasets must be considered as a factor in any difference between the results of the two datasets.

III. Implicit Pressure

Because of the nature of the interviews of Dataset 2, whereby participants either read or heard a short sentence aloud, and then rushed to type it into their smartphones while they still retained it in their mind, all while I waited for them to complete this task, there was a detectable sense of pressure on behalf of the participant to finish the task quickly. This pressure was implicit, as I did not provide any explicit time limit and did nothing to rush them, but nevertheless it meant that the majority of participants wrote relatively quickly, and ultimately in a manner that would closely resemble how they might usually write when communicating with friends and others. In Dataset 1, however, it is much less likely that this kind of pressure would have been present, and for longer comments, some users may have been inclined to check through what they have written before posting (though this was patently not always the case).

IV. Media Type

Though smartphones are now used for all kinds of digital communication, towards the earlier part of my data from Dataset 1 it would have been likely that a reasonable proportion of participants were using computers instead. In Dataset 2, however, the exercise was conducted entirely through smartphone, which has potential effects on the data, such as the fact that digital keyboards are smaller and often require a click to switch between different modes (such as alphabetical and numerical), which could have a bearing particularly with regards to the use of numerical graphemes such as <7>, <8> or <5>.

V. CMC Genre

Ultimately, the interviews of Dataset 2 simulated a form of digital communication closer to synchronous CMC than was the case in the Facebook comments of Dataset 1, which are non-synchronous, or potentially semi-synchronous. This is reflected through implicit pressure in Dataset 2 and ultimately is likely to also contribute to the differences between the data collected for each dataset. Though there is nevertheless plenty of similarity between the data in both datasets, considering the same broad genre of CMC, the same geographical location and linguistic community and the fact that the dates of the collected data are not significantly distant, the above factors must nevertheless be taken into account when considering Dataset 2 relative to Dataset 1, particularly where we find major divergences.

9.2 The Writing of Consonants

We now use the high-token, highly-repeated and semantically transparent data we are afforded by Dataset 2 to further map out the graphemic space of Tripolitan LQA CMCR, as well as to review our findings so far using new data. Though the scope of this study does not allow for the definition of an exhaustive catalogue of the orthographical representations of every sound within LQA, we will nevertheless be able to use this approach as a framework for better understanding the variation that we have identified within LQA CMCR, and by focusing on the most common and highly variable features, we will be able to use the Lexemic-Aggregational model to predict variability within words on the basis of the points and degrees of variation within its phonemic make-up. We begin by considering the consonants we have discussed thus far in the context of Dataset 1 and examining how these are represented in Dataset 2.

Table 9.1 – *Consonants Frequencies - Dataset 1*

/ʃ/	**<sh>**	**<ch>**
Dataset 1	**55%**	**45%**

/ħ/	**<7>**	**<h>**
Dataset 1	**71%**	**29%**

We have already developed overall frequencies for two of our central features (Table 9.1), having determined in Chapter 7 that /ʃ/ is split between StF <ch> and StE <sh> realisations and in Chapter 8 that /ħ/ is split between specific <7> and ambiguous <h>, noting that these overall ratios cannot be taken alone, however, but must be understood in the context of the other factors influencing users' choice of representation for them. We also discussed the representation of the velar fricatives /x/ and /ɣ/ (see 7.2.2 I), for which sounds and others we will now be able to determine clearer ratios in the course of our analysis of the data of Dataset 2, which now follows.

9.2.1 French vs. English: The Voiceless Alveolar Fricative

Table 9.2 – *Both Datasets*

	Dataset 1		**Dataset 2**	
	Tokens	%	Tokens	%
<ch>	**635**	45%	**451**	**40%**
<sh>	**782**	55%	**681**	**60%**
Total	*1,417*		*1,132*	

We find in Dataset 2 a similar overall split between <sh> and <ch>. Despite a higher frequency further in favour of <sh> over <ch>, the overall pattern remains the same: a visible preference for the English-derived form <sh> whose popularity therefore stands between 55% and 60% across the datasets. The results from both datasets are contrary to Abu Elhija's (2012) findings, who determines a ratio of 57% in favour of <ch> to 43% for <sh> (Abu Elhij'a, 2012, 95), though her data consist of a total of 107 tokens (compared to our 1,417 for Dataset 1 and 1,132 for Dataset 2) and, more importantly, is likely to consist of data from Beirut and other locations, (though she does not specify which region her participants hailed from; ibid: 71), where StF is more likely to be perceived as prestigious (see 11.2.1), and once again highlighting the importance of the focus of this manner of study on a specific and clearly-specified community (see 5.3.3). In our case, we also find that the same Dataset 2 ratio of 60:40 for <sh>:<ch> broadly holds for words containing consonant /ʃ/ but not vowel /u/ (and thus neither <u> or <ou>, which we examine individually in 9.3.1). Simplifying the variation of other phonemes within these words, we produce the following table showing the variation of <sh> and <ch> within the resulting word-sets (for which a full table can be found in the appendix, Table 9.3X):

Table 9.3 – Word-Sets containing **/ʃ/** but not **/u/** - *Dataset 2*

<*shi>	<*chi>	Total Tokens	Tokens per Participant	Word Meaning
86	**60**	146	3[18]	*"Something"*
59%	***41%***			

<*shaghle>	<*chaghle>	Total Tokens	Tokens per Participant	Word Meaning
57	**38**	95	2	*"Thing* [f.]*"*
60%	***40%***			

<*lesh>	<*lech>	Total Tokens	Tokens per Participant	Word Meaning
53	**29**	81[19]	2	*"Why"*
65%	***35%***			

The only real difference to the expected ratio appears in the word <lesh>/<lech> (/le:ʃ/, "*why*"), where we see the anticipated 60:40 ratio skew towards <sh> by a further 5% at 65:35. This is also the only instance where the /ʃ/ appears in the word-final position, wherein there is potential semantic ambiguity in the form <lech> that is risks a reading of /letʃ/ using StE orthographical associations, coupled also with the expectation in StF writing of an additional <e> following word-final <ch> (*vache, quiche,* etc), making word-final <ch> for /ʃ/ also unusual in a StF reading. Interestingly, only a single participant opted to fix this by rendering the word <leche> (producing two tokens thereof), while others make use of the variable orthographic resources of LQA CMCR and simply utilise <sh> instead, or else simply retain <ch>, as was the case a majority of cases (29 tokens). It is interesting that a number of users who generally use StF <ch> are potentially aware enough of StF (and even StE) conventions to know that <ch> is problematic as a word-final grapheme, but who resolve this by utilising the flexibility of LQA CMCR to revert to the less-problematic <sh>. Moreover, that most continued to use <ch> in the word-final position without apparent interference by orthographical rules outside of LQA CMCR indicates that for most, new

[18] This indicates how many tokens of this word each participant is expected to produce. In this case, one token is not produced by participant 31, which is why the total tokens is 146 and not the expected 147 (given 3 tokens across 49 participants). Tokens are often not produced or produced in too divergent a manner to be useful, which is why the total tokens is not always exactly equivalent to the expected total given tokens per participant across total participants.

[19] Here there are notably fewer tokens (82 instead of the expected 98) because in addition to one missing token, there are also 15 instances where variant /le:/ appears instead (as <le> or <leh>), a phonetic variant with the same meaning "*why*", but as it does not contain /ʃ/ it is of no use here. It is often the case that some tokens which are useful for the analysis of one sound are not useful for others, depending on how they appear and what phonemes are missing for each token.

(informal and non-standardised) conventions have developed, wherein its users are no longer directly influenced by the rules of either StF or StE, but instead are drawing on the conventions available within LQA CMCR. Thus we conclude with two clear ratios for this sound from both datasets, as well as an in-depth awareness of its graphemic variability and of some of the factors governing this variation. We return to the discussion of StF and StE influence on LQA CMCR in our discussion of /u/ in 9.3.1 to follow later in this chapter.

9.2.2 Voiceless Pharyngeal Fricative: <7> vs. <h>

I. A New Ratio for Dataset 2

Our discussion of the voiceless pharyngeal fricative /ħ/ in Dataset 1 centred around the ratio of 71:29 for <7> to <h>, where we used divergence from this consistent ratio as a means of detecting the effect of various factors. There are a total of 14 words in Dataset 2 which contain the voiceless pharyngeal fricative, some of which appear more than once per session, for a sum of 22 tokens per participant, producing a total of 1,061 tokens across the dataset (a table showing the spread of this data over Dataset 2 can be found in the appendix under Table 9.5X).

Table 9.4 – *Both Datasets*

	Dataset 1		**Dataset 2**	
	Tokens	%	Tokens	%
<7>	**764**	71%	**651**	**61%**
<h>	**307**	29%	**410**	**39%**
Total	*1,071*		*1,061*	

We find in Dataset 2 an entirely new ratio of **61:39** instead of 71:29, more in favour of <h> than was the case in Dataset 1, even if <7> remains the more popular of the two variant graphemes as before. This is despite there being (quite by chance) an almost identical number of tokens, with a difference of only 10 tokens between the two datasets. The divergent ratios across datasets can be explained in a number of ways. The physical media being used to produce tokens is likely to be the primary factor (see 9.1.2 IV): posting a Facebook comment using a computer conceivably allows individuals more time and care to distinguish <7> from <h>, relative to the smartphones used in my experimental interviews, particularly given the requirement on most smartphones to switch to a separate keyboard

screen in order to access numbers, meaning that two additional clicks are required for each switch, while on computer keyboards numbers are generally as accessible as any alphabetical letter. This is in addition to the implicit time pressure within the interviews (9.1.2 III) and thus ultimately the difference in CMC genre between the two datasets (9.1.2 V). Even so, we cannot discount the possibility that there has been a shift in conventions between the data collected for Dataset 1 (spanning the years 2012-2015) and that for Dataset 2, collected in 2016 (9.1.2 I). This might combine with other factors, such as the overall popularity of smartphones over desktop or laptop computers for CMC also contributing to an overall change in the conventions of the orthography, where <h> may have become more common than previously, irrespective of whether the immediate input is via computer or smartphone. Even further, we might even posit that <h> as a conventional representation of the sound has grown in popularity as a result of the very factors we identified in the previous chapter, whereby over time writing becomes less transcriptional and relies more on conventions. In this way, the various factors we determined to encourage the use of <h> over <7> might have had a long-term effect on the ratio we began with. However, without further study, this possibility cannot be pursued further, and we cannot determine whether this new ratio is indeed an artefact of conventionalisation or merely the result of the pragmatic context of the data of Dataset 2, bearing in mind also the constant variation of non-standard writing and even of conventionalisation. We conclude again that the non-standard nature of LQA CMCR very much means that conventionalisation is not a static but ongoing and changeable process that is not subject to the same pressures a standardised orthography, which we do not expect to necessarily move in a single direction, nor retain any ratio in the longer term. With this in mind, and using for the analysis of Dataset 2 the new ratio of 61:39, we now re-examine some of the same factors we discussed in 8.2.

II. Word Position

Table 9.5 – Representation of /ħ/ by Word Position - *Both Datasets*

	Dataset 1			Dataset 2		
	<7>	**<h>**	***<h>% +/- from 71:29***	**<7>**	**<h>**	***<h>% +/- from 61:39***
Word Initial	69%	31%	+2%	61%	39%	0%
Word Medial	72%	28%	-1%	62%	38%	-1%
Word Final	**77%**	**23%**	**-6%**	**60%**	**40%**	+1%
Total	***71%***	***29%***		***61%***	***39%***	

The preference for <7> over <h> in the word-final position we saw in Dataset 1 (8.2.2) is not replicated in Dataset 2. In Dataset 1 grapheme <7> appeared 6% more than predicted by the 71:29 ratio; in Dataset 2, no position differed by more than 1%, including the word-final position (see Table 9.5X in the appendix for each word per word-position). Again here we cannot discount the potential effect of the smartphone-mediated, time-pressured nature of Dataset 2, leading participants to prioritise writing speed over taking the time to specify using <7>, even in otherwise sensitive contexts such as the word-final position. We also note the fact that of 138 the total 186 tokens of word-final /ħ/ in Dataset 2 are instances of the word <*mbere7>/<*mbereh> (meaning *"yesterday"*, appearing three times per individual session), a word for which the ending <h> is potentially less problematic given the lack of any real semantic or phonetic misreading, meaning again that the pressure to delineate with <7> is lessened, potentially also by word-frequency convergence. This is perspective is bolstered by the data for <*mne7>/<*mneh> ("[*They*] *are well, good"*), which provides the remaining 48 word-final /ħ/ tokens in Dataset 2, and for which word (alone) we do find an additional 4% <7> than the 61:29 ratio of Dataset 2 predicts, meaning that it is possible that the word-final effect would have been replicated in Dataset 2 had the tokens come from a wider pool of words showing the sound in this position.

Table 9.6 – Words with Word-Final /ħ/ - *Dataset 2*

<7> Form	**Tokens**	**%**	**<h>%** [*from 61:39*]	**%**	**Tokens**	**<h> Form**
***mne7**	**31**	**65%**	**-4%**	**35%**	**17**	***mneh**
***mbere7**	**82**	**57%**	**+4%**	**43%**	**61**	***mbereh**

III. Semantic Overlap

We determined semantic clarity to be an important consideration in Dataset 1 (8.2.3), wherein a proportion of users preferred to distinguish their intended semantic meaning using the specific form <7> rather than ambiguous form <h> in cases where <h> is more easily misread as /h/ (or not read at all). We found this to be true for the form <hada>, which read with the pharyngeal fricative means "*someone*", while with a glottal reading instead indicates "*that, that one* [m.]". This same word appears four times per participant in Dataset 2, and so has 195 tokens within the dataset. The effect here is slightly less pronounced, with an increase of 2% in favour of specific <7>, either meaning that the effect of semantic pressure is not a major factor in the context of Dataset 2, or else the lower tendency to use <7> here can again be explained as the result of the smartphone-based and time-constrained nature of the data-gathering process for Dataset 2, which broadly assigns more value to the expediency of producing <h> over the specificity of <7>, and ultimately makes more use of conventional rather than transcriptional writing.

Table 9.7 – *Dataset 2, "Someone"*

<7> Form	Tokens	%	<h>% [*from 61:39*]	%	Tokens	<h> Form
***7ada**	**122**	**63%**	**-2%**	**37%**	**73**	***hada**

IV. A Synthesis of Factors

As in the previous chapter, here again we conclude that even where the effect of individual factors is more difficult to discern, these factors can be more clearly identified when working in tandem (or even in opposition). Forms like <habibi>, where multiple factors are at play, give us the most information about how users of LQA CMCR represent this sound (and, in turn, about what factors might affect conventionalisation within the orthography). In the case of <habibi>, we were limited in our analysis of Dataset 1 (8.2.4) by the low number of tokens of the word (though even then we still saw a significant effect in the reversal of the ratio from 71:29 all the way to 30:70). Now in Dataset 2, we have a full 48 tokens of the word (which appears once per session), and thus can now make firmer conclusions about how this word functions. As discussed prior, <habibi> is widely-used and recognisable even within the Anglosphere where it appears frequently with a relatively stable spelling of <habibi>. We predicted a reversal of the ratio for this word on the basis a

combination of factors affecting grapheme-choice: its common appearance in the Roman script, the fact that no semantic ambiguity is introduced by the use of <h> instead of <7> (its glottal realisation not being semantically meaningful), and its relative frequency of use (within synchronous communication between friends and family, hence its lower token count in the semi-synchronous communication mostly between strangers in Dataset 1). The frequency factor alone, as we saw in 8.2.1, and as we have developed within our Lexemic-Aggregational model, is capable of leading to convergent conventionalisation in cases where there is sufficient invariability in the individual positions of a given word.

Table 9.8 – *"My dear"* – Dataset 2

Total	<7> Form	Tokens	%	<h>% [*from 61:39*]	%	Tokens	<h> Form
38	**7abibi**	**16**	**42%**	**19%**	**58%**	**22**	**habibi**
1	7abebe	0	0%		100%	**1**	habebe
1	7abebi	1	100%		0%	**0**	habebi
1	7abibe	1	100%		0%	**0**	habibe
5	7bb	2	40%		60%	**3**	hbb
1	7b	0	0%		100%	**1**	hb
1	7abb	0	0%		100%	**1**	habb
48	***7abibi**	**20**	**42%**	**19%**	**58%**	**28**	***habibi**

We saw no clear convergence effect for this word in Dataset 1 due to the low total tokens, but now with our 48 tokens of Dataset 2, we see the predicted convergent form <7abibi>/<habibi> emerging, concurrent with an exact reversal in ratio (that we also saw in Dataset 1), in this case 58% in favour of <h> (instead of the usual ratio 61% in favour of <7>, which here is no longer preferred). The combination of high-frequency use, the common use of this orthographic form outside of a purely LQA CMCR context, the minimal points of high variation and the lack of semantic ambiguity within its construction leads to both orthographic convergence and preferential use of conventional <h>. We see therefore the clearest effects when our various factors combine, understanding the ratios of our two datasets (71:29 and 61:39) to only be useful in tandem with an understanding of the further socio-linguistic factors governing LQA CMCR users' choice of graphemic representation.

9.2.3 The Velar Fricatives: <kh>/<gh> vs. <5>/<8>

Table 9.9 – *Dataset 2*

/x/			/ɣ/		
<kh>	**<5>**	***Tokens per Participant***	**<gh>**	**<8>**	***Tokens per Participant***
157	135	6	97	81	4[20]
54%	***46%***		***54%***	***46%***	

The velar fricatives /x/ and /ɣ/ occur (by design) more frequently in Dataset 2, allowing us to better examine their variable representation between digraphs <kh> and <gh> and numerical graphemes <5> and <8>. We see in Table 9.9, quite strikingly, that the ratios between numerical and digraphic representations for both the voiced and voiceless velar fricatives are identical, despite appearing in different words and at different relative rates (a total of six words containing /x/ and four of /ɣ/ per participant in each interview session).

Table 9.10 – *Dataset 2*

	<kh> <gh>	**<5> <8>**	**Mixed**
Participants	21	18	10

We also find that it is relatively rare for users to mix digraphic and numerical resolutions: of the total 49 participants, 18 used <5> and <8> exclusively, and 21 used <kh> and <gh> exclusively, while the remaining 10 participants mixed between digraphic and numerical across the velar fricatives. Considering the degree to which we have observed other forms (such as <sh>/<ch>, <ou>/<u>, and <h>/<7>) being used interchangeably by the same individuals and sometimes even within the same texts, it is notable here that one resolution is used so exclusively. In Chapter 7 (7.2.2. I) we posited the possibility that numeric forms <5> and <8> are likely to group with French rather than English orthographical features, on the basis that digraphs <kh> and <gh> are used more frequently in StE than StF. We found a possible effect for the representation of /x/ but not for /ɣ/ and could not draw meaningful conclusions on the basis of the low token count, but we can now conduct a new analysis for these variables using Dataset 2. Considering how strongly the numerical or digraphic

[20] The low total tokens for /ɣ/ considering its 4 tokens participants is due to the presence of minority forms (8 tokens of <g>, 4 tokens of <3'>), as well missing tokens and likely typos such as one instance of <5er> instead of <8er>

representations cluster together respectively, we now divide the remaining participants into two groups, one comprising the 18 participants who used numerical graphemes only, and the 21 participants who used the digraphs only, within which we test for the ratios of StE and StF-derived orthographical features (as compared to the overall ratios of <sh>/<ch> which we discussed in 9.2.1, and <u>/<ou> which we will discuss in depth later in 9.3.1).

Table 9.11 – French and English Features & the Velar Fricatives - *Dataset 2*

Ratios Across All Dataset 2

<sh> *English*	**<ch>** *French*
681	451
60%	***40%***

<u> *English*	**<ou>** *French*
308	226
58%	***42%***

DS2 Ratios for <kh>/-<gh> Group (English)

<sh> *English*	**<ch>** *French*
335	156
68%	***32%***

<u> *English*	**<ou>** *French*
152	81
65%	***35%***

DS2 Ratios for <5>-/<8> Group (French)

<sh> *English*	**<ch>** *French*
189	223
46%	***54%***

<u> *English*	**<ou>** *French*
81	119
41%	***60%***

Testing for both the primary divergent features between English and French, we see a meaningful effect: in the digraphic group <kh>/<gh>, the frequency of English features is higher than predicted, with <sh> at 68% and <u> at 65%. More tellingly still, the <8>/<5> grouping actually sees the strong preference in Dataset 2 for English-based features overturned, with French <ch> at 54% and French <ou> at 60%, indicating very clearly that StF features are popularly used by the same individuals who represent the velar fricatives with numerals instead of digraphs. This in turn confirms our previous hypothesis that the variation in the representation of the velar fricatives is related to the divergence between StE and StF sound-symbol correspondences, and as such are part of the harmonic effect we posited in Chapter 7. As we determined in our discussion of these harmonic preferences,

only a sub-section of users combine harmonic features in this way, and despite this preference showing strongly when selected for, it is not possible to construct a full sub-orthographical perspective (see 7.4). Nevertheless, it is an important factor to consider in the context of our frequency ratios and what factors affect them, in the same manner as other effects such as semantic clarity or word-position. Besides this, it is also notable that the velar fricatives themselves (in the majority of instances) group together either as numerical or digraphic representations, even by individuals who intermix other correspondences derived from StF and StE writing, indicating that a link exists for Tripolitan LQA speakers between the two sounds and their graphemic resolutions.

9.2.4 Gemination & Reduplication

We close our treatment of consonants with a feature we have not yet examined: the representation of geminate consonants. In Table 9.12 below, we see all words from Dataset 2 that exhibit geminated consonants, divided by whether this gemination is represented by a single (<C>) or reduplicated (<CC>) grapheme:

Table 9.12 –Representations of Gemination – *Dataset 2*

	<C>	**<CC>**	***<C> %***	
/ʔlaj.ji/	38	7	*84%*	*"On me, to me"*
/rəd.dəl.li/	29	14	*67%*	*"Answer me, get back to me"*
/rəd.dəl.li/	37	6	*86%*	*"Answer me, get back to me"*
/t'aj.jəb/	37	10	*79%*	*"Alright; so"*
/xal.li:ni/	39	9	*81%*	*"Le me, allow me"*
/əj.je:m/	41	7	*85%*	*"Days"*
/mən.non/	36	11	*78%*	*"From them, of them"*
Total	**257**	**64**	***80%***	
	80%	**20%**		

In addition to the overall 80:20 ratio (in favour of single rather than reduplicated consonantal graphemes) that emerges from within this data, we see in the case of /rəd.dəl.li/ a word with two instances of gemination (/d.d/ and /l.l/), leading to variation in both positions. Due to the relatively low variability in other positions and the fact that gemination is predicted at 80:20 overall, we still see a convergent form emerging as <redeli>

with a majority of 16 tokens (with the next most popular spelling among the remaining 27 tokens showing only 5 tokens).

Table 9.13 – Graphemic-Phonemic Breakdown – *"Get back to me"* – Dataset 2

IPA	Variant	Variant%
r	**r**	**100%**
ə	**e**	**86%**
	i	12%
	ø	2%
d.d	**d**	**67%**
	dd	33%
ə	**e**	**65%**
	i	23%
	ø	12%
l.l	**l**	**86%**
	ll	14%
i	**i**	**91%**
	e	7%
	ee	2%

Outside the two geminate consonants, the greatest variation comes from the middle schwa vowel, which in fact sees higher variation (65% in favour of <e>) than the first schwa vowel (86% in favour of <e>), along with the fact that the first geminate consonant /l.l/ sees a more variable ratio at 67:33 than both the second geminate (at 86:14) and also the overall 80:20 ratio. This is in fact the most highly variant ratio in all our data, as the other geminated consonants in Table 9.12 above vary only between 78% and 85% in favour of single-consonantal graphemic representation, and as such most likely indicates an effect within words that contain two instances of gemination. This is the only instance of such a word in Dataset 2, but we hypothesise that the first of two geminated consonants is likelier to see greater reduplication (and thus higher variation) than otherwise predicted. We see here (and in the difference between the two schwas of this word too) a potential limitation in how far we are able to use our frequency-system to predict orthographic convergence,

short of describing every possible position for each phoneme and producing a full frequency charting for each– an endeavour that unfortunately lies outside the scope of this study. Nevertheless, we conclude with not only a ratio for the representation of geminates, but also a potential factor affecting this ratio in the form of double-geminated words.

9.2.5 Consonants: Conclusions

We have therefore established ratios (in some cases from both datasets) for the representation of the consonants we have identified to be the most variable within the writing of LQA CMCR. We have also identified a range of further factors that affect and alter these ratios, whether on the basis of the factors we predicted to affect conventionalisation, or indeed on the basis of such things as repeated sounds across different word-positions. These generalised ratios provide a powerful tool for understanding variation in the particular context of convergent conventionalisation and whether (and to what degree) the process is likely to take place in any given construction based not only on the frequency of appearance of that construction but specifically also on the number of variable phonemes within it, and how much they vary. We finally note that the remaining consonants used in LQA (and written in LQA CMCR) are largely invariable (including <r>, <t>, <d>, <z>, <s>, <k>, <f>, <m>, <n>, <l>, <h> and others), in part as a result of the boundaries established by the basis of LQA CMCR on the standard orthographies of StF and StE which largely have a single resolution for each of these sounds, which is adopted into the orthography of LQA CMCR without further problematisation. Even the voiced pharyngeal fricative /ʕ/, which has no representation in StE or StF writing is simply resolved using the character <3> without running into the same issues as its voiceless cousin /ħ/ as it is not additionally associated with a standard Roman script grapheme. There are occasional instances of /ʕ/ being written using vowel reduplication (such as <aa> for /ʕa/) that is reminiscent of syllabaries, but this is a rare resolution in Tripolitan LQA CMCR which does not truly impact the variation seen in the orthography outside of a small number of outliers. The only remaining question here is that of the emphatic consonants, which we hypothesised (see 6.2.5 II) to be used in Tripolitan LQA speech but not represented in writing except very rarely, contrary to Abu Elhij'a's (2012) conclusion that they are absent in both spoken LQA and written LQA CMCR, bearing in mind that this may hold true in other variants of LQA (such as the Beiruti LQA

urban capital dialect Abu Elhij'a is likely to have studied). While certainly spoken Tripolitan LQA like most other LQA (and many Levantine QA) dialects does feature the sound change from the emphatic /kˤ/ of SA to glottal stop /ʔ/ in most positions, we still expect the other emphatic consonants (/tˤ/, /sˤ/, /ðˤ/ and /dˤ/) to be marked in the speech of Tripolitan LQA speakers. That these sounds are not represented in writing is already abundantly clear in our work, as we have not encountered any attempt at orthographically marking emphatic consonants whatsoever; as for whether they are marked in spoken Tripolitan LQA, this will be addressed in Chapter 10 (10.3.2) to follow. For our present purposes, it suffices to say that both emphatic and non-emphatic consonants are written alike.

9.3 The Writing of Vowels

In Chapter 7, and using Dataset 1, we determined frequencies for the vowel /u/ on the basis of French <ou> and English <u> (7.2), and we have also devised ratios for the frequency of vowel omission depending on vowel length (7.3). Now we reconsider this work not only in the context of Dataset 2, but also in the specific framework of our frequency-based Lexemic-Aggregational approach. We will also determine specific frequencies for each of the remaining major vowel sounds and their graphemic variations, noting that we cannot rely on the total graphemic counts as we did with many of the consonants, since factors like vowel length are seldom distinguished in writing, in addition the complexity of things like word-position, which are more impactful in the case of vowels. We therefore rely more extensively on identifying words containing the relevant sounds and aggregating the representations of each sound, though we begin by reviewing our work on StF and StE.

9.3.1 French vs. English: /u/ and Word-Final /e/

I. The Representation of /u/ and /u:/

Table 9.15 – *Both Datasets*

	<u>	**<ou>**	
Dataset 1	915	1,466	***Total*** *2,381*
	38%	***62%***	

	<u>	**<ou>**	
Dataset 2	308	226	***Total*** *534*
	58%	***42%***	

The phoneme /u/ was most frequently represented in Dataset 1 with French-derived <ou>, showing 62% usage compared to the 38% of <u> (see 7.2.1 and Table 7.7). This is not so in Dataset 2: in fact, this ratio is overturned, where <u> now appears with a majority frequency of **58%** (308 tokens), with <ou> at 42% (226 tokens). This transformation in popularity most likely runs along the same lines as the (still higher) popularity of <sh> over <ch> in Dataset 2 (discussed in 9.2.1 prior), and both of these shifts in favour of StE-derived forms are due to the younger ages of the participants in the experimental interviews (and thus, their leaning towards StE over StF), itself likely linked to the gradual shift from StF to StE as the primary second language of Lebanon (see the discussion in 6.1.2). It is difficult to determine whether this is an effect of a generalised orthographical shift over time, or a result of the relatively younger ages of the participants in Dataset 2, though even if we conclude that it is generational, the general tendency towards <u> among younger users, if confirmed more widely, naturally still indicates a shift in the written conventions of Tripolitan LQA CMCR over time (even if the usage of members of older generations does not itself shift), ultimately being a natural part of the organic orthographical flexibility and changeability of non-standard writing. Nor can we discount the fact that, despite the high number of tokens (which for most words matches or exceeds the tokens available in Dataset 1), the pool of individuals contributing spellings in Dataset 2 is significantly smaller, with a far higher share of tokens for each word coming from the same individuals (totalling 49), which serves as another explanation for any variations from Dataset 1. Ultimately, in Dataset 2 the two highest-popularity graphemic resolutions for both /u/ and /ʃ/ are StE-derived <u> and <sh>, while in Dataset 1 we observed a split between StF-derived <ou> and StE-derived <sh>. Finally, we also note that we have found no real variation between /u/ and /u:/, nor any attempt to distinguish between them in either dataset.

Table 9.16 – *"What" – Both Datasets*

		Harmonic Forms		
	Mixed	**French**	**English**	Mixed
	<shou>	**<chou>**	**<shu>**	**<chu>**
Dataset 1	21	**49**	**45**	25
	15%	***35%***	***32%***	*18%*
Dataset 2	37	**53**	**92**	30
	17%	***25%***	***44%***	*14%*

We examine the new dynamics of our new dataset further by again analysing how the primary features /ʃ/ and /u/ combine, as we did for Dataset 1 using the word "*what*" which comprises these two features precisely and exclusively as **/ʃu:/**. We recall that there was a clear tendency in Dataset 1 for the use of harmonic forms (see again 7.2.1), and though mixed forms were also clearly present, they showed about half the popularity of the harmonic forms in our original dataset. Comparing this to Dataset 2, it is not surprising that we find that StE-harmonic form <shu> is now significantly more frequent than StF harmonic form <chou>, given that both StE features are highest-frequency in Dataset 2. In light of this, however, it is quite remarkable that harmonic StF form <chou> still appears in second place at 25%, despite both its constituent graphemes <ch> and <ou> being the least-popular forms for their respective phonemes. When this 25% frequency is compared to the 14% of <chu> (using overall most popular form StE <u>) and the 17% of <shou> (using overall most popular form StE <sh>), we see clear indication of the strength of the orthographical link between the forms derived from the same standard orthography: individuals who do not opt for harmonic (and highest-popularity in both phonemes) <shu> are likelier to opt for harmonic (but least-popular in both individual phonemes) resolution <chou> than they are to opt for non-harmonic resolutions <shou> and <chu> even if these contain one highest-popularity and one lowest-popularity graphemic resolution each. We see thus that the tendency towards harmonic forms remains largely intact in this new data even in spite of this greater preference for StE conventions across both variables, and thus we have now observed this harmonic tendency across both our datasets in the case of the word *"what"*.

II. The Representation of Word-Final /e/

We have previously considered the use (or even perceived need to use) a word-final <h> to be a factor in the preference for marking the voiceless pharyngeal fricative with a specific <7> in the word-final position, given that <h> is also used in word-final constructions such as <eh> for /e/(see 8.2.2). We also hypothesised that the use of <eh> instead of <e> in this word-final position is potentially preferred in StE convention but not StF, which hypothesis we now further explore using Dataset 2. The word-final /e/ shows the following overall ratios, showing a clear preference for <e> over <eh> across the dataset:

Table 9.17 – Word-Final /e/ across All Tokens – *Dataset 2*

	<eh>	**<e>**
Dataset 2	83	144
	37%	***63%***

Among the words containing word-final /e/ (for which a breakdown can be found in Table 9.17X in the appendix) is the word /ʃaɣle/ (meaning "*thing, something* [f.]*"*), which is useful to our analysis as it also contains /ʃ/, allowing us to directly investigate the link between <sh>/<ch> and word-final <eh>/<e>. Unlike /ʃu:/ which contains no other sounds, this construction contains another important variable /ɣ/, which here we simplify to <gh> (in addition to simplifying vowels to <a> and <e> with no omission) in order to create four word-sets based around the variables we are interested in (<sh>/<ch> and <e>/<eh>). Running the same analysis on these word-forms, as we have done prior for <shou> and its variants, and so considering <chaghle> and <shaghleh> to be fully harmonic spellings and <chaghleh> and <shaghle> to be mixed spellings, we find the following:

Table 9.18 – *"Thing, Something [f.]" Word-Sets - Dataset 2*[21]

Mixed	**French**	**English**	Mixed
<*chaghleh>	<*chaghle>	<*shaghleh>	**<*shaghle>**
12	23	22	**35**
13%	*25%*	*24%*	***38%***

Mixed form <*chaghleh> is least popular, harmonic forms <*chaghle> and <*shaghleh> have nearly identical frequencies of 25% and 24% respectively, but the most popular form

[21] Excluding 2 tokens of <cha8li> and 1 of <cha8lh>, given their alternative endings.

by far is the mixed form <*shaghle>, at 38% frequency. Similarly to how English <sh> and French <ou> were the most popular respective forms in Dataset 1 (and thus indicated possible conventionalisation of these features beyond the original standard orthographies they derive from; see Table 7.7 and discussion in 7.4), here too the word-ending <e>, as the most popular individual form, does not most frequently couple with StF features, but rather has come to be used independently from its original orthographic derivation of StF. That the word-form <*shaghle> is the most common form, therefore, is a result of the overall popularity of both StF-derived word-final <e> and StE-derived <sh> independently. That harmonic forms <*chaghle> and <*shaghleh> have a similar share at 25% and 24% respectively (and more than mixed form <*chaghleh> at 13%) may also suggest that here too there remains a leftover effect of orthographic harmony, though we cannot be certain of this simply because this is also feasibly explained by the fact that each of these forms contains one of the most-common features for their respective positions (<e> in <*chaghle> and <sh> in <*shaghleh>). Ultimately, we add the ratio of **63:37** in favour of <e> to our ratios for <ou> and <u>, while noting the harmonic effect is less pronounced in this case, and <e> is the preferred form overall. Ultimately, we reach the same conclusions with regards to StE and StF features, which retain in certain cases and for certain users a harmonic connection to a single standard orthography, but where the overall tendency in Tripolitan LQA CMCR is towards admixture as resources within a new, emergent non-standard orthography, within which the use of these resources- irrespective of their derivation- becomes conventionalised as a result of high-frequency convergence, though as we also witness in the change from Dataset 1 to Dataset 2, conventions (unlike standardised prescriptions) are highly variable and changeable over time.

9.3.2 Short Vowels & Schwa

In Chapter 7 (section 7.3), we used simple, single and short-vowel common forms to measure vowel omission in Dataset 1. Now with Dataset 2 we are able to examine in more detail every major individual vowel sound, as well as identifying more clearly where (and to what extent) omission takes place across all words and not only the common marker words (such as /ʕal/ or /bəl/), in addition to which we also review the change in frequency in these marker words between the datasets. We begin with a detailed analysis of short vowels as

they appear in Dataset 2, before discussing the role of vowel omission, and finally moving on to an analysis of long vowels.

I. Short Vowels /i/ and /ɪ/

Table 9.19 – Word-Initial /i/ – *Dataset 2*

	<e>	<i>	
/iza/	8	41	*"If"*
/iʒi/	32	17	*"I come"*
Total /i/	**40**	**58**	
	41%	**59%**	

Table 9.20 – Word-Medial /i/ – *Dataset 2*

	<e>	<i>	
/tiʒi/	31	18	*"You come"*
Total /i/	**31**	**18**	
	63%	37%	

In both the word-initial and word-medial position for /i/ we see a sparse number of words and tokens, primarily on account of the fact that in the majority of cases this short /i/ turns to schwa; we posit here that it is the effect of the voiced fricatives /z/ and /ʒ/ following the sound that leads to the retention of the realisation /i/ (though we need further data to be certain of this conclusion). We see that word-initially, there is a slight preference for <i> by 59%, whereas word-medially we see a similar preference, except for <e> at 63%- both of which are low majorities and likely thus to result in a reasonable degree of variation (and conversely, less likely to lead to frequency-convergence for words containing these sounds).

Table 9.21 –Word-Final Short Vowel **/i/** – *Dataset 2*

	<i>	<e>	
/tiʒi/	46	3	*"[You] come"*
/iʒi/	46	3	*"[I] come"*
/xal.li:ni/	45	3	*"Let me, allow me"*
/fi:ni/	45	3	*"I can"*
/ħaje:ti/	43	4	*"My life"*
Total /e:/	**225**	**16**	
	93%	7%	

In the word-final position, short /i/ converges far more clearly on grapheme <i>, with only a minority of representations as <e>, though based on recent experience, the Beiruti LQA pronunciation of this sound as /e/ (such as in /ħaje:t**e**/ instead of /ħaje:t**i**/) is becoming increasingly exaggerated, functioning as a short-hand marker of the prestige of the urban capital Beiruti LQA dialect and thus being adopted by LQA speakers across the country. On the basis of anecdotal observation, this is also widely mimicked orthographically in the use of <e> for the writing of this phoneme across all Roman script LQA writing, and we expect it to already be impacting the writing even of Tripolitan LQA CMCR. Should this same analysis be run again on data from the present day, we should expect to find different results for word-final /i/, with noticeably higher <e> tokens than appears here from our 2015 data. While change in spoken language (prestige-based or otherwise) occurs even within SLC, the non-standard nature of an orthography such as LQA CMCR allows for conventional representation of these changes in writing without a concurrent loss of prestige associated with abandoning the standard orthographical form (even if it is to mimic a prestige spoken form).

Table 9.22 – Short Vowel /ɪ/ – *Dataset 2*

	<e>	**<ø>**	**<i>**	
/se:kɪt/	44	3	1	"*[He is] quiet*"
/we:ħɪd/	65	26	2	*"One, someone"*
/mbe:rɪħ/	95	32	6	*"Yesterday"*
Total /e:/	**204**	**61**	**9**	
	75%	22%	3%	

When /i/ appears as the final vowel before a final consonant, and is itself preceded by long vowel /e:/, we do not find /i/ but rather a realisation closer to /ɪ/ (as in /se:kɪt/ "[*He is*] *quiet"*, /we:ħɪd/ *"one, someone* [m.]*"* and /mbe:rɪħ/ ("*yesterday"*). Though we have thus far used generalised and not phonetically precise notation, the distinction in this case requires us to classify this occurrence separately (and more accurately) as /ɪ/. We note that this vowel sees a high omission rate at 22%, even if the majority representation remains <e> at 75%, and where <i> forms only a minor alternative appearing in only 9 tokens (and so at 3%).

II. Short Vowel /o/

Table 9.23 – Short Vowel /o/ - *Dataset 2*

	<o>	<u>	<ou>	<ø>	
/mən.n**o**n/[22]	39	6	1	0	*"From them"*
/mənk**o**n/	39	3	1	4	*"From you [pl.]"*
/maʕk**o**n/	39	4	2	4	*"With you [pl.]"*
/bənʔ**o**z/	41	2	1	0	*"I get a fright"*
Total /e:/	**158**	**15**	**5**	**8**	
	85%	8%	3%	4%	

Short vowel /o/ appears in four words of Dataset 2, its presence in LQA being primarily a reflex of SA /u/ in the word-final position, usually in the second-person masculine plural inflexion such as <مِنْكُمْ> (*"from you"*, SA /minkum/, becoming /mənk**o**n/ in LQA) or first-person verb-endings such as <آكُل> (*"I eat",* SA /ʔa:kul/, becoming /ʔe:k**o**l/ in LQA), and where in both cases this sound is reflected by a diacritic <ُ> in SA writing. This leads to the representation of the sound by users of LQA CMCR as <u> or <ou> (a total of 11%), even if this sound is almost never realised as /u/ in Tripolitan LQA. The majority do utilise <o> at 85%, however, and we note that word omission rates are low for this sound, with only eight tokens total (thus 4%).

III. Short Vowel /a/ - Word Final & Initial

Table 9.24 – Word-Initial /a/ – *Dataset 2*

	<a>	<2a>	<2ø>	
/**a**na/	49	0	0	*"Me, I"*
/**a**hwe/	36	10	1	*"Café, coffee"*
/**a**ħsan/	91	7	0	*"Better"*
/**a**xb:ar/	36	11	0	*"News"*
Total	**212**	**28**	**1**	
	88%	**12%**	**0%**	

[22] By way of example, /mən.non/ appears here with 46 total tokens despite showing 47 tokens in Table 9.12 due to one token appearing as <mnin>, with the <i> either anomalous or a typo and therefore of little use for this table as it is not replicated for any other word, but for the use of a non-reduplicated <n> for gemination the form <mnin> was nevertheless useful in the context of Table 9.12 and is therefore included there. This is a not an uncommon occurrence and leads to occasional discrepancies in total token counts when the same words are used for the analysis of different sounds. The word /bənʔoz/ in the same table shows variants <bn2az> (two tokens) as well as one each for <bin2az> and <benaaz>, which indicate phonetic variants and so are omitted in this table, but are used in Table 9.28A where the schwa (which is unaffected by this variation) is examined.

The only real variation for word-initial /a/ (Table 9.24 above) is the additional marking of the glottal stop with a <2>, which occurs (in 12% of cases) likely due to the fact that the glottal stop is marked in in this position in SA writing using the *hamza* <ء>. Word omission does not occur because critical semantic information would be lost in this position, with the exception of a single token appearing as <2hwe> (/ahwe/ (*"coffee, café"*), where the <2> is used as an implicit marker of a following vowel, thus allowing for omission, though this only occurs in one token out of 241. The word-final position (Table 9.25 below) is even more clear-cut, as vowel omission would not only fail to mark semantic information, but additionally cannot be implied by the use of <2> as there is no glottal stop, leading to a full 439 tokens where word-final /a/ is marked precisely as <a>.

Table 9.25 – Word-Final /a/ – *Dataset 2*

	<a>	**<ø>**	
/ħada/	195	0	*"Someone"*
/rəħna/	49	0	*"We went"*
/iza/	49	0	*"If"*
/bale:ha/	49	0	*"Without it, her* [f.]*"*
/hana/	48	0	*"Felicity"*
/ana/	49	0	*"Me, I"*
Total	**439**	**0**	
	100%	**0%**	

IV. Short Vowel /a/ - Word Medial

We find greater rates of omission in the word-medial position, though here variation occurs strictly between either the use of <a> or the omission of the vowel, with no alternative representations present in the data (a pattern that we have seen hold for /a/ in all positions because of the lack of any real alternative using the Roman script as based on StF and StE). The overall ratio of 91:9 in favour of <a> over <ø> (omission) that we see in Table 9.26 below is complicated, however, by the different ranges of omission that we see across the 19 words (with a total of 30 total tokens showing word-medial /a/ per individual participant):

Table 9.26 – Word-Medial /a/ – *Dataset 2*

	<a>	<ø>	<ø>%	
/hana/	48	0	*0%*	*"Felicity"*
/dal/	49	0	*0%*	*"Stay* [v., imp.]"
/xal.li:ni/	48	0	*0%*	*"Let me, allow me"*
/sale:me/	49	0	*0%*	*"Health, well-being"*
/ħaje:ti/	47	1	*2%*	*"[My] life"*
/ħada/	190	5	*3%*	*"Someone"*
/baħər/	95	3	*3%*	*"Sea"*
/maʕi/	44	2	*4%*	*"With me"*
/ʕajle/	46	2	*4%*	*"Family"*
/ɣaj.jru/	47	2	*4%*	*"Other than him/it* [m.]"
/t'aj.jəb/	46	3	*6%*	*"Alright, so"*
/hal/	101	6	*6%*	*"This"*
/maʕkon/	45	4	*8%*	*"With you* [pl.]"
/ʃaɣle/	87	8	*8%*	*"Thing* [f.]*"*
/maʕ/	84	11	***12%***	*"With"*
/ħabi:bi	42	6	***13%***	*"My darling"*
/bas/	38	6	***14%***	*"But, only"*
/ʕal/	129	30	***19%***	*"On the"*
/ʕam/	133	46	***26%***	***"****I am [doing]"*
Total	**1,368**	**135**		
	91%	9%		

Omission for medial /a/ varies between 0% and 8% for the majority of words (941 tokens), with only the final five words in Table 9.26 (the remaining 426 tokens) showing between 12% and 26% omission. Moreover, we note that the final three words (highlighted: /bas/, /ʕal/ and /ʕam/, showing highest omission, at 14%, 19% and 26%) were among the common marker words we used in 7.3.1, being single-vowel, morphologically simple and semantically useful words for which we proposed a conventionalised use of high omission as a result of a lowered pressure to specify the vowel, given their high frequency and familiarity. That we find the highest omission of Dataset 2 in these same words confirms our approach (which we further follow up on shortly in 9.3.3). We also find further evidence that higher familiarity leads to higher omission where the /a/ of /ħabi:bi/ shows a similar omission rate to the common marker words (at 13%), which we attribute to the same high-frequency effect (being the same effect we discussed in the context of the use of <h> for this word's

/ħ/ in 9.2.2 IV). Finally we also recall (7.3.1, Figure 7.13D) that the form /hal/, despite its morphological similarity to our common high-omission words and its common meaning ("*this*"), nevertheless showed much lower omission despite having the highest token count in Dataset 1. Our conclusion that this word is anomalous (most likely due to the semantic ambiguity of its omitted form <hl>) is also vindicated by Dataset 2, for which /hal/ again appears with low omission at 6%, exactly the same rate it showed in Dataset 1:

Table 9.27 – "*This, that*" /hal/ – *Both Datasets*

/ha**l**/

	<a>	**<ø>**	***<ø>%***
Dataset 1	159	10	***6%***
Dataset 2	101	6	***6%***
	94%	6%	

Apart from these common forms, our analysis of vowel omission for lower frequency, lower familiarity words in 7.3.1 was limited by a lack of sufficient repeated tokens of such words in Dataset 1. Now, using the highly-repeated words afforded by Dataset 2, we can determine that the omission rate of /a/ is considerably lower in the non-common forms (that constitute the 941 tokens that see 0%-8% omission), even in the context of the time-pressure and smartphone-based data of Dataset 2, and is indeed closer to the omission of other short vowels, including /o/ (for which we found 4% omission in 9.3.2 II prior).

V. Schwa

Table 9.28A – Schwa – *Dataset 2*

	<e>	<i>	<ø>	<ø>%	
/rəd.dəl.li/	37	5	1	*2%*	*"Answer me, get back to me"*
/əħki/	42	6	1	*2%*	*"Speak* [v., imp.]"
/jrəd/	88	2	8	*8%*	*"[He] replies"*
/a:ʕəd/	42	1	6	*12%*	*"[He] is sitting"*
/rəd.dəl.li/	28	10	5	*12%*	*"Answer me, get back to me"*
/t'aj.jəb/	40	2	6	*13%*	*"Alright, so"*
/ʃəfli/	33	5	7	*16%*	*"Check, see for me"*
/rəħna/	39	2	8	*16%*	*"[We] went"*
/təħki/	38	3	8	*16%*	*"[You] talk"*
/mən.non/	31	3	13	*28%*	*"From them"*
/bənʔoz/	30	3	15	*31%*	*"I get a fright"*
/kəl/	88	6	50	***35%***	***"****Every"*
/bəl/	44	7	48	***48%***	*"In the"*
/wəl/	18	1	18	***49%***	*"And the"*
/mən/	28	6	63	***65%***	***"****From"*
Total	**626**	**62**	**257**		
	66%	**7%**	**27%**		

Table 9.28B – Schwa, Excluding Common Words – *Dataset 2*

<e>	<i>	<ø>	<ø>%
448	**42**	**78**	***14%***
79%	**7%**	**14%**	

We see for schwa too a wide range of omission, though with an overall higher frequency of omission across the board compared to word-medial /a/, consistent with our conclusions in in 7.3.1 that schwa sees higher omission overall. The highest-omission forms for the schwa range from 35% to 65%, and again show the same common marker words that showed the highest omission in Dataset 1. Discounting these forms (as in Table 9.28B), general omission for schwa appears at a rate of 14%. Additionally, we note that while in the case of non-omission, the most popular resolution is <e> (with a total of 66% overall frequency), there is nevertheless also a third variant in the form of <i> that appears as a minority form at 6% total frequency. We now move on to reconsider the new data we have for these same forms in Dataset 2 (alongside that of the /a/ forms) in our re-examination of vowel omission as a whole.

9.3.3 Vowel Omission Revisited

I. Vowel Omission in High-Frequency Forms

Having reviewed using Dataset 2 the primary positions within which we expect vowel omission and found that the forms with the highest omission in Dataset 2 are some of the same common forms we examined in Dataset 1, we now compare overall word omission frequencies between Dataset 1 and Dataset 2.

Table 9.29 – Schwa Omission– *Both Datasets*

Dataset 1

	<e>	<i>	<ø>	<ø>%	Tokens
/wəl/	20	8	11	***28%***	*39*
/kəl/	46	8	21	***28%***	*75*
/bəl/	56	16	45	***38%***	*117*
/mən/	83	47	79	***38%***	*209*
Total	**205**	79	**156**		
	47%	18%	**35%**		

Dataset 2

	<e>	<i>	<ø>	<ø>%	Tokens
/wəl/	18	1	18	***49%***	*37*
/kəl/	88	6	50	***35%***	*144*
/bəl/	44	7	48	***48%***	*99*
/mən/	28	6	63	***65%***	*97*
Total	**178**	20	**179**		
	47%	5%	**48%**		

Comparing the same four common schwa forms across both datasets, we see a clear rise in omission rates in Dataset 2, in keeping with our expectations considering the nature of the experimental interviews of the dataset (implicit time-pressure, use of smartphone and simulation of an informal texting context), leading the overall omission for schwa vowels to rise by 14% from **35%** in Dataset 1 to **48%** in Dataset 2. It is also notable that the frequency of <e> as a representation for schwa remains identical at 47% in both datasets, with instead fewer instances of <i> making up the higher omission in Dataset 2, though without further investigation it is not entirely possible to determine the (unlikely) possibility that only those producing <i> switch to omission, not least because the participants in Dataset 2 are entirely different individuals to those whose writing we examined in Dataset 1. In terms of high-frequency forms showing higher omission, we note first of all that the token-count for Dataset 2 is meaningless in this regard, being a representation of pre-determined sentences prepared for the interview; therefore we consider instead the token count for these words in Dataset 1, reflecting a more natural spread of frequencies for these words. In this regard, we see that while the same scale is retained in Dataset 2 for /kəl/, /bəl/ and /mən/ (which also scale upwards in omission rates in Dataset 2 in accordance with their frequencies in Dataset 1), the form /wəl/ sees much higher omission in Dataset 2 compared to the other

Dataset 2 words, despite having the lowest token count in Dataset 1 (and having a lower omission rate there of 28%). Here we are forced to conclude that /wəl/ is (for whatever combination of factors) more commonly abbreviated among the 49 participants of Dataset 2 (and potentially more commonly used by them in their overall LQA CMCR writing), hence its higher rate of omission here, though the fact that it appears only once per interview session also means that it has the lowest tokens overall for Dataset 2, indicative not of the actual frequency of use among individuals, but instead potentially indicating a statistical effect for why its omission rate does not synchronise with the rest of the forms in Dataset 2. It is also possible that the omitted form <wl> has become more conventional in the intervening period between Dataset 1 and Dataset 2.

Table 9.30 – Short-Vowel Omission– *Both Datasets*

	Dataset 1				Dataset 2				
	<a>	**<ø>**	***<ø>%***	***Tokens***		**<a>**	**<ø>**	***<ø>%***	***Tokens***
/ʕal/	58	12	***17%***	70	/ʕal/	129	30	***19%***	159
/ʕam/	100	27	***21%***	127	/ʕam/	133	46	***26%***	179
/bas/	95	39	***29%***	134	/bas/	38	6	***14%***	44
Total	**253**	**78**			**Total**	**300**	**82**		
	76%	**24%**				79%	**21%**		

In the case of short-vowels, we compare the frequencies of the three common words that appear in both Dataset 2 and Dataset 1 (omitting other forms from the common short vowel forms of Dataset 1, including anomalous /hal/). We see here, however, a drop of 3% in the overall omission rate. Unlike the schwa forms, the omission of which in Dataset 2 largely followed the same scale of frequency predicted by the token-counts of Dataset 1, short vowel omission does not follow the same pattern. The word /ʕal/, which had the lowest token count (70) and lowest omission rate (17%) in Dataset 1 shows a similar omission rate of 19%, and /ʕam/ which had a high 127 tokens in Dataset 1 and an omission rate of 21% now shows the highest omission rate in Dataset 2 at 26%. In this case we identify /bas/ as the anomalous form in Dataset 2, which despite having shown the highest tokens and highest omission (29%) in Dataset 1 (134), now shows only 14% omission in Dataset 2. We cannot conclude that vowel omission in Dataset 2 scales in fact with the token count of Dataset 2 rather than Dataset 1, given the insignificance of the token counts of Dataset 2.

This data is better understood in the context of the low token count specifically for the word /bas/ (at 44 tokens) leading to an anomalous rate of 14% omission most likely by chance. In this way, we understand the individual omission rates for /ʕal/ rising from 17% to 19% from Dataset 1 to Dataset 2, and for /ʕam/ rising from 21% in Dataset 1 to 26% in Dataset 2, thus demonstrating the same effect we saw for schwa and which we predicted due to the nature of the data of Dataset 2, whereby omission for short vowels also rises in this new dataset proportionally to both the short vowel data of Dataset 1, as well as to the rise in omission rates of schwa between the datasets. We therefore propose that, absent any other factor we have not successfully identified, /bas/ should show the same effect given enough tokens.

Table 9.31 – Long-Vowel Omission– *Both Datasets*

Dataset 1

	<V>	<ø>	<ø>%
/he:k/	70	6	*8%*
/kti:r/	70	3	*4%*
Total	**140**	**9**	
	94%	**6%**	

Dataset 2

	<V>	<ø>	<ø>%
/he:k/	46	3	*6%*
/ke:n/	41	2	*5%*
/axba:r/	48	1	*2%*
Total	**135**	**6**	
	96%	**4%**	

Finally, as in Dataset 1, we find again see only rare cases of omission for long vowels, with a total of only 6 tokens in Dataset 2 showing long vowel omission, compared to the 9 tokens in Dataset 1. As such, we conclude that even in the case of a context that encourages higher omission in schwa and short vowels, long-vowel omission frequency is largely unaffected by the nature by which individuals produce orthographical forms. The 2% drop from Dataset 1 to Dataset 2 is likely anomalous, considering the very low number of tokens for omitted low vowels in both datasets, rather than indicating less tendency to omit long vowels in the experimental context of Dataset 2. As such, the tiered view we took of omission in 7.3.1 is retained here: schwa omission (as well as being most common) is most variable depending on context, with short vowel omission, in addition to being second-most common, rises by a lesser degree depending on context, and long-vowel omission, which is rarest, is largely non-variable by context.

II. Vowel Omission in Non-Frequent Forms

Table 9.32 – Vowel Omission per Phoneme *–Dataset 2*

	<V>	<ø>
/o/	96%	4%
/a/	95%	5%
/ə/	86%	14%
/ɪ/	78%	22%

Finally, we look to omission outside of the common forms examined in 7.3.1 and 9.3.3 I, noting that while vowel omission generally occurs across the board, it is far less prevalent outside of the common forms, appearing at low rates of 4% and 5% for short vowels /o/ and /a/, and at 14% for schwa (based on Table 9.28B for schwa and the respective tables for the other short vowels). Based on Dataset 2, outside of the common forms it is most commonly found as a representation of /ɪ/, where it is omitted at a rate of 22%, largely equivalent to its omission in the case of the /a/ of high-frequency common words. In this way we develop our understanding of vowel omission to be a generalised feature throughout the four vowel sounds of Table 9.32 above, in addition to being conventionally used at much higher frequencies in the case of the common forms, where high familiarity encourages higher omission and, in the resulting loss of phonetic detail, less transcriptional writing.

9.3.4 Long Vowels

I. Long Vowel /a:/

Table 9.33 – Long-Vowel /a:/ *–Dataset 2*

	<a>	<e>	<ø>	<o>	
/**a:**ʔəd/	48	1			*"He is sitting"*
/axb**a:**r/	48		1		*"News"*
/ħ**a:**li/	48			1	*"Myself"*
/ħ**a:**lak/	49				*"Yourself"*
Total /a:/	**193**	**1**	**1**	**1**	
	100%	0%	0%	0%	

We see very little variability in the representation of long vowel /a:/, with only one-off tokens (one of <e>, one of <o> and one showing omission), while all 193 other tokens show <a>. Like short /a/, there is no real variation in the writing of long vowel /a:/, though also no

means of distinguishing vowel length. The Tripolitan LQA pronunciation of long vowel /a:/ tends towards /o:/, particularly noticeably in Old Tripolitan pronunciation, though this is only marked once in all of Dataset 2 with an <o>, and that was in the case of a participant who chose to ironically exaggerate both their speech and writing in a characteristically Tripolitan manner. That no others write <o> instead of <a>, however does not mean that the spoken vowel is not realised in some cases closer to /o:/ than to /a:/, but that it is either not being perceived as such by the individuals producing the sound, or else that they choose to avoid the non-prestigious representation of the Old Tripolitan dialect in writing; we explore this further in 10.3.4.

II. Long Vowel /i:/

Table 9.34 – Long-Vowel /i:/ *–Dataset 2*

	<i>	**<e>**	**<ee>**	
/fi:ni/	46	1	1	*"I can"*
/fi:/	79	1	0	*"Is, is there"*
/ʃi:/	145	1	0	*"Something"*
/xal.li:ni/	45	2	1	*"Let me, allow me"*
/mni:ħa/	47	0	1	*"Good [f.]"*
/ħabi:bi/	38	2	0	*"My darling"*
Total /e:/	**400**	**7**	**3**	
	98%	1%	1%	

Long vowel /i:/ is similarly straight-forward, with <i> as its only real representation, and <e> and <ee> appearing only with minor tokens. We also again note that /i:/ is not distinguished from its short form /i/.

III. Long Vowel /o:/

Table 9.35 – Long-Vowel /o:/ *–Dataset 2*

	<o>	**<ou>**	**<u>**	
/əljo:m/	62	12	10	*"Today"*
	74%	14%	12%	

The long vowel /o:/ in LQA (aside from where /a:/ is realised as /o:/ in Old Tripolitan, as discussed above) primarily derives from SA diphthong /aw/ (such as SA <خَوف> /xawf/

turning to LQA /xo:f/, *"fear"*). We find one example of this sound in Dataset 2 in the word /əlj**o:**m/ (deriving from SA /alj**aw**m/, both meaning "*today*"). Though the majority of individuals represent this sound with grapheme <o> (again not distinguishing long /o:/ from short /o/), a minority opt for forms <ou> or <u>, likely for etymological rather than phonetic reasons, given that the SA orthographical form <اليوم> features the character <و> which plays the role of both the diphthong /aw/ (in conjunction with a preceding diacritic marking the sound /a/), but which also (without preceding diacritics) represents long /u:/. A minority of users of LQA CMCR therefore retain the etymological conflation of /o:/ and /u:/ by writing <ou> or <u> instead of <o>, (as we also saw for the short form /o/ in 9.3.2 II). In frequency-convergence terms, <o> remains the most popular graphemic resolution.

IV. Long Vowel /e:/

Table 9.36 – Long-Vowel /e:/ *–Dataset 2*

Trip. LQA	**<e>**	**<ei>**	**<ø>**	**<a>**	**<ay>**	***SA***	
/x**e:**r/	100	32			3	/x**aj**r/	*"Good, goodness"*
/ɣ**e:**r/	35	8			1	/ɣ**aj**r/	*"Other, different"*
/b**e:**t/	41	8				/b**aj**t/	*"House"*
/ʔl**e:**k/	41	4	3		1	/ʔal**aj**k/	*"On you [m.]"*
/mb**e:**rɪħ/	128	3	5	7		/al b**a:**riħa/	*"Yesterday"*
/əj.j**e:**m/	42	2	1	4		/aj.j**a:**m/	*"Days"*
/bal**e:**ha/	42	3		4		/bal**a:**ha/	*"Without it/her [f.]"*
/sal**e:**me/	39		2	8		/sal**a:**ma/	*"Health, well-being"*
/h**e:**k/	39	4	3	1	2	/h**a:**kaða/	*"Like this, thus"*
/ħaj**e:**ti/	39	2	4	3		/ħaj**a:**ti/	*"My life"*
/s**e:**kɪt/	38	2		8		/s**a:**kit/	*"[He is] quiet"*
/mn**e:**ħ/	39	3		5		/ml**a:**ħ/	*"They are well"*
/w**e:**ħɪd/	74	3	5	13		/w**a:**ħad/	*"One, someone [m.]"*
Total /e:/	**697**	**74**	**23**	**53**	**7**		
	82%	**9%**	3%	**6%**	0%		

In the case of long vowel /e:/, we see a clear majority for the form <e> (at 82% of all variation). However, we also see a secondary form in <ei> at 9%, and a tertiary form in <a> at 6%. The form <ei> is notable in that it marks long /e:/ apart from short /e/ (unlike other vowels for which length has no means of being distinguished). This <ei> form appears most prominently in the words where the long /e:/ sound in Tripolitan LQA derives from the

assimilation of the SA diphthong /aj/ (as in SA /bajt/ becoming Tripolitan LQA /be:t/). The final column in Table 9.36 marks the etymological SA root of /e:/ in each word, where we see that 52 of the total 74 <ei> tokens (70%) appear where the /e:/ is etymologically derived from SA /aj/ (for which we also find 5 tokens of <ay>). While some dialects of LQA retain the /aj/ diphthong in speech, this is generally never the case in Tripolitan LQA, meaning that these are not likely to be transcriptional representations. Table 9.37 below shows only words where the /e:/ derives from SA /aj/, where we find <ei> has a frequency of 19%.

Table 9.37 – Long-Vowel /e:/ derived from SA /aj/ – *Dataset 2*

Trip. LQA	**<e>**	**<ei>**	**<ø>**	**<a>**	**<ay>**	***SA***
/xe:r/	100	32			3	**/aj/**
/ɣe:r/	35	8			1	**/aj/**
/be:t/	41	8				**/aj/**
/ʔle:k/	41	4	3		1	**/aj/**
Total /e:/	**217**	**52**	**3**	**0**	**5**	
	78%	**19%**	1%	0%	2%	

More interesting still is the possibility that <ei> is becoming generalised in the repertoire of Tripolitan users of LQA CMCR, given that it appears with 19 tokens even for representations of /e:/ that derive instead from SA /a:/ (the second major SA source for LQA /e:/, as SA /aj/ and /a:/ converge to /e:/ in spoken Tripolitan LQA). Even if the origins of the convention <ei> are etymological in nature, it is entirely feasible to see <ei> emerging as a convention for distinguishing the long vowel /e:/ in all its manifestations, though its low overall rate of use at 9% means that it cannot be considered a major variant.

Table 9.38 – Long-Vowel /e:/ derived from SA /a:/ – *Dataset 2*

Trip. LQA	<e>	<ei>	<ø>	<a>	<ay>	*SA*
/mbe:rɪħ/	128	3	5	7		/a:/
/əj.je:m/	42	2	1	4		/a:/
/bale:ha/	42	3		4		/a:/
/sale:me/	39		2	8		/a:/
/he:k/	39	4	3	1	2	/a:/
/ħaje:ti/	39	2	4	3		/a:/
/se:kɪt/	38	2		8		/a:/
/mne:ħ/	39	3		5		/a:/
/we:ħɪd/	74	3	5	13		/a:/
Total /e:/	**480**	**22**	**20**	**53**	**2**	
	84%	**4%**	**3%**	**9%**	**0%**	

Table 9.38 above shows words where /e:/ derives instead from SA /a:/, in which <ei> has a low frequency of 4%, in tandem with an expected rise in the use of <a>, up to 9% (from the 6% in the first, combined Table 9.36). This too is likely to be etymological rather than phonetic, given that /a:/ pronunciations such as /sala:me/ are unlikely in spoken Tripolitan LQA, which we confirm using our recordings in 10.3.1. Unlike <ei>, the use of <a> is not generalised, but only appears for the words in which the long /e:/ derives from SA /a:/; Table 9.37 showing /aj/-derived words shows exactly zero <a> tokens for /e:/ (which in turn reinforces the meaningfulness of the generalisation of <ei> irrespective of derivation as a potentially emergent convention for distinguishing /e/ and /e:/- though if so, still a fledgling one). In summary, there is a degree of etymological retention of SA written conventions even in the Roman script of LQA CMCR writing, though only to a limited extent (given the overall 9% and 6% for the etymological features in the combined Table 9.36). Moreover, in the case of <ei> there exists potential for an etymological form to become a useful convention for representing /e:/ instead of /e/. The vast majority (82%) of representations of long /e:/ appear as <e>, mirroring /o:/ and /a:/ in the non-delineation of vowel length, and ultimately, representing spoken LQA forms rather than etymological SA written forms. Conventionalisation, therefore, can occur not only as a result of the resolution of phonetic transcriptional variation but also of etymologically derived variants, which can either become conventionalised as primary representations, or else are replaced by non-etymologically derived conventions.

9.3.5 Conclusions

In the course of this chapter, we have not only reviewed much of our analysis in Chapters 7 and 8 in light of the new written data from Dataset 2, but also used this same data to develop a fuller understanding of the frequencies of phonemic-graphemic resolutions among users of LQA CMCR, allowing us to understand (and predict) the convergence (or non-convergence) of words via high-frequency usage on a partially statistical basis in conjunction with the understanding we have developed for the unique factors affecting these choices and informing these ratios and in many cases altering them. While such an approach therefore provides a starting point for understanding conventionalisation, it cannot alone be used to represent the entirety of the structure of LQA CMCR as used in Tripoli, nor to represent the entirety of the conventionalisation that takes place within it. We will use utilise these frequencies- and our fuller understanding of conventionalisation- in Chapter 11, where we use this model to predict convergence in various words, with various degrees of success. First, however, we examine one final feature: the link between spoken LQA and written LQA CMCR, and thus our final research question (RQ5), which we now turn to in Chapter 10.

Chapter 10: Writing & Speech

10.1 Dataset 2: Spoken Data

10.1.1 A Novel Analytical Approach

We have thus far understood conventionalisation as the tension between transcriptional and conventional writing, whereby specific graphemic forms are used more frequently in the transcription of uncommon or less familiar words that their writers assume to be less immediately clear, whereas less specific, more ambiguous graphemic forms are resorted to in the context of common and familiar forms for which there is an expectation of more immediate clarity. This is an underlying framework for the function of non-standard orthographies more generally, as discussed at length in Chapter 4. Our practical understanding of transcriptionality has thus far, however, been largely limited to considering the choice of variant graphemic representations of the same phonetic realisations, as we have not had the capability to discern differences in spoken forms. As such, we have understood transcriptional writing to be the use of clearer graphemic resolutions such as <7> and conventional writing to be the use of more ambiguous (but also more orthographically conventional) resolutions such as <h>- even while these graphemes both represent the same sound /ħ/. Now we are able to take a new approach in which we consider the fuller extent of transcriptional writing including the graphemic representation of alternative spoken realisations. In this way, we understand transcriptional writing to be the most transparent graphemic representation of the individual's phonetic repertoire, whereas here conventional writing is the use of graphemic forms that do not directly represent the spoken realisation of the same form for the individual in question. We thus combine our frequency-based understanding of transcription and convention as per RQ1 with a supplementary understanding thereof on the basis of RQ5 and the *divorce* between speech and writing (Jaffe, 502; Kress, 18; see 4.3.2), which is typical of standard orthographies but which nevertheless takes place to a limited extent within the writing of non-standard orthographies too, and for which we have posited a scale between fully transcriptional and fully conventional (see again 4.3.2). In this way, grassroots conventions afford users of Type 2/NSR orthographies the choice to not write fully transcriptionally. Prestige also plays a central role in the divorce between speech and writing, again not only within SLC but so too in our non-standard context (which we also saw in Elhij'a, 2012, and

her description of the use of a conventionalised <2> by Palestinian QA speakers to mark the urban form they did not themselves produce in speech; see 5.3.3). We anticipate other factors to also play a role, such as the etymological factor we examined in the previous chapter for the writing of the phoneme /e:/, variously written as <ei> or <a> depending on its SA etymon (see 9.3.4 IV), and we are now in a position to confirm whether it is the orthographic or phonetic SA form that motivates such choices using our recorded data (which we do in 10.3.1). We are now also able to examine the extent to which spoken LQA elements are indeed represented in writing, including the emphatic consonants (discussed in 6.2.5 II and 9.2.5) and Tripolitan speech elements (discussed in 6.1.2 and 9.3.4 I). Thus we can split graphemic and phonetic variation into the following three categories:

- Words that differ **graphemically** *and* **phonetically**
  (i.e. <badi> or <bedi> graphemic variation, and /bad.di/ or /bəd.di/ phonetic variation)

- Words that differ **graphemically** *but not* phonetically
  (i.e., the graphemic forms <7ayati> and <7ayeti>, both pronounced as /ħaje:ti/).

- Words that differ **phonetically** *but not* graphemically
  (i.e. <dal> has variable pronunciations /dal/ or /dˁal/, but is generally only written as <dal>).

Variation on an exclusively graphemic basis consists of various representations for the same sound, while variation on a phonetically exclusive basis consists of phonetic variation that is not represented graphemically in the writing of LQA CMCR. Where variation can take place in both graphemic and phonetic realisations, what we have understood to be *transcriptional writing* is when this variation occurs both graphemically and phonetically in tandem. For words that differ phonetically particularly across the wider spectrum of LQA (including the prestige dialect of Beiruti LQA), cases where both the graphemic and phonetic Beiruti LQA form is utilised by Tripolitan LQA speakers is again an example of transcriptional writing; where, however, the prestige forms informs only the orthographical realisation of Tripolitan LQ CMCR users without impacting their local pronunciation of the word, we are able to identify prestige-based conventional writing instead. We begin our analysis in 10.2 by examining these prestige forms, before going on to analyse the other phonetic variation that has come up in our work so far in 10.3.

10.1.2 Methodology for Spoken Data

Before beginning our analysis, we briefly describe the two exercises by which the spoken data of Dataset 2 was gathered.

I. Transliteration Exercise

In this exercise the participant is presented with vernacular sentences written in the classical Arabic-script LQA CMCA orthography. They are asked to repeat each sentence out loud, and then asked to write it out again in the Roman script orthography of LQA CMCR. This is done on a phone to make the conditions as similar as possible to the usual environment in which this orthography is usually employed. In this way we are able to compare individuals' own pronunciations with the way they choose to render words in writing. The primary aim of this exercise is to determine to which degree individuals are reflecting their own speech-patterns in their text-writing.

II. Reproduction Exercise

In this exercise, the participant is presented orally with a sentence in the general LQA of Tripoli, then asked to write it as if they were texting it (in the same conditions as in the above exercise). Here, the sentence is not given to them using the LQA CMCA script, but my own pronunciation with which I read it aloud, which in some cases might influence the way in which the participants render the sentences. This can then be compared with the results of the first exercise to determine the degree of individuality present within individuals' renderings of words and how it might be affected by external sources. Ultimately, in cases where there is a marked difference between results in this exercise and the above, there would then be evidence of a conscious rendering of words based on how they sound, rather than an internalised writing system uninfluenced by the oral quality of the words being written.

10.2 Prestige Forms: Between Beiruti & Tripolitan LQA

10.2.1 <Bedi> vs. <Badi>

The word "*want*" appears twice in Dataset 2, in the first person variably as <*bedi> or <*badi> ("*I want"*), and in the second person masculine variably as <*bedak> or <*badak>

("*you want*"). The Beiruti LQA pronunciation is generally realised with the first vowel as /a/ (thus /bad.di/ and /bad.dak/), but in Tripolitan LQA the initial vowel is schwa (/bəd.di/ and /bəd.dak/). Orthographically, the majority of tokens for "*I want*" (which appears in the first part of Dataset 2, produced by transliteration of Arabic-script LQA CMCA sentences) show <*badi> at 32 tokens, with only 4 showing the <*bedi> form that indicates the schwa pronunciation of Tripolitan LQA. Examining the actual pronunciation made by each individual for each token, however, paints a very different picture: as we see in Table 10.1 below, of the 32 <*badi> tokens, a small majority of 17 were realised with the Tripolitan LQA /bəd.di/ pronunciation, while the remaining 15 were phonetically consistent with the graphemic form and were pronounced /bad.di/. Thus while a number of participants did indeed use the Beiruti LQA form in their speech as well as their writing, more retained the Tripolitan LQA form in their speech even while writing Beiruti form <*badi>. Of the few who did write <*bedi>, three were consistent in their pronunciation of it as /bəd.di/, though curiously one participant wrote Tripolitan <bedi> but said Beiruti /bad.di/ in an anomalous reversal of the overall pattern.

Table 10.1 - Spoken vs. Written Tokens – *"I want"*
Part 1 – ***Prompt****:* <بدي> [B-D-Y]

Written	Tokens	Spoken	Tokens
<*badi>[23]	32	/bəd.di/	17
		/bad.di/	15
<*bedi>	4	/bəd.di/	3
		/bad.di/	1

In addition to these total 36 tokens, we see in Table 10.2 below that vowel-omitted form <*bdi> appears with 11 tokens. We expect the majority of these to represent schwa rather than <a>, based on our frequency rates calculated in the previous chapter (9.3.2 V) combined with the overall phonetic popularity of form /bəd.d.i/, and we find that this is indeed the case, as 8 of the vowel-omitted tokens represent /bəd.di/, while the remaining 3 were accompanied by a /bad.di/ pronunciation.

[23] We use again the same convention we have utilised throughout, whereby an asterisk marks a generalised form simplified to focus on the variation we are interested in. In this case <*badi> consists of variants such as <badi> and <baddi> (but not <bdi> and <bddi>, which belong to the <*bdi> group); a specific token with that precise spelling is indicated where there is no asterisk.

Table 10.2 – Spoken vs. Written Tokens (with Vowels Omitted) – *"I want"*
Part 1

Written	Tokens	Spoken	Tokens
<*bdi>	11	/bəd.di/	8
		/bad.di/	3

Rearranging the data by phonetic form, we see in Table 10.3 below that Tripolitan pronunciation /bəd.di/ has the highest total spoken tokens at 28, but a majority of orthographical representations appear as <*badi> (with a minority of <*bdi> and still smaller minority of <*bedi> written tokens); on the other hand, /bad.di/ is the less popular pronunciation at 19 tokens, and retains the <*badi> orthographical form for its representation, with only four alternative tokens (3 vowel omission, one anomalous written token of <*bedi>).

Table 10.3 – Variant Spellings of Binary Phonetic Forms – *"I want"*
Part 1

(*Tripolitan*) /bəd.di/

<badi>	17	61%
<bdi>	8	29%
<bedi>	3	11%
Total	**28**	

(*Beiruti*) /bad.di/

<badi>	15	79%
<bdi>	3	16%
<bedi>	1	5%
Total	**19**	

This data is fully consistent with our hypothesis that the writing of this word is becoming conventionalised on the basis of the Beiruti LQA form, which to a certain degree affects the pronunciation of a proportion of our Tripolitan participants, but which more visibly impacts not their phonetic but orthographic realisation of the word, thus indicating a clear prestige effect of the Beiruti LQA form and the emergence of a convention for the writing of this word that is no longer transcriptionally reflecting individuals' spoken realisation, but is rather realised with a conventionalised spelling as <*badi> irrespective of how it is being pronounced. This is reinforced by what appears to be only a minor change in orthographical representation depending how the word is realised in speech, given that both sub-tables in Table 10.3 see a similar spread of orthographical forms, where the only indication of the different pronunciations is the rise in omission (by 13%) and slight rise in <*bedi> forms for

/bəd.di/ and the subsequent fall in popularity of these alternate forms in the /bad.di/ table but within which there is still a similar spread of percentages overall. We now further investigate the effect of phonetic realisation on orthographical representation using the second form: <*bedak> and <*badak>.

10.2.2 <Bedak> vs. <Badak>

The second person masculine form of the same word (*"you want"*) shows the same manner of variation, appearing as <*bedak> or <*badak> orthographically and /bəd.dak/ or /bad.dak/ phonetically. While the <*bedi>/<*badi> form appeared in Part 1 of the interviews (where it was read out from Arabic-script CMCA by participants, then typed out in Roman-script CMCR by them), the <*bedak>/<*badak> form appears in Part 2, where participants heard the word in my pronunciation (using Tripolitan LQA /bəd.dak/), and were then asked to reproduce it in CMCR writing, without recording their own spoken realisation. While this means that we cannot compare the tokens of <*bedak>/<*badak> to the speech of those who produced them, we can observe whether the total number of orthographical representations of /ə/ rises as a result of my /bəd.dak/ pronunciation compared to the totals of <*bedi>/<*badi> in 10.2.1.

Table 10.4 – Part 1 vs. Part 2 – *"I want"* / *"You want"*

Part 1			**Part 2**		
Prompt: <بدي> [B-D-Y]			***Prompt:*** /bəd.dak/		
<*badi>	32	**68%**	**<*badak>**	27	**57%**
<*bedi>	4	**9%**	**<*bedak>**	12	**26%**
<*bdi>	11	**23%**	**<*bdak>**	8	**17%**

We see in Table 10.4 a clear effect resulting from my /bəd.di/ pronunciation that participants were prompted with in Part 2 of the interviews: the <a> forms drop by 11%, and even schwa forms fall by 6%, all in favour of <e> forms that rise by a full 17% as a result of the clear Tripolitan LQA prompt, versus the indeterminate LQA CMCA prompt in Part 1 which did not indicate either LQA spoken variant, appearing as <بدي> (B-D-Y) with the initial vowel unmarked. Nevertheless, it is perhaps still more telling that the majority did not alter their use of <a> irrespective of the form in which they were presented with the word, consistent with the fact that the majority of <*badi> forms in Part 1 were coupled with

/bəd.di/ pronunciations by the same individuals. We recall briefly here by analogy our discussion of StE and StF-derived features, where although we did find a clear harmonic effect for a minority of users, we were forced to determine that overall, this harmonic link is no longer meaningful as a means of determining the orthographical composition of LQA CMCR (see the discussion in 7.4). Here, too, we find a similar effect, where some individuals alter their orthographical production on the transcriptional basis of the link to the phonetic form in a clearly observable manner, but the fact that the majority retain the <a> form is stronger indication still that the majority of participants are in fact writing these words conventionally and not transcriptionally. We take this one final step further by examining the specific participants who changed from writing <a> in Part 1 to <e> or omission in Part 2 as observed in Table 10.4 above.

Table 10.5 – Orthographic Forms from Part 1 to Part 2 – *"I want" / "You want"*

#	P1	P1 pron.	P2
5	<bdi>	/bəd.di/	<bedak>
7	<baddi>	/bəd.di/	<beddak>
9	<bdi>	/bəd.di/	<bedak>
10	<baddi>	/bəd.di/	<biddak>
16	<badi>	/bəd.di/	<bedak>
21	<bdi>	/bəd.di/	<bedek>
31	<badi>	/bad.di/	<bedak>
36	<badi>	/bəd.di/	<bedak>
39	<badi>	/bəd.di/	<bedak>
43	<badi>	/bəd.di/	<bdak>
47	<bdi>	/bad.di/	<badk>

In Table 10.5 above we see which orthographic and phonetic forms each participant produced in Part 1, as well as which orthographic form they produced in Part 2 in response to my spoken prompt. Some of the changes observed take place between omitted and un-omitted forms: #5, #9 and #21 (all of whom used the Tripolitan pronunciation) produced omitted forms in Part 1, but wrote <bedak> or <bedek> in Part 2. For the omitted form <bdak> produced by #43 in Part 2, we cannot discount that the omitted vowel was intended to be <a>, given that they produced <badi> in Part 1, and conversely for #47, we cannot discount that their <bdi> in Part 1 was not omitting <a> rather than <e> considering both their production of <badk> in Part 2 and their own /bad.di/ pronunciation. The remaining six

participants (#7, #10, #16, #31, #36 and #39, highlighted with light yellow in Table 10.5) clearly change from an <a> form to an <e> form between Parts 1 and 2, as likely influenced by my pronunciation, the majority of whom themselves also produced the phonetic form /bəd.di/ in Part 1, with the only exception being a single participant #31, who changed orthographic forms from <badi> to <bedak>, while actually pronouncing the word /bad.di/ in Part 1. In the case of this participant, we see an example of an individual being directly influenced by my schwa pronunciation and changing their spelling to match it, despite their own pronunciation being different, whereas the rest, though they wrote conventional form <badi> while themselves saying /bəd.di/, changed their writing to match the clear production of /bəd.di/ as spoken by another person (in this case myself).

10.2.3 <We7ed> vs. <Wa7ad>

Just as with written forms <*badi> and <*bedi> and spoken forms /bad.di/ and /bəd.di/, the word meaning "*one, someone*" can be pronounced variably as /wa:ħad/ and /we:ħɪd/, and generally sees variable orthographical forms <*wa7ad> or <*we7ed> as a result. Here, /we:ħɪd/ is the Tripolitan LQA form while /wa:ħad/ is the Beiruti LQA form of the capital, and thus can again be hypothesised to have a prestige bearing in both its pronunciation and in its orthographical realisation as <*wa7ad>.

Table 10.6 – Variant Spellings of Binary Phonetic Forms – *"One, someone"*
Part 1 – ***Prompt***: <واحد> [W-A-Ħ-D]

(*Tripolitan*) /we:ħɪd/			(*Beiruti*) /wa:ħad/		
<*we7ed>	37	**77%**	<*we7ed>	0	0%
<*wa7ed>	8	17%	<*wa7ed>	0	0%
<*w7d>	3	6%	<*w7d>	0	0%
<*wa7ad>	0	0%	<*wa7ad>	1	**100%**

We find in Part 1, however, only a single instance where the Beiruti LQA pronunciation of /wa:ħad/ is realised phonetically, which is also represented transcriptionally with the only spelling of <*wa7ad> in all of Part 1. The remaining 48 spoken tokens appear as Tripolitan LQA /we:ħɪd/, indicating that the phonetic influence of the Beiruti form here is significantly less prevalent among the LQA speakers of Tripoli both phonetically and orthographically. The same prestige-based conventional effect we saw for <*bedi>/<*badi> is not replicated

here, but rather that transcriptional writing remains in place, with the orthographic forms generally representing the spoken forms directly. The only apparent exceptions to this are the 8 total tokens of <wahed>, <wahd>, <wahd> and <wa7ed>, which appear to indicate a middle-ground pronunciation of /wa:ħɪd/, but which is much more likely the result of the purely orthographic effect we discussed in 9.3.4 IV, where long vowel /e:/ is variably represented with an <a> at an overall frequency of about 6%, up to 9% in the case where the /e:/ derives from SA /a:/, as is the case with /we:ħɪd/. Unlike the case of <*bedi> and <*bedak> in Dataset 2, and indeed unlike <*wa7ad>/<*we7ed> itself in Dataset 1 where Beiruti form <*wa7ad> accounted for 20 out of a total 33 tokens (but for which we have no phonetic data; see 8.2.1 II), in the case of Dataset 2 the variation for this word is largely shared between transcriptional and etymological, with no visible prestige effect, though on account of Dataset 1, it is feasible that one does exist. We further examine the etymological effect with regards to the writing of /e:/ as <a> further in 10.3.1 to follow shortly.

Table 10.7 – Orthographic Forms from Part 1 to Part 2 – *"One, someone"*

Part 1

Prompt: <واحد> [W-A-ħ-D]

Written	Spoken	Count	
<*we7ed>	/we:ħɪd /	37	**76%**
<*wa7ed>	/we:ħɪd/	8	16%
<*w7d>	/we:ħɪd/	3	6%
<*wa7ad>	/wa:ħad/	1	2%

Part 2

Prompt: /we:ħɪd/

Written	Count	
<*we7ed>	40	**85%**
<*wa7ed>	4	9%
<*wa7d>	2	4%
<*wa7ad>	1	2%

This second appearance of this word in Dataset 2 is in Part 2 of the interviews, where it is read orally to participants using my Tripolitan LQA /we:ħɪd/ pronunciation. The single person who produced <*wa7ad> orthographically and /wa:ħad/ phonetically is the same person to produce <*wa7ad> again in Part 2, where it remains the only instance of this form. Moreover, the rise of the percentage of forms favouring <*we7ed> from 76% to 85% is likely a direct reflection of my clear vocal prompt of /we:ħɪd/ replacing the orthographical prompt of Part 1 (which appeared as <واحد> in the Arabic script, indicating [W-A-ħ-D] and therefore likely inducing an etymological effect), leading to the fall in both <*wa7ed> and <*wa7d>. This manner of variability again indicates the transcriptional construction of this word, rather than the availability of a conventionalised manner of writing.

10.3 Phonetic Variability in Tripolitan LQA

10.3.1 Vowel Variation: Phonetic or Etymological?

I. Vowel /e:/ with Grapheme <a>

In addition to our discussion of the grapheme <a> in the context of <wa7ed> just prior, we examined in 9.3.4 IV the overall tendency for users of LQA CMCR to render the long vowel /e:/ using <a> in 6-9% of instances, which we posited to be an etymological effect of SA orthography rather than a representation of any true phonetic variation in the realisation of this sound, bearing in mind that the change from SA /a:/ to LQA /e:/ is one of the prime markers of the Lebanese QA dialect, setting it apart from neighbouring dialects such as Syrian, Jordanian and Palestinian QA. We now use the instances of these words that appear in Part 1 to confirm the hypothesis that the spelling of /e:/ as <a> is an example of orthographical and not phonetic variation, which Table 10.8 emphatically demonstrates, wherein we find not a single pronunciation of /e:/ as /a:/ in our data; even those who write <a> produce the phonetic form /e:/, confirming our hypothesis.

Table 10.8 – Orthographic vs. Spoken Tokens - Vowel /e:/ & Grapheme <a>
Part 1

	/e:/	/a:/	
<*seket>	40	0	*"[He] Is quiet"*
<*saket>	8	0	
<*mne7>	42	0	*"[They] are good"*
<*mna7>	5	0	
<*saleme>	39	0	*"Health, wellbeing"*
<*salame>	8	0	

Total	/e:/	/a:/
<e>	121	0
<a>	21	0

II. Vowel /e:/ with Grapheme <ei>

The same is true for the realisations of /e:/ as <ei>, which we also hypothesised to be orthographical, this time based on the etymological root of /aj/ in SA and its corresponding orthographical form <ي> ([Y], with a previous diacritic marking /a/ to produce /aj/; though we also saw that this <ei> form might be becoming generalised as a conventional manner of

distinguishing /e:/ from /e/; see 9.3.4 IV). Examining (in Table 10.9 below) the two cases of /ɣe:r/ and one of /xe:r/ which appear in Part 1 of our interviews, we clearly see that, irrespective of the LQA CMCR orthographical form produced, all participants produce /e:/ and not a realisation like SA /aj/ from which the LQA sound derives, thus confirming again the etymological nature of this as orthographical (and not phonetic) variation.

Table 10.9 – Orthographic vs. Spoken Tokens - Vowel /e:/ & Grapheme <ei>
Part 1

	/e:/	/aj:/	
<*kher>	69[24]	0	
<*kheir>	20	0	*"Good, goodness"*
<*khayr>	3	0	
<*gher>	35	0	
<*gheir>	8	0	*"Other, different"*
<*ghayr>	1	0	

Total	/e:/	/aj:/
<e>	104	0
<ei> / <ay>	32	0

10.3.2 The Emphatic Consonants

There are two words containing emphatic consonants in Dataset 2, the first being /tˤaj.jəb/ (*"alright, so"*, and its alternate form /tˤab/), while the second is one of two pronunciations of "*stay, remain*", either as emphatic /dˤal/ or non-emphatic /dal/.

Table 10.10 – Emphatic /tˤ/ - Written and Spoken Tokens – *"Alright, so"*
Part 1 – ***Prompt:*** <طيب> [Tˤ-Y-B]

Written	Tokens	Spoken	Tokens
<*tayeb>	48	/tˤaj.jəb/	48
		/taj.jəb/	0

For the word *"alright, so"*, the spread of orthographical representations appears more or less precisely as predicted by our Lexemic-Aggregational frequency principles (see 11.1.6), with no indication of any kind of representation for /tˤ/ apart from that of /t/. Examining the

[24] These are not the same totals we find in Table 9.36 in section 9.3.4 IV because we are only able to examine the 2 tokens of <kher> per individual that appear in Part 1 alongside a phonetic realisation; the final token per individual appears in Part 2, and thus has no accompanying spoken data.

spoken tokens, we find that every phonetic realisation of this word is emphatic (with zero cases of /taj.jəb/, as might occur in prestige capital dialect Beiruti LQA).

Table 10.11 – Emphatic /dˤ/ - Written and Spoken Tokens – *"Stay, remain"*
Part 1 – ***Prompt:*** <ضل> [Dˤ-L]

Written	Tokens	Spoken	Tokens	
<dal>	40	/dˤal/	30	**75%**
		/dal	10	25%
<dall>	9	/dˤal/	7	**78%**
		/dal	2	22%

The word "*stay, remain*" can be realised in Tripolitan LQA either with emphatic /dˤ/ as /dˤal/ or non-emphatic /d/ as /dal/, with no difference in meaning. The most popular form is <dal> at 40 tokens, and though we might have hypothesised <dall> to be a possible representation of the emphatic form (given the pharyngealisation effect of the emphatic consonants on the sounds that follow), we find (in Table 10.11) a largely even spread, with both <dal> and <dall> realised as /dˤal/ at frequencies between 75-78%, and as /dal/ between 22-25%. Though the emphatic form appears to be the majority phonetic realisation, it is important to note that the written CMCA prompt participants received showed the emphatic form delineated (as <ض>) in the Arabic script, likely affecting many of the phonetic realisations. Even so, that the spread of phonetic forms between <dal> and <dall> is largely the same means that this orthographical couplet is not used for distinguishing the emphatic and non-emphatic sound in this word, and so ultimately here, too, there is no means of distinguishing emphatic /dˤ/ from non-emphatic /d/ in LQA CMCR writing.

Table 10.12 – /kˤ/-retention vs. /kˤ/ to /ʔ/ - "*Coffee, café*"
Part 1 – ***Prompt:*** <قهوه> [Kˤ-H-W-H]

<*ahweh>	/ahweh/	47
<kahweh>	/ahweh/	1
<kawheh>	/kˤawheh/	1

A common feature of LQA is the dropping of the emphatic /kˤ/ of SA to glottal stop /ʔ/ in LQA. Dataset 2 has two words in which this sound occurs: "*[he] is sitting*" (LQA /a:ʕəd/, SA /kˤa:ʕid/) and *"coffee, café"* (LQA /ahwe/, SA /kˤahwa/). No written tokens for *"[he] is*

sitting" show any indication of the consonant /kˤ/, and all spoken tokens consist of a word-initial /ʔa:/. In the case of *"coffee, café",* however, we see in Table 10.12 above two written tokens of <kahweh> in addition to the 47 of <*ahweh> (and variants), one of which is accompanied by a pronunciation of /kˤahwe/ that retains the word-initial emphatic. The single <kahweh> token pronounced as /ahweh/ is likely another one-off instance of etymological writing, retaining in LQA CMCR the representation of <ق> ([Kˤ]). The other <kahweh> token, for which the same individual also produced a /kˤahweh/ pronunciation is highly unusual for Tripolitan LQA (this emphatic /kˤ/ is more typically retained only in the Druze LQA dialects of Mount Lebanon, culturally and geographically disconnected from Tripolitan LQA), and in this case the use of <k> is not etymological but transcriptional, reflecting the individual's pronunciation. In the case of both <kahweh> tokens however, etymological and transcriptional, no measure is taken to represent the specific emphatic form /kˤ/, but instead the same <k> that also represents non-emphatic /k/ in general LQA CMCR usage is utilised.

10.3.3 Written & Spoken Variability in Alternative Couplets

We finish this section by examining the further phonetic variation in Dataset 2 in comparison with the orthographic representations that accompany it, specifically in the case of words with alternative pronunciations, where this analysis allows us to determine what written variation is entirely orthographical, and what written variation is transcriptionally representative of an equivalent phonetic variation, which we do by examining four sets of couplets consisting of alternative forms for the same words.

I. /fi:ni/ and /fij.ji/

Table 10.13 – Orthographic Realisations vs. Spoken Tokens – *"I can"*
Part 1 – ***Prompt****:* <فيني> [F-Y-**N**-Y]

Written	Tokens	Spoken	Tokens
<*fini>	45	/fi:ni/	**45**
		/fij.ji/	0
<fiyi>	1	/fi:ni/	**1**
		/fij.ji/	0
<feye>	1	/fi:ni/	0
		/fij.ji/	**1**
<fiyi feeni>	1	/fi:ni/	0
		/fij.ji/	**1**

The word "*I can*" is most frequently realised in Tripolitan LQA as /fi:ni/, but has an alternate form pronounced /fij.ji/ with the same meaning, more popular in Beiruti LQA and likely derived therefrom (where it occurs closer to /fij.je/). The majority of instances (a total of 45) show the Tripolitan LQA form in both writing (as <fini> or <fine>) and with the vocalisation /fi:ni/. There are three outliers, the first being one written token of <fiyi> which is nevertheless pronounced as Tripolitan LQA /fi:ni/, recalling our conclusions for <*bedi>/<*badi>, though as this is a single token we cannot draw further conclusions. The other written token appears as <feye> and is coupled with a Beiruti pronunciation of /fij.ji/, thus being a transcriptional reflection of spoken speech in writing. Finally, we have the curious token of <fiyi feeni>, whereby the participant produced both orthographical forms, along with the spoken form /fij.ji/. We recall that the interview process consisted of the participant first reading a line of Arabic-script LQA CMCA (in which this word was written as <فيني>, reflecting pronunciation /fi:ni/), and thereafter wrote the same sentence out using LQA CMCR. In this case, this participant read CMCA form <فيني> ([F-Y-N-Y]) as /fij.ji/, presumably in their own personal dialect, but when it came to producing an orthographical token, after replicating this personal dialect transcriptionally as <fiyi>, then re-wrote it as <feeni>. We might assume they had intended to delete the first token, but the implicit pressure of the process led to them failing to do so, thus providing us with an interesting insight into their potential thought process, as well as giving us clear indication of the interplay between these two forms in the minds of speakers of LQA.

II. /leːʃ/ and /leː/

Table 10.14A – Orthographic Realisations vs. Spoken Tokens – *"Why"*

Part 1 – ***Prompt****:* <ليش> [L-Y-**SH**]

	/leːʃ/	/leː/
<*l**esh**>	**44**	**0**
<*l**eh**>	**5**	**0**

Table 10.14B – Orthographic Realisations – *"Why"*

Part 2 – ***Prompt****:* /le**ːʃ**/

<*l**esh**>	**38**
<*l**eh**>	**8**

The word "*why*" appears variably as either /leːʃ/ or /leː/ in Tripolitan LQA. In Part 1 of the interviews, the /leːʃ/ form was prompted using CMCA (<ليش> [L-Y-SH]), which was reflected in a majority of tokens (44) as <*lesh> and variants (including <leish>, <leich>, and others), as well as showing 5 tokens of <*leh>. Every phonetic token appeared as /leːʃ/, with the /ʃ/ clearly pronounced, as we see in Table 10.14A. In Part 2, where participants were prompted with my own pronunciation (/leːʃ/), 38 tokens of <*lesh> were produced, and 8 tokens of <*leh>, as we see in table 10.14B. Of these <*leh> tokens, only two were produced by the same participants who also produced <*leh> in Part 1 (participants 12 and 13), indicating that aside from these two cases where <leh> might be the preferred form, there is an overall general variability between the <*lesh> and <*leh> forms, noting in particular also the possibility of mistyped forms, given that both <lesh> and <lech> are one letter away from <leh> and in the implicit pressure of the interviews (replicating the general nature of synchronous CMC), the possibility of a missed <s> or <c> leading to <leh> cannot be discounted.

III. /ħada/ and /ħadan/

Table 10.15A – Orthographic Realisations vs. Spoken Tokens – *"Someone"*
Part 1 – ***Prompt****:* <حدا> [Ħ-D-**A**]

Written	Tokens	%	Spoken	Tokens
<*7ada>	46	**94%**	/ħad**a**/	45
			/ħad**an**/	1
<*7adan>	3	**6%**	/ħad**a**/	3
			/ħad**an**/	0

Table 10.15B – Orthographic Realisations – *"Someone"*
Part 2 – ***Prompt****:* /ħad**a**/

	Tokens	%
<*7ad**a**>	134	**92%**
<*7ad**an**>	12	**8%**

The word for "*someone*, *somebody*" is most commonly realised as /ħada/ in Tripolitan LQA, though an alternate form /ħadan/ exists, deriving from the grammatical marking of the word in SA in the context of the case of the word that follows, though in the case of Tripolitan LQA it is used irrespective of grammatical marking and has become a standalone variant of the word, generally perceived to be an older local form which does not generally appear in Beiruti LQA. We find however no clear prestige effect in the way this word appears in our Dataset 2, particularly given the incongruity with which orthographical and phonetic variants interplay as we see in Table 10.15A, where one token of <*7ada> couples with a /ħadan/ pronunciation, and the three <*7adan> written tokens are realised as /ħada/. Thus the single phonetic realisation of /ħadan/ that appears in our data is not specified as such, though a prestige effect, if it were to be significant, should mean that <*7adan> is almost never realised, particularly when the pronunciation itself is indicating the conventional form /ħada/. There is a slightly higher ratio of <*7adan> forms in Part 2 (where the prompt was spoken form /ħada/), but the rise is a minor 2%, and the fact that most participants were producing the /ħada/ spoken form in Part 1 indicates that the spoken realisation of this word has little impact on its orthography, with the <*7ada> form generally more prevalent.

IV. /mbe:rɪħ/ and /mbe:rħa/

Table 10.16 – Orthographic Realisations vs. Spoken Tokens – *"Yesterday"*
Part 1 – ***Prompt:*** <مبارح> [M-B-A-**R-Ħ**]

Written	Tokens	%	Spoken	Tokens
<*mbere7>	93	95%	/mbe:**rɪħ**/	91
			/mbe:**rħa**/	2
<*mber7a>	5	5%	/mbe:**rɪħ**/	1
			/mbe:**rħa**/	4

Finally, the word for "*yesterday*" is usually realised as /mbe:rɪħ/, but an alternative form /mbe:rħa/ also exists in Tripolitan LQA. As we see in Table 10.16, the majority of written and spoken forms indicate <*mbere7> with pronunciation /mbe:rɪħ/, though two tokens of <*mbere7> are in fact realised phonetically as the less common phonetic form /mbe:rħa/, and one token of orthographic form <*mber7a> is realised phonetically as the more common spoken form /mbe:rɪħ/. The remaining four spoken tokens of /mbe:rħa/ are also orthographically marked as <*mber7a>, showing a transcriptional orthographical adjustment in line with the phonetic variation.

10.3.4 Old Tripolitan Speech Elements

Table 10.17 – Full Text Produced by Participant 21 – *Translations in Appendix*

Part 1

1	Tayeb bas 5alini chuf iza fini eji ma3koN 3al **ba7ar**
2	Chu bdi dal 3am e7ki m3 **7olo** ? Lch ma **7adan** 3m yred 3layii
3	Kl we7ed mnkon seket choi hal cha8le haydi **ya**
4	5er **inchala** ? Mbere7 re7na 3al ahwe ma ken fi chi
5	8er hk chu l **a5bar** ? **Inchala** l yom a7san mn mbere7 ? Wl 3ayle mne7?
6	Bel hana 7abibi nchufak b 5et w salame **nchala**

Part 2

7	Bedek tji 3al **ba7ar** wala la chefli w reedeli **5abar**
8	3m t7ki ma3i wala ma 3a **7alak** lch ma **7ada** 3m yred 3lk
9	Kl ma bchuf we7ed mnon bn2oz
10	**Echbak** mbere7 bt2li chi wl yom chi 8ayroo
11	Eh mni7 a7san mn balehaa 5er **inchala**
12	Ana b hal iyem **a3ed** bl bet kl 7ayeti ma bchuf 7ada wala 7ada bichufni

In Table 10.17 above, we see the full interview text produced by Participant 21 (male, aged 18-21). For Part 1, we can also consult the accompanying recordings produced by the same

participant, while Part 2 was produced in response to my own phonetic reading of the sentences. Participant 21 is of such particular interest here because they took it upon themselves- upon hearing that this was a study of the Tripolitan dialect of LQA- to produce the most emphatically and stereotypically Tripolitan LQA they could muster, both in the spoken recordings as well as the written CMCR extracts they provided. For the sake of our study, of equal if not greater interest are the points where this individual- striving for the most Tripolitan participation possible- failed to mark their Tripolitan spoken tokens in their writing. Highlighted in red are the orthographic tokens they produced with deliberately Tripolitan transcription, whereas in blue we highlight the written tokens this participant produced using more conventional LQA CMCR orthographic forms (closer to what we have called New Tripolitan LQA, rather than Old Tripolitan LQA; see 6.1.2), despite the Old Tripolitan LQA form being possible to represent transcriptionally in writing, such as for example their token of <a5bar>, for which a spelling of <a5b**o**r> would better reflect their Old Tripolitan vocalisation of /axb**o:**r/ (instead of generalised LQA /axb**a:**r/, consistent with SA and general LQA /a:/ to Old Tripolitan LQA /o:/ as discussed in 9.3.4 I).

Table 10.18 – Written vs. Spoken Tokens - Performative Old Tripolitan LQA
***Part 1** – Participant 21*

Writing	*Pronunciation*	
Token	**Indicated**	**Actual**
<ba7ar>	/baħar/	/baħar/
<7olo>	/ ħo:li/	/ ħo:li/
<7adan>	/ħadan/	/ħada/
<inchala>	/ənʃa:la/	/ənʃo:la/
<a5bar>	/axba:r/	/axbo:r/
<Inchala>	/ənʃa:la/	/ənʃo:lo/
<nchala>	/ənʃa:la/	/ənʃo:lo/

Using the recording of Part 1, we determine precisely how these orthographical forms were vocalised by Participant 21, and find that with the exception of <7adan> being produced as the more common /ħada/, all other words with an obvious Old Tripolitan orthographical realisation that Participant 21 did not use were phonetically realised in their Old Tripolitan form. Just as the token <a5bar> does not fully reflect the participant's /axb**o:**r/ pronunciation, all three instances of /ənʃ**o:**lo/ produced vocally appear orthographically

variously as <inchala> and <nchala>, with the /**o:**/ again unmarked in writing. Even in the unique case of an individual attempting to produce both spoken and written tokens of Old Tripolitan we find a great deal of spoken variation unmarked in writing, despite an effort to mark Old Tripolitan forms such as <ba7ar> (Old Tripolitan /baħ**a**r/ instead of LQA /baħ**ə**r/ for "*sea"*) and <7olo> (intended as <7oli> and thus Old Tripolitan /ħ**o:**li/ instead of LQA /ħ**a:**li/ for *"myself"*). Additionally, the tone of voice Participant 21 takes has an aggressive edge, including minor alterations to the sentences such as orthographic and phonetic flourishes like <ya> (/ja:/) at the end of sentence 3, a Tripolitan LQA grammatical marker signalling (in this case playful) confrontation, in keeping with the rest of the excessive performance of not only the accent itself but the attitude associated with it, recalling our discussion from 6.1.2 of the perception of Old Tripolitan LQA as a macho, aggressive and thus masculine-prestigious register, in this case largely played for laughs.

All things considered, however, the degree to which the Old Tripolitan register was successfully marked by this participant in their orthographical forms is ultimately sparse, in comparison certainly to their vocal performance, reflecting our hypothesis that speakers of Old Tripolitan LQA tend to underestimate their Old Tripolitan spoken forms in writing, most likely because of the conventional use of generalised LQA orthographic forms. Though we do not have recorded tokens for the forms in Part 2, it is still telling that <ba7ar> ("*sea*") and <echbak> were still produced despite my prompting with a clear /baħər/ for the former, and /ʃəbak/ for the latter, where the participant produced <echbak> indicating an alternate form /əʃbak/ (*"what's wrong with you?"*), again adding an aggressive tone as well as likely being perceived as archaic. This form does not appear elsewhere in our data but, is presumably associated by Participant 21 with stereotypical Old Tripolitan LQA, despite it being in fact generally more associated with rural dialects and even Beiruti LQA, yet it was utilised within Participant 21's performative interview for its perception as an unusual form (and in it potentially not being an especially Old Tripolitan form, recalls another discussion from 6.1.2 whereby the younger generation learns Old Tripolitan for prestige reasons via third parties, rather than from their parents and immediate surroundings). Overall, my New Tripolitan LQA prompts appear to have led to fewer tokens of Old Tripolitan LQA CMCR in Part 2, where even the token <7**a**lak> (indicating New Tripolitan /ħ**a:**lak, "*yourself"*) is produced in a New Tripolitan orthographical form despite the appearance in Part 1 of the Old Tripolitan

form <7olo> (indicating Old Tripolitan /ħ**o:**li/, "*myself*"). Here <7ada> also reverts to the more conventional <n>-less form, while <5abar>, <inchala> and <a3ed> are not written in the possible transcriptional Old Tripolitan forms of <*5ab**o**r> (/xabor/), <*inch**o**l**o**> (/ən**ʃo:**lo) and <**2o**3ed> (/**o:**ʕəd/). Here we can also recall Hinrichs' (2004) findings of how some JC speakers use a limited number of JC words to indicate a JC reading of a passage without needing to delineate each JC pronunciation; in this way, we can also interpret the sparse number of Old Tripolitan LQA tokens in the passages of Participant 21 as a means of indicating a wider Old Tripolitan pronunciation for the entirety of the passage.

Table 10.19 – Old Tripolitan LQA Phonetic Forms – Other Participants

Part 1

Participant #	*Gender*	/ənʃo:lo/	/ħo:li/	/axbo:r/
1	*Male*	I		
8	*Male*	I	I	
11	*Female*	I	I	
13	*Female*	I		
16	*Male*	I		
18	*Male*	I		I
19	*Male*		I	
27	*Male*		I	I
29	*Female*		I	I
30	*Female*	I	I	
35	*Male*		I	
40	*Female*		I	
47	*Female*		I	

While Participant 21's forms (both spoken and written) were intentionally performative, elements of Old Tripolitan LQA also appear in the speech of Tripolitan LQA speakers more generally. Using the recorded data, we observe (in Table 10.19 above) that other than Participant 21, a total of thirteen participants produced Old Tripolitan LQA phonetic forms as /ənʃo:lo/ (*"God-willing",* general LQA /ənʃa:la/), as /ħo:li/ (*"myself",* general LQA /ħa:li/) or as /axbo:r/ (*"news",* general LQA /axba:r/), or some combination of these three. We firstly note that this effect falls outside the gendered masculine prestige of Old Tripolitan LQA; here, there is an almost even gender split for those who use Old Tripolitan /o:/ in their speech (six female, seven male), indicative of the fact that this is very different to what

Participant 21 engaged in: it is not performative, but in fact an element that these individuals are unlikely to be aware of within their speech, hence why we find in our written data not a single orthographic representation of the phonetic variation observed in Table 10.19. In this way we observe another limitation of transcriptional writing, whereby the choice of representing or not representing Old Tripolitan forms is not a question of prestige (whether a desire to avoid stigma, or to perform masculinity), but in this case because the phonetic variation that takes place here is not recognised by those producing it. This is in addition to the fact that, despite Participant 21's deliberate attempt to transcriptionally represent the strong Old Tripolitan spoken LQA he utilised vocally, much of the spoken variation was not successfully represented orthographically. While we certainly have found a great deal of transcriptionality in the various orthographical forms that vary in harmony with spoken variation throughout this analysis, not every phonetic quality of spoken Tripolitan LQA is accurately transcribed even in cases where the prestige factor is overturned by a motivation to produce fully transcriptional Tripolitan LQA writing, indicating at least a minor degree of conventional writing underpinning the entire non-standard orthography of Tripolitan LQA based on a more generalised LQA. This manifests in the writing of <a> even for individuals making pronunciations closer to /axbo:r/ than /axba:r/, and thus this limitation of further potential variability is evident in our analysis in 9.3.4 I, where the long vowel /a:/ appeared in 193 tokens with the grapheme <a> across all of Dataset 2, and only once with <o>- with that singular token deriving precisely from the <7olo> (intended as <7oli>) of Participant 21.

10.4 Conclusions

We have observed transcriptional writing in a number of instances through the close alignment of graphemic and phonetic variation, demonstrating a conscious awareness- among a certain number of LQA speakers at least- of the production of graphemic forms that mirror the spoken forms these speakers produce. Conversely, conventional writing takes place where there is a divorce between phonetic realisation and orthographic representation, within which we observe a loss of transcriptionality. This can occur in the form of etymological realisations such as <a> and <ei> for /e:/, where latter form <ei> could potentially develop into a conventional representation of long vowel /e:/ as apart from

short vowel /e/, and in this way even etymologically-based non-transcriptionality might lead to the emergence of conventionalised forms, though presently the form <ei> remains a minority representation (as seen in 9.3.4 IV). It is, however, through the conventional writing of specific words (in particular the form <*badi>) that we observe through orthographic-phonetic variation the clearest instance of conventionalised usage, where the prestige of the Beiruti LQA form and its CMCR representation leads to widespread use of the orthographical form <*badi> even while the majority of those producing this form realise the same word phonetically as /bəd.di/. It is here that we find the clearest example of orthographic-phonetic divergence, and thus most emphatically answer our fifth research question:

5. (How) can we observe conventionalisation on a basis of phoneme-grapheme divergence?

Transcriptional writing occurs in words that differ orthographically and phonetically in the same way, whereas where phonetic and orthographic divergence occurs separately but systematically (such as the use of prestige or etymological forms), we see the emergence of conventional and conventionalised forms. Again, it is through the same tension between transcriptional and conventional writing that we are able to understand variation and conventionalisation within our non-standard LQA CMCR orthography. Words can also vary only phonetically without this variation being reflected orthographically, such as in the emphatic consonants which users of LQA CMCR generally do not distinguish in writing, as well as in the non-representation of Old Tripolitan LQA forms, even when they occur clearly in speech, indicating a degree of conventionalised limitation to the use of LQA CMCR whereby the majority of forms are anchored to their generalised LQA orthographical forms whereby transcriptional re-writing on the basis of these particular, low-prestige realisations (which are likely to often not be consciously discerned as different by those producing them) does not occur, and thus defines a certain degree of boundary to possible variation within the writing of LQA CMCR.

Chapter 11: Conventionalisation in Tripolitan LQA CMCR

11.1 The Make-Up of LQA CMCR: Research Questions

We have examined the non-standard writing of LQA CMCR of Tripoli in detail, focusing in particular on the most variable features and the contexts and reasons for their variation. We addressed the research questions that we formed on the basis of previous work in the field of grassroots conventionalisation, and as such understood not only the variation that exists, but also its potential re-arrangement via the emergence of conventional forms. In this way we have conducted a novel examination of conventionalisation within a Type 2/NSR orthography that is not directly tethered to a single standard form, but instead built on the basis of a number of standard orthographies, which users of LQA CMCR draw upon for the resources that form their LQA CMCR orthographic repertoire. We conclude our study by first summarising the resolutions to our research questions in 11.1, before moving on in 11.2 to contextualising our findings within the field of conventionalisation and standardisation that we explored in depth in the first chapters of this thesis.

11.1.1 - RQ1: The Frequency Convergence Effect

1. **(How) does high-frequency usage of specific words lead to conventionalised spellings?**

We have replicated to some degree the findings of Deuber and Hinrichs (2007) whereby high-frequency words see convergence upon conventional forms, and further demonstrated how in a Type 2/NSR context, these convergent forms are based on the highest-popularity graphemic resolutions for each phonemic position. Using this principle, we have developed a broader model for predicting high-frequency convergence in the form of our Lexemic-Aggregational methodology, the usefulness of which we finally evaluate in 11.1.6 to follow. We have also observed other, more specific effects that occur in the case of high-frequency words, primarily predicated on high-frequency meaning high familiarity and therefore a reduced need among users of the orthography to specify the form being produced. We saw this in the tendency towards the use of ambiguous <h> over specific <7> in the cases of

words that appear with high frequency (8.2.1), where the need to distinguish /ħ/ from /h/ is reduced. This same effect applies for vowel omission, where we saw (in 7.3.1 I and 9.3.3 I) that high-frequency forms are consistently likely to see higher frequencies of omission, again as a result of the reduced ambiguity as compared with omission in less common words, where specification is preferred at a higher frequency. We understand the writing of high-frequency forms to be more conventional, indicated with less specific graphemes in contrast to the specifically phonetic and therefore transcriptional writing of less frequent forms.

11.1.2 - RQ2: The Semantic Overlap Effect

2. (How) does the need to maintain semantic clarity affect conventionalisation?

Semantic clarity can have a modifying effect on both examples of the word-frequency effect discussed above. The potential ambiguity of the vowel-omitted form <hl> leads to it being the single anomalous form that sees high-frequency use and yet a substantially lower rate of omission than predicted by other forms, an effect we observed to be remarkably consistent in both Dataset 1 and Dataset 2 (9.3.3 I). The motivation for semantic clarity can, in some cases, also overturn the expected frequency-familiarity effect for the voiceless pharyngeal fricative, while in other cases the frequency-effect can overturn the semantic clarity effect (8.2.3). Semantic clarity is therefore understood within the expectations of users of LQA CMCR, and how clear they perceive the form they are producing to be for its intended readers (whether this expectation is accurate or otherwise). While users tend to write high-familiarity words (as in RQ1) with lower specification and higher ambiguity, this can be counteracted in cases where the use of more ambiguous forms leads to multiple valid semantic readings of the same orthographical construction, such as where an orthographic form using <h> has equally valid but different semantic realisations depending on whether <h> is read as /h/ or /ħ/. Semantic clarity was considered an important factor in the orthographical choices of users of Jamaican Creole (Hinrichs, 2004) but not Nigerian Pidgin (Deuber and Hinrichs, 2007); we find that it is certainly a factor for users of LQA CMCR, and moreover, in our case, functions not where ambiguity is created between non-standard forms and the equivalent standard forms of their lexifier (StE, as in the case of Hinrichs' JC

users), but rather, in the Type 2/NSR context, occurs internally as a result of certain graphemic combinations, or in the case of graphemes like <h> to which ambiguity is introduced when these graphemes are used to represent more than a single LQA phoneme.

11.1.3 - RQ3: The Effect of French & English Orthographies

3. (How) does conventionalisation take place on the basis of the sound-symbol correspondences of the standard English and French orthographies?

Bilingualism plays a central role in the emergence of new orthographies, including the grassroots emergence of non-standard orthographies (Sebba, 2007; see 4.2). The non-standard writing of LQA CMCR is rooted in the two Roman script standard orthographies of StE and StF that are most familiar to speakers of LQA (6.1), and which provide the majority of the sound-symbol correspondences used in the writing of LQA CMCR. We explored in Chapter 7 the possibility of sub-dividing the writing of LQA CMCR into two sub-orthographies, each premised on one of the two Roman script orthographies that inform LQA CMCR, but concluded that this is not a viable approach. Whilst there does exist for a certain number of users a harmonic relationship between the use of StF or StE-rooted features, resources form both StE and StF writing have become part of the overall repertoire of LQA CMCR, and are used largely interchangeably. Conventionalisation, instead, takes place within the admixture of these features, the frequencies of which we observed to vary between Dataset 1 (in Chapter 7) and Dataset 2 (in Chapter 9). We understand these conventions as variable features in a dynamic, changing non-standard orthography without any clear resolution outside of the frequency-preferences within which conventionalisation is to be understood, in contrast to the codified and largely unchangeable standard writing of SLC (4.3.3). Unlike examples like Haiti, where French orthographical features have come to index a native Haitian identity that is threatened by the incursion of orthographical forms based instead on standard English (2.3.3), the admixture of conventions deriving from both StE and StF reflects the weaker social effect of either orthography for users of Tripolitan LQA CMCR, both of which are regarded as foreign, though we might discern a generational effect whereby members of the newer generation are more likely (though not certain) to prefer

StE-based rather than StF-based forms on the basis of the growing prevalence of StE and the fading relevance of StF in the Lebanese and particularly Tripolitan context (see 6.1.2).

11.1.4 - RQ4: The Effect of standard Arabic writing

4. (How) does standard Arabic writing affect the writing of LQA CMCR?

Despite being written in the Roman script, the repertoire of LQA CMCR contains a number of features deriving directly from the standard Arabic-script orthography of SA. Chief among these is vowel omission (7.3.1 and 9.3.3), deriving from the unwritten short vowels of SA writing, though we have established that vowel omission in LQA CMCR does not mirror the unwritten vowels of SA writing directly, but rather has been adopted as a generalised convention, for which individual tendencies have developed among users of LQA CMCR based not on codified rules, but instead conventionalised frequencies. Additionally, we have identified a number of graphemic choices as etymological, such as the writing of /e:/ with an <a> that reflects the <ا> of the Arabic script (as well as the SA pronunciation /a:/). The same LQA sound /e:/, in cases where it derives instead from SA /aj/ is often written as <ei> in LQA CMCR, for which we have observed a potentially emergent convention through the generalisation of <ei> to all /e:/ sounds, even those deriving from SA /a:/ and <ا>. Though <ei> does not appear with great frequency, it nevertheless provides a useful way to differentiate long /e:/ from short /e/ and provides a clear-cut example of a new LQA CMCR convention deriving from the SA writing despite the change in script. This coincidentally echoes a very similar effect observed by Deuber and Hinrichs (2007) in the case of Nigerian Pidgin users writing <ey> for the sound /e/, itself derived from standard English <ey> as in words such as *they* (which /ej/ sound turns to /e/ for speakers of NP; see 5.2.2 II). We ultimately understand conventionalisation as the resolution not only of phonetic variation, but also of etymologically-derived variation either by increasingly conventional use of such variants or else a loss of popularity in favour of non-etymologically derived conventions (even while these remain accessible as resources for users of LQA CMCR).

11.1.5 - RQ5: The Interplay Between Writing and Speech

5. (How) can we observe conventionalisation on a basis of phoneme-grapheme divergence?

Using the voice recordings of Dataset 2, we conducted a unique investigation in Chapter 10 of the distinction between spoken and written language in a manner not done by any of the previous studies in the field of grassroots conventionalisation. In this way we identified the prestige effect of Beiruti LQA on both the speech but especially the writing of Tripolitan speakers of LQA, and observed emerging written conventions in LQA CMCR through the conventional use of the prestige Beiruti orthographic forms such as <badi>, even in cases where the local pronunciation /bəd.di/ is retained. In this divergence between phonetic and orthographical realisation we observe a concurrent shift from transcriptional to conventional writing as a result of sociolinguistic pressures, where the written form begins to represent less of the phonetic detail of the spoken form. We also used our spoken data to ascertain the etymological (rather than phonetic) nature of spellings such as <ei> and <a> for /e:/, as well as determining through variant phonetic forms that there exists in LQA CMCR both heavily transcriptional writing (where spoken and written forms diverge in unison) and less transcriptional (and thus more conventional) writing where written and spoken variation do not occur together.

11.1.6 – Lexemic-Aggregational Analysis

We now finally examine both the usefulness and limitations of our Lexemic-Aggregational model by using the overall frequencies developed in Chapter 9, which we summarise in our own phonetic-graphemic table below (Table 11.1). Unlike the tables of Yaghan (2008) and Abu Elhij'a (2012, 2014; see 5.3.3 for both), our table is unique first in not aiming to represent singular graphemic resolutions per phoneme in the mould of SLC expectations of standard invariance, and secondly in representing a highly specific local spoken variant and its localised CMCR writing, rather than attempting to generalise across an entire national QA variant (as Abu Elhij'a) or indeed across all of Roman script writing of QA online (as Yaghan). We now apply the frequencies summarised in the table below to three different words from Dataset 2 in order to compare the predicted convergent form with the actual tokens that appear for these words. This allows us to observe the extent to which this approach can successfully predict convergence, as well as the importance of the other factors delineated through our research questions, and finally also other effects that limit the ability to predict convergence through a contextless frequency methodology alone.

Table 11.1 – Grapheme Frequency per Phoneme in Tripolitan LQA CMCR

A. Vowels

Grapheme / Frequency		
/a/ Word Initial	**<a>**	**<2a>**
Dataset 2	**88%**	12%

/a/ Word-Medial	**<a>**	**<ø>**
Dataset 2	**95%**	**5%**

/a/ Word-Final	**<a>**
Dataset 2	**100%**

/a:/	**<a>**	*Other*
Dataset 2	**99%**	**1%**

/ə/ Low-Freq.	**<e>**	**<ø>**	**<i>**
Dataset 2	**79%**	14%	7%

/ə/ High-Freq.	**<e>**	**<ø>**	**<i>**
Dataset 1	**47%**	35%	18%
Dataset 2	47%	**48%**	5%

/e/ Word Final	**<e>**	**<eh>**
Dataset 2	**63%**	37%

/e:/	**<e>**	**<ei>**	**<a>**	*Other*
Dataset 2	**82%**	9%	6%	3%

/i/ Word-Initial	**<i>**	**<e>**
Dataset 2	**59%**	41%

/i/ Word-Medial	**<e>**	**<i>**
Dataset 2	**63%**	37%

/i/ Word-Final	**<i>**	**<e>**
Dataset 2	**93%**	7%

/ɪ/	**73**	**<e>**	**<i>**
Dataset 2	**75%**	22%	3%

/i:/	**<i>**	*Other*
Dataset 2	**98%**	2%

/u/ /u:/	**<u>**	**<ou>**
Dataset 1	38%	**62%**
Dataset 2	**58%**	42%

/o:/	**<o>**	**<ou>**	**<u>**
Dataset 2	**74%**	14%	12%

/o/	**<o>**	**<u>**	**<ø>**	**<ou>**
Dataset 2	**85%**	8%	4%	3%

B. Variable Consonants

	Grapheme / Frequency	
/ʃ/	**<sh>**	**<ch>**
Dataset 1	**55%**	45%
Dataset 2	**60%**	40%

/ħ/	**<7>**	**<h>**
Dataset 1	**71%**	29%
Dataset 2	**61%**	39%

/x/	**<kh>**	**<5>**
Dataset 2	**54%**	46%
/ɣ/	**<gh>**	**<8>**
Dataset 2	**54%**	46%

/C.C/	**<C>**	**<CC>**
Dataset 2	**80%**	20%

I. Successfully Predicted Convergence

Table 11.2 – Predicted vs. Realised Orthographic Form – *"Alright, so"*

Predicted			Actual				
IPA	**Grapheme**	**%**	**IPA**	**Grapheme**	**%**	**Tokens**	**%Diff.**
/tˢ/	**t**	**100%**	**/tˢ/**	**t**	**100%**	***48***	**0%**
/a/	**a**	**95%**	**/a/**	**a**	**94%**	***45***	**-1%**
	ø	5%		**ø**	6%	***3***	**+1%**
/j.j/	**y**	**80%**	**/j.j/**	**y**	**77%**	***37***	**-3%**
	yy	20%		**yy**	21%	***10***	**+1%**
				i	2%	***1***	**+2%**
/ə/	**e**	**79%**	**/ə/**	**e**	**83%**	40	**+4%**
	ø	14%		**ø**	13%	6	**-1%**
	i	7%		**i**	4%	2	**-3%**
/b/	**b**	**100%**	**/b/**	**b**	**100%**	48	**0%**

We see to the left the generalised predicted frequencies for each phoneme of the word /tˢaj.jəb/ (*"alright, so"*) on the basis of the total frequencies derived from Dataset 2, while to the right we see the actual frequencies with which each phoneme appeared within this specific word in Dataset 2. The Lexemic-Aggregational model here predicts the emergent forms nearly perfectly, with only minor differences (<e> overrepresented by 4%, <a> and <y> underrepresented by 1% and 3% respectively). The form <*tyeb> would indicate /tje:b/, meaning "*clothes*", and so does not appear at all, potentially reducing omission of <a>, though only by 1% because omission can still occur in other forms, such as <tyb>. Taking the most frequent graphemic choice by phoneme, our predictive table suggests the emergence of <tayeb> as a convergent form, and this is precisely what we find in Dataset 2:

Table 11.3 – Token Breakdown - *"Alright, so"*

<tayeb>	**31**
<tayyeb>	8
<tayb>	3
<tyb>	3
<tayyib>	2
<taieb>	1

II. Limited Prediction of Convergence

Table 11.4 – Predicted vs. Realised Orthographic Form – *"Yesterday"*

Predicted			Actual				
IPA	Grapheme	%	IPA	Grapheme	%	Tokens	%Diff.
/m/	m	100%	/m/	m	98%	*128*	-2%
				n	1%	2	+1%
				ø	1%	1	+1%
/b/	b	100%	/b/	b	100%	*131*	0%
/e:/	e	79%	/e:/	e	89%	*117*	+10%
	ei	14%		ei	2%	*3*	-12%
	a	7%		a	5%	*6*	-2%
				ø	4%	*5*	+4%
/r/	r	100%	/r/	r	100%	*131*	0%
/ɪ/	e	75%	/ɪ/	e	73%	*95*	-2%
	ø	22%		ø	23%	*30*	+1%
	i	3%		l	4%	*6*	+1%
/ħ/	7	60%	/ħ/	7	57%	*75*	-3%
	h	40%		h	43%	*56*	+3%

In more complex cases, our model is less successful, though it still predicts a convergent form, in this case a couplet on account of the presence of the voiceless pharyngeal fricative in the word /mbe:rɪħ/ meaning "*yesterday*". The <m> we took to be invariable in fact shows minor tokens of <n> as well as omission, which are the result not of graphemic but phonetic variation (with /nbe:rɪħ/ and /be:rɪħ/ being alternative pronunciations). The overall expected frequency of <ei> falls because the /e:/ of this word derives from SA /a:/ and not /aj/, indicating that we might improve our model by dividing our predicted forms for /e:/ depending on the etymological origin of the word in question. For /ɪ/, our model predicts the graphemic frequencies almost perfectly, with variations of 1% and 2% only, while the higher <h> over <7> is fully in line with our expectation of the high-frequency familiarity effect of convergence, as this word also shows the expected convergent couplet predicted by our table as <mbere7>/<mbereh>:

Table 11.5 – Token Breakdown - /ħ/-Split – *"Alright, so"*

mbere7	**52**	**34**	**mbereh**
mber7	11	14	mberh
mberi7	3	3	mberih
mbr7	2	3	mbrh
mbeire7	3	0	mbeireh
mbare7	2	1	mbareh
nbare7	2	0	nbareh
bare7	0	1	bareh

III. Unsuccessful & Unpredictable Convergence

Table 11.6 – Predicted vs. Realised Orthographic Form – *"God-willing"*

Predicted			**Actual**				
IPA	**Grapheme**	**%**	**IPA**	**Grapheme**	**%**		
IPA	Variant	Variant%	**IPA**	Variant	Variant%	**Tokens**	**%Diff.**
/ə/	**e**	**79%**	**/ə/**	**e**	**5%**	***10***	**-74%**
	ø	14%		**ø**	**75%**	***149***	**+61%**
	i	7%		**i**	20%	***40***	**+13%**
/n/	**n**	**100%**	**/n/**	**n**	**100%**	***199***	**0%**
/ʃ/	**sh**	**60%**	**/ʃ/**	**sh**	**57%**	***114***	**-3%**
	ch	**40%**		**ch**	43%	***85***	**+3%**
/a:/	**a**	**100%**	**/a:/**	**a**	**100%**	***199***	**0%**
/l.l/	**l**	**80%**	**/l.l/**	**l**	**27%**	***53***	**-53%**
	ll	20%		**ll**	73%	***146***	**+53%**
/a/	**a**	***100%***	**/a/**	**a**	**46%**	***91***	**-54%**
	ah	*0%*		**ah**	**54%**	***108***	**+54%**
	øh	*0%*		**øh**	0%	**0**	**0%**

The limitations of our model are apparent in the word /ənʃa:la/, meaning "*God-willing*", where our predictions are- at least in some positions- very far from what we find in the data, as a result of a series of factors that occur in this word all at once. The vastly lower <e> (74% less than predicted) leads to the overrepresentation of omission by 61% and of <i> by 13%. This is most likely an etymological effect based on the writing of this word in standard

Arabic with <إن>, mirrored most closely in the Roman script by the form <in>. The overrepresentation of word-initial omission is due to phonetic variation, where this word is frequently pronounced as /nʃa:la/, without sounding the initial schwa. Finally, the overturned ratio in favour of reduplicated <ll> over single <l> for the geminate consonant /l.l/ is partly etymological given the standard Arabic writing of the second part of the word (*"Allah",* meaning God) using the ligature <الله> which consists of reduplicated <لل> (<ll>), as well as being likely affected by the conventional use of the form <Allah> in standard English and other languages (as discussed in the context of the voiceless pharyngeal fricative in 8.2.4). This also explains the dramatically lower incidence of <a> for the word-final /a/, for which <ah> becomes vastly overrepresented, by 54%. Altogether, the sum of these effects means our Lexemic-Aggregational model is unable to predict how this word is written by users of LQA CMCR without accounting for the specific factors at play. This, in addition to the presence of /ʃ/ with its dual <ch> and <sh> realisations (one of the few phonemes our model predicts accurately in this instance), means that there is no emergence of any single convergent form for this word, but instead four pairs of competing forms which we see below contrasted along the axis of the representation of /ʃ/:

Table 11.7 – Token Breakdown - /ʃ/-Split - *"God-willing"*

Tokens	**<sh> forms**		**<ch> forms**	
57	nshallah	**27**	**30**	nchallah
48	nshalla	**24**	**24**	nchalla
32	inshallah	**28**	**4**	inchallah
30	nshala	**23**	**7**	nchala
14	nshalah	10	4	nchalah
5	enshallah	1	4	enchallah
5	enshala	0	5	enchala
4	inshala	1	3	inchala
4	inshalla	0	4	inchalla

11.2 Conventionalisation & Standardisation in LQA CMCR

11.2.1 Academic & Native Perspectives of LQA CMCR

We proposed in Chapter 4 two primary ways in which users of non-standard orthographies can resolve the communicative difficulties of transcriptional, self-orthographical writing:

either reversion towards the standard form (in the case of Type 1/SR), or else by the grassroots emergence of new conventions (see 4.4). While we have seen elements of SA writing adopted and transposed onto the Roman script writing of LQA CMCR, the change of script has resulted in additional close relationships with the writing of StE and StF, which contribute to the pool of resources available to users of LQA CMCR. As a result, the resources of LQA CMCR derive from a variety of different orthographical sources that contribute in different ways and to different, often competing extents, and yet none of which can be taken as the sole source or indeed *standard reflex* of this non-standard system, given that the rewriting from SA into the Roman script undoes any potential standard relationship while the writing of StE and StF provides a loose orthographical basis but no standard relationship of any kind either. This is primarily what marks it as a Type 2/NSR orthography. While we have understood LQA CMCR to be *non-standard* in an academic context, we recall (from 4.4.2) that the division of a linguistic space into standard and non-standard is an ideological, cultural and political construct that originated in Europe (3.2.3) and spread across the globe primarily through colonialism and the perceptions of prestige for standard language and writing through the adoption of SLC. This is, however, a single *axis of differentiation,* and Gal (2018) has demonstrated what other such axes look like (3.2.2). The label *non-standard* also comes with the inherent implication of at least one of two things: either a move away from a standard orthography (as in Type 1/SR), or else an anticipated development of a standard by way of standardisation, and in many cases both at once. LQA CMCR neither derives directly from any single standard form, nor is there any reason to believe that it is likely to undergo any of the standardisational pressures that would lead to the emergence of a standard form, considering the cultural context and the attitude of most speakers of Arabic towards the value of SA and SA writing (see 1.3.1 and 5.3.2). As a result, the labelling of writing such as that of LQA CMCR as *non-standard* is only meaningful as far as academic convenience is concerned, given the characteristics shared by *unstandardised* orthographies that do not fall within the narrow confines of standard writing. We understand LQA CMCR as an orthography which provides its users with a rich collection of orthographic resources allowing for both expressive and effective communication on a spectrum from transcriptional to conventional, within which exists both variation and some amount of conventions emerging by grassroots means. This is not,

however, a meaningful ideological categorisation; we better understand non-standard writing not in direct opposition to standard writing, but as *writing that is not standardised*.

There are also important implications within this discussion with regards to how speakers of Tripolitan LQA perceive their speech and in particular their CMCR writing. Even if we understand LQA CMCR to be outside the standard versus non-standard differential divide, we cannot declare that it is used in a culture that exists outside of SLC, in which case we argued (in 4.4.2) for the full rejection of terms of *standard* and *non-standard* as un-useful and prejudiced on the premises of the western arrangement of linguistic variation. Speakers of LQA and users of LQA CMCR, however, are certainly participants in SLC, whether in the historical prestige of SA, or whether in the form of MSA, moulded in the image of the western standard (see 1.2.4), and so can be expected to subscribe to notions of *standardness* and *correctness*. The non-standard orthographies of QA such as LQA CMCR, however, are uniquely positioned, ironically as a result of the inherent perception of them not being *proper* orthographies. Yaghan (2008, see 5.3.2) argues that the writing of QA with the Roman script is perceived as being error-free, within which 'typos' (by which he means misspellings) do not exist as such. It is in fact the initial rejection of LQA CMCR as a *proper* orthography that allows its users the freedom of linguistically and socially expressive writing, largely free from perceptions of *correctness* and removed from diachronous ideological considerations, beyond the synchronous question of whether this writing should be used in the first place- which has long been answered in how widespread LQA CMCR has become, whether or not its use is ideologically *accepted* even by those who use it. As SA continues to fulfil most of the functions of the standard- particularly the *prestige functions* thereof- the use of QA dialects and writing is afforded a type of freedom more usually associated with cultures entirely outside of SLC. Though low-prestige judgements of non-standard forms certainly do exist, they do little to hinder the communicative needs that are met by CMCR (and, increasingly, CMCA). There are of course exceptions, such as two participants in Dataset 2 who refused to use numerical graphemic solutions like <3> or <7> on a principled basis (speaking to one of them after the interview, they informed me that they did not find these forms *proper*), and instead insist on highly ambiguous reduplicated vowel resolutions such as <aa> instead of <3a> for the syllable /ʕa/ (see 9.2.5). Others use something similar to what Panovic (2018) calls *script-fusing,* in this case essentially being

orthographic reflections of what is otherwise known as *translanguaging* (see 5.1.1 II) whereby Arabic-script forms are inserted into their writing, particularly in pious phrases such as "*God-willing*" (<إن شاء الله>) or any mention of *Allah* (*s.w.t.*) <الله>, though they do not go to the full extent of using both Arabic and Roman script productions in the same words, such as <**aحmad**> (Panovic, 2018, 85). Even these multi-script sentences, like the rejection of <3> and <5>, are rare and idiosyncratic exceptions to the overall tendencies of users of LQA CMCR, who utilise the fullness of the resources available to them, with little ideologically-motivated limitation. Despite the SLC context within which it very much exists, the notion of *correctness* is largely glossed over in the writing of LQA CMCR, nor is there any real desire for *prescriptive power* among most users (certainly in the Tripolitan context), who are content with the prescriptive power of the SA that they, like speakers of other QA dialects, consider to be a native possession (Albirini, 2016, 33).

The availability of resources rooted in both StF and StE allows some degree of identity performance whether conscious or otherwise in the choice of which sub-orthographical series of resources are utilised, and is usually accompanied by more typical translanguaging in the use of StF and StE words alongside LQA ones. The role of identity performance, however, is lower in the case of LQA CMCR than Hinrichs (2004) found for JC, and Deuber and Hinrichs (2007) for NP, but this should not be surprising, given that those are Type 1/SR orthographies with greater flexibility and optionality in which forms are used, whereby the degree of expressivity is a choice predicated on the distance any individual is prepared to take from the standard form in order to express their dialectal variation. For the majority of Tripolitan LQA speakers, the conventions underpinning LQA CMCR are practical choices made primarily for the sake of communicability, recalling Lüpke's (2018) discussion of the use of orthographic features in a West African context without accompanying perceptions of identity related to the languages from which these features derive (see 2.4.2). Though it is certainly feasible that identity can be performed by means of choice between StF or StE features, this is more typical of the writing of Beiruti LQA, where French orthographical features are valued with high prestige, as the French language itself is; Tripolitans, on the other hand, are likely to perceive both French and English as *foreign* (Shaaban & Ghaith, 2002). Thus a preference for the use of conventions from one or the other in a LQA CMCR is, in the first instance, most likely to be based on the language an individual is most familiar

with, in addition to the admixture that has occurred as part of the assimilation of features from both StE and StF as available orthographic resources within LQA CMCR.

11.2.2 Not Standardisation but Conventionalisation

In the course of our work, we have made a series of important distinctions in order to develop a refined understanding of the various forces at work within our thesis, beginning with separate (but closely interrelated) understandings of language and writing, and therefore between the standardisation of language and the standardisation of writing. We have also defined standardisation and conventionalisation as separate (but overlapping) processes, and within non-standard writing specifically, have made a novel and important distinction between Type 1/SR and Type 2/NSR. Joseph's (1987) differentiation between *standard language* and *language standards* (3.1.1) has been central to the distinction we have developed between conventionalisation and standardisation. What Joseph calls *relative language standards* can emerge within any language, but *standard language* is necessarily imposed. We have added to this Gal's view (3.2.2) that any language arranged along any axis of differentiation that can be *normalised*, and indeed, *conventionalised*-processes not exclusive to the domain of SLC. These same conventions *can* become codified through standardisation, by the process which Milroy (2001) calls the *imposition of uniformity* (3.2.1), should the language become subjected to the right pressures. Within this context we relabelled what previous studies have termed grassroots *standardisation* as grassroots *conventionalisation* (5.2), as in the studies of Hinrichs (2004), Deuber and Hinrichs (2007) and Rajah-Carrim (2008). We further understand conventionalisation as the interplay of transcriptional versus conventional forms (first discussed in 4.3.2), which we envisage as a spectrum between fully transcriptional and fully conventional writing. While fully conventionalised writing is tantamount to the codified writing of standard orthographies, within non-standard writing we find a co-existence of transcriptional and conventional writing, usually as two sets of resources for users of the orthography to draw upon, depending on both linguistic and social context. In this way, we understand the process of grassroots conventionalisation as the organic, user-driven emergence of these conventional forms.

I. Conventionalisation through Convergent Forms

From within a SLC perspective, the emergence of high-frequency preferential forms for each phoneme can be interpreted as a means by which further uniformity can be imposed by means of elimination of the variation that characterises these conventions, such as a choice between <7> or <h>, <sh> or <ch> and <u> or <ou>, or a strict ruleset for where vowel omission can or cannot take place. Such an approach assumes that standardisational pressure is imminent, or even inevitable, while in reality, it is neither. Absent such an external pressure, there is no reason to assume- or even desire- any reduction of variation. Instead, a rich repertoire of both convergently conventional forms and expressively transcriptional forms characterise this orthography, in keeping with its primary role as an online language of CMC utility, wherein its flexible nature is not restrictive but rather encourages communicative expressivity. Unlike the ideological impositions typical in SLC, the emergence of conventional forms need not mean the replacement of transcriptional ones, but instead consists of the addition of new resources to enrich the orthographic choice available. Moreover, the frequency-based convergence that our Lexemic-Aggregational Model is based on belies the influence of other factors on the orthographical choices of users of LQA CMCR, including etymological and semantic motivations, which as we saw in 11.1.6, can make any statistical model of convergence fall short of successful prediction. These orthographical motivations must always be accounted for independently, whether it is an etymological spelling retaining a SA form or the need for semantic clarity overriding any predicted spelling. While our Lexemic-Aggregational approach is capable of predicting convergence in many cases, individual graphemic-phonetic popularity is far from the only factor affecting the choices of users of LQA CMCR.

II. Conventionalisation through Prestige Forms

The emergence of prestige-based conventional forms also functions on an axis between transcriptional and conventional writing. While frequency-convergence sees the reduction of graphemic variation for the representation of the same phonetic realisations, prestige-motivated conventions see a reduction in the phonetic detail of orthographic forms, whereby spellings no longer represent the phonetic Tripolitan LQA pronunciation, with words instead conventionally indicated using orthographical forms based on the Beiruti LQA pronunciation, and in which we therefore see a more dramatic reduction of

transcriptionality in writing. We have observed this most clearly in the word most commonly written as <badi>, representing Beiruti LQA /bad.di/ despite the majority of Tripolitan LQA speakers realising this same word for "*I want"* instead as /bəd.di/, and which is represented more transcriptionally in Tripolitan LQA CMCR as <bedi>. In this, we observe both our primary means of conventionalisation overlapping: the form <badi> is *conventional* in being a representation of a pronunciation not generally used by most of those writing it; on the other hand, <bedi> is a closer phonetic realisation of the spoken form, and thus more *transcriptional*, and yet it is itself still more *conventional* than other, more transcriptional orthographical realisations of the Tripolitan LQA spoken form (such as for example <biddi>), given that <bedi> utilises the highest-frequency graphemic forms for each phoneme and yet, unlike <badi>, retains the phonetic detail of the localised Tripolitan realisation. In this we observe both different types as well as different degrees of (co-existing) conventions. From our discussion of disglossia (1.3.2) and particularly prestige in the Arabic-speaking world (1.3.3), we know the capital urban dialect commands a certain degree of prestige and fills some of the prestige-functions that SA does not. Prestige, just as with conventionalisation itself, is not exclusive to standard language or SLC, and is part of Joseph's description of synecdochal emergence, occurring as a result of the hierarchisation universal in linguistic behaviour (Joseph, 1987, 60; 3.1.1). Beiruti LQA is therefore associated with high prestige, though as we saw in 6.1.2, such perceptions are complicated and vary depending on other social and sociolinguistic factors. Nevertheless, we have observed a clear effect of the Beiruti LQA prestige-dialect on both the speech and writing of Tripolitan speakers of LQA and users of LQA CMCR, where it is another means by which written conventions organically develop.

11.2.3 The Flexible Resources of Unstandardised Expression

The unstandardised nature of LQA CMCR means that its conventions are never fully settled, but always in flux. We observed this in the shift between Dataset 1 and Dataset 2, where the ratio of <7> to <h> shifted between 71:29 to 61:39, or indeed in the case of /ʃ/ and the shift in preference from <ch> in Dataset 1 to <sh> in Dataset 2. Whether this is a result simply of the different sets of individuals that contributed to each dataset and the demographic differences between them, or the result of a shift in time, it nevertheless points to fluctuation and dynamism. Absent any imposition of uniformity and a process of

standardisation, a unstandardised orthography is not subject to the same pressures that limit the possibility of variation. Nor is grassroots conventionalisation a one-time event, but instead an ongoing and ever-shifting process. This results in an orthography that is not only flexible within the resources- conventionalised and otherwise- that are available to its users, but also in their arrangement over time. The CMC nature of LQA CMCR both contributes to this fluctuation and, at the same time makes this flexibility and changeability highly useful to its users for the primary means of communication within which this orthography is utilised, allowing flexibility of expression that can also shift in social context, between ambiguity with friends and family who are familiar with the specific conventions utilised by an individual who might otherwise draw on more conventional forms for communication with strangers. The available resources of LQA CMCR do not only allow for, but to some degree necessitate transcriptional expressivity and the use of local colloquial forms in writing, co-existing with conventional forms that emerge from frequency-convergence, and which are in flux and modified by a number of factors, from semantic to etymological, as well as with conventions derived from prestige forms, which, when used, represent further reduction still in the transcriptional nature of this writing. All these resources are available for users of LQA CMCR, without the standard language concept of *correctness* governing their use in any real way, as a result both of the unstandardised nature of the orthography, as well as its Type 2/NSR nature. Conventionalisation in LQA CMCR leads to a richer array of available resources, without the reductive effects of standard language culture, and for the speakers of LQA in Tripoli who utilise this orthography, this leads to a convenient, effective and expressive means of self-expression, quite unlike that of the axis of differentiation premised on the duality of standard and non-standard, but instead, existing largely outside of it.

11.3 Epilogue: Future Study

We have encountered in the course of our work a number of questions that merit further study. Further experimental work of a similar nature to ours can be conducted to probe how the features observed within this study have changed in the time since our data was collected. Within this exist possibilities for further following ratios such as that of <h> to <7>, or whether the balance between StF and StE features has continued to shift in favour of the latter. Such a study might be able to ascertain whether the conventions identified in

our thesis have persevered, continued to develop, or have been reversed entirely, particularly fledgling features such as <ei> as a generalised representation for /e:/. It would also allow for the study of new conventions, such as the expected replacement of convergent grapheme <i> for word-final /i/ with <e> instead, on the basis of the growing phonetic exaggeration of the Beiruti LQA pronunciation of /e/ and its widespread use in television and other Lebanese media. The gender-based dialectal differentiation observed in 6.1.2 was in its infancy at the time of data-collection, but also provides a potentially interesting avenue for further work on the basis of whether the masculine-prestige associated with Old Tripolitan LQA and the feminine-prestige associated with Beiruti LQA is also orthographically observable, potentially leading to more representations of Old Tripolitan speech orthographically than the single instance we found in our data, as well as clearly gendered conventional resources. Additionally, within the Tripolitan LQA landscape, there are also possibilities for studies centring around the distinction between the orthographical productions of Old Tripolitans and New Tripolitans. Within the same scope of Tripoli, a further, more detailed delineation of the graphemic inventory of Tripolitan LQA CMCR would also be of interest, and with it a refining of the Lexemic-Aggregational Model we have developed. The introduction of new methodologies is also possible, including a greater focus on the attitudes of the individuals interviewed towards their use of LQA and LQA CMCR, determining to which extent Tripolitan attitudes are in line with those of other speakers of LQA and of other QA dialects, and whether there is notable change in the attitudes towards writing in the Roman script. Specific features of the writing of LQA CMCR also bear further study, such as the phenomenon of vowel omission, particularly in cases where the omitted vowels allow for strategies of *indeterminacy* such as those proposed by Hillewaert (2015), which she identifies in the context of Kenyan CMC writing as a means of circumventing societal and prestige pressures pertaining to certain vernaculars while still maintaining their coded use. Such an approach to cases where vowel omission leads to such indeterminate orthographical forms in LQA CMCR or the CMC of other QA variants has potential to be of much interest, whether in Hillewaert's context of societal pressure or other cases such as where vowel omission can be strategically utilised to avoid marking gender, sometimes used where the gender of the recipient of a message is not known to the writer (for example in forms like <bedk>, where both masculine /bəd.d**a**k/ and feminine /bəd.d**ɪ**k/ readings are equally viable).

More generally, the study of both LQA as well as QA dialects has been constrained by attitudes informed by SLC, and I believe that the field would benefit greatly from the adoption of an approach that acknowledges the unique nature of non-standard writing on its own terms, rather than as a temporary and undesirable precursor to an inevitable standardisation. A new approach, wherein variation is understood in terms of available resources, free from preconceived prescriptions of uniformity, will allow for the development of a mature sociolinguistics of the CMC writing of QA to complement the mature sociolinguistics of writing proposed by Blommaert (2013). Within this exists a scope for a variety of studies, such as work which incorporates Arabic script CMCA data, or indeed which compares the use of CMCA (as Type 1/SR) and CMCR (as Type 2/NSR) within the confined context of Tripoli, or any other constrained locale. There are great possibilities beyond LQA for the study of grassroots conventionalisation of other Type 2/NSR orthographies, which would also complement our findings by determining whether the factors we have described are unique to the LQA context or generalisable across all Type 2/NSR writing. Finally, in the context of our understanding that the grassroots conventionalisation we observe in a CMC context is, by and large, the same process that languages and orthographies undergo in a historical synecdochal stage also opens up the possibilities of using living languages whose non-standard orthographies are undergoing this process as a means of better understanding how the organic development of written (and even, by analogy, spoken) conventions emerge, which can be generalised to a broader understanding of the phenomenon and potentially applied to unobservable historical instances of the same process, one example being the Middle English orthographic situation where the non-standard and uncodified variability resembles that which we have found in LQA CMCR and other conventionalised but unstandardised orthographies.

References

Abdallah, S. (2008). Online chatting in Beirut: sites of occasioned identity-construction. *Ethnographic Studies, 10*, 3-22. http://doi.org/10.5449/idslu-001104695

Abdel-Jawad, H. R. E. (1981). Lexical and phonological variation in spoken Arabic in Amman. https://repository.upenn.edu/dissertations/AAI8207924

Abdel-Jawad, H. R. E. (1987). Cross-dialectal variation in Arabic: competing prestige forms. *Language in Society*, *16*(3), 359–67. https://www.jstor.org/stable/4167860

Abdulaziz, M. H. (1986). Factors in the development of modern Arabic usage. *International Journal of the Sociology of Language, 62*, 11–24. https://doi.org/10.1515/ijsl.1986.62.11

Abu Elhij'a, D. A. (2012). Facebook written Levantine vernacular languages. *The Levantine Review*, *1*(1), 68–105. https://doi.org/10.6017/lev.v1i1.2157

Abu Elhij'a, D. (2014). A new writing system? Developing orthographies for writing Arabic dialects in electronic media. *Writing Systems Research,6*(2), 190-214. https://doi.org/10.1080/17586801.2013.868334

Al-Muhannadi, M. (1991). *A sociolinguistic study of women's speech in Qatar* [Unpublished doctoral dissertation]. University of Essex. https://ethos.bl.uk/OrderDetails.do?uin=uk.bl.ethos.292085

Al-Tamimi, Y., & Gorgis, D. (2007). Romanised Jordanian Arabic e-messages. *International Journal of Language, Society and Culture, 21*, 1–12.

Al-Wer, E. (2014). Language and gender in the Middle East and North Africa. In S. Ehrlich et al (Eds.), *The handbook of language, gender, and sexuality* (pp. 396–411). Wiley-Blackwell. https://doi.org/10.1002/9781118584248.ch20

Albirini, A. (2011). The sociolinguistic functions of codeswitching between standard Arabic and dialectal Arabic. *Language in Society, 40*(5), 537–562. https://doi.org/10.1017/S0047404511000674

Albirini, A. (2016). *Modern Arabic sociolinguistics: Diglossia, variation, codeswitching, attitudes and identity*. Routledge. https://doi.org/10.4324/9781315683737

Álvarez-Cáccamo, C., & Herrero Valeiro, M. J. (1996). O continuum da escrita na Galiza: Entre o espanhol e o português. *Agália: Revista da Associaçom Galega da Lingua, 46*, 143-56. http://agal-gz.org/faq/lib/exe/fetch.php?media=agalia:n46_o_continuum_da_escrita_na_galiza.pdf

Androutsopoulos, J. K. (2000). Non-standard spellings in media texts: the case of German fanzines. *Journal of Sociolinguistics, 4*(4), 514–533. https://doi.org/10.1111/1467-9481.00128

Androutsopoulos, J. (2006). Introduction: Sociolinguistics and computer-mediated communication. *Journal of sociolinguistics*, *10*(4), 419–438. https://doi.org/10.1111/j.1467-9841.2006.00286.x

Androutsopoulos, J. (2009). "Greeklish": transliteration practice and discourse in the context of computer-mediated digraphia. In A. Georgakopoulou & M. S. Silk (Eds.), *Standard Languages and Language Standards: Greek, past and present* (pp. 221–49). De Gruyter Mouton. https://doi.org/10.1515/9781614511038.359

Androutsopoulos, J. (2015). Networked multilingualism: some language practices on Facebook and their implications. *International Journal of Bilingualism*, *19*(2), 185–205. https://doi.org/10.1177%2F1367006913489198

Aoun, J., Benmamoun, E., & Choueiri, L. (2010). *The Syntax of Arabic*. Cambridge University Press. https://doi.org/10.1017/CBO9780511691775

Appadurai, A. (1996). *Modernity at Large*: *Cultural Dimensions of Globalization*. University of Minnesota Press.

Auer, P. (2005). Europe's sociolinguistic unity, or: a typology of European dialect/standard constellations. In N. Delbecque (Ed.), *Perspectives on variation: Sociolinguistic, historical, comparative* (pp. 7–42). De Gruyter. https://doi.org/10.1515/9783110909579.7

Badawi, S. A. (1973). *Mustawayat al-lugha al-ʿ arabiyya al-muʿasira fi misr* [Levels of Contemporary Arabic in Egypt]. Dar al-Ma'arif.

Baker, P. (1997). Developing Ways of Writing. In A. Tabouret-Keller et al (Eds.), *Vernacular literacy: A re-evaluation*, (pp. 93–141). Oxford University Press.

Bassiouney, R. (2006). *Functions of code switching in Egypt: Evidence from monologues*. Brill. https://doi.org/10.1163/9789047417132

Baym, N. K. (1998). The emergence of on-line community. In S. Jones (Ed.), *CyberSociety 2.0. revisiting computer-mediated communication and community* (pp. 35–68). Sage. https://doi.org/10.4135/9781452243689.n2

Baym, N. K. (2003). Communication in online communities. In K. Christiansen & D. Levinson (Eds.), *Encyclopedia of community* (Vol. 3, pp. 1016–1017). Sage. https://doi.org/10.1075/eww.19.2.14tab

Berjaoui, N. (2001). Aspects of the Moroccan Arabic orthography with preliminary insights from the Moroccan computer-mediated communication. In M. Beisswenger (Ed.), *Chat-Kommunikation: Sprache, Interaktion, Sozialitat & Identitat in synchroner computervermitelter Kommunikation* (pp. 431–468). Bidem-Verlag.

Besnier, N. (1988). The linguistic relationships of spoken and written Nukulaelae registers. *Language 64*(4), 707–736. http://doi.org/10.2307/414565

Bidaoui, A. (2017). Revisiting the Arabic diglossic situation and highlighting the socio-cultural factors shaping language use in light of Auer's (2005) model. *International Journal of Society, Culture & Language*, *5*(2), 60–72. http://www.ijscl.net/article_27577.html

Bishai, W. B. (1966). Modern Inter-Arabic. *Journal of the American oriental society*, *86*(3), 319–323. http://doi.org/10.2307/597040

Blanc, H. (1960). Style variations in Arabic: A sample of interdialectal conversation. In C. A. Ferguson (Ed.), *Contributions to Arabic linguistics* (pp. 81–156). Harvard University Press.

Blau, J. (1981). *The renaissance of modern Hebrew and modern standard Arabic: Parallels and differences in the revival of two Semitic languages*. University of California Press. https://doi.org/10.1017/S0026318400011780

Blommaert, J. (2008). *Grassroots literacy: Writing, identity and voice in Central Africa*. Routledge. http://dx.doi.org/10.4324/9780203895481

Blommaert, J. (2013). Writing as a sociolinguistic object. *Journal of Sociolinguistics*, *17*(4), 440–459. https://doi.org/10.1111/josl.12042

Bloomfield, L. (1933). *Language*. Holt, Rinehart and Winston.

Bolander, B., & Locher, M. A. (2010). Constructing identity on Facebook: Report on a pilot study. In K. Junod & D. Maillat (Eds.), *Performing the Self* (165–185). Narr.

Bou Tanios, J. (2016). *Language choice and Romanization online by Lebanese Arabic speakers.* [Unpublished Master's thesis], Universitat Pompeu Fabra. http://hdl.handle.net/10230/27669

Canut, C., & Dumestre, G. (1993). Français, bambara et langues nationales au Mali. In D. Robillard & M. Beniamino (Eds.), *Le français dans l'espace francophone* (p. 219–228). Honoré Champion.

Castells, M. (2000). *The Rise of the Network Society* (2nd ed.). Blackwell. http://doi.org/10.1002/9781444319514

Caton S. 1991. Diglossia in North Yemen: a case of competing linguistic communities. *Southwest Journal of Linguistics 10*(1), 143–59

Chejne, A. (1969). *The Arabic language: Its role in history*. University of Minnesota Press. https://www.jstor.org/stable/10.5749/j.ctttv9cb

Clifton, J. M. (2013). Dialects, orthography and society. *Work papers of the summer institute of linguistics, University of North Dakota session*, *53*(1), 1–8. http://doi.org/10.31356/silwp.vol53.01

Collot, M., & Belmore, N. (1996). A new variety of English. In S. C. Herring (Ed.), *Computer-mediated communication: Linguistic, social, and cross-cultural perspectives* (pp. 13–28). John Benjamins. https://doi.org/10.1075/pbns.39.04col

Coulmas, F. (2013). *Writing and society: An introduction*. Cambridge University Press. https://doi.org/10.1017/CBO9781139061063

Crystal, D. (2001). *Language and the Internet*. Cambridge University Press. https://doi.org/10.1017/CBO9780511487002

Daniels, P. T. (2013). The Arabic Writing System, in J. Owens (Ed.), *The Oxford Handbook of Arabic Linguistics* (pp. 412-432). Oxford University Press. http://doi.org/10.1093/oxfordhb/9780199764136.013.0018

De Saussure, F. (1978). *Cours de linguistique générale*. (W. Baskin, Trans.). Collins. [Original work published 1916].

Deuber, D., & Hinrichs, L. (2007). Dynamics of orthographic standardization in Jamaican Creole and Nigerian Pidgin. *World Englishes*, *26*(1), 22–47. https://doi.org/10.1111/j.1467-971X.2007.00486.x

Diem, W. (1974) *Hochsprache und Dialekt im Arabischen.* Steiner.

Diringer, D. (1968). *The Alphabet: A Key to the History of Mankind*. Funk & Wagnalls.

Donaldson, C. (2015). The social life of orthography development. *Working Papers in Educational Linguistics (WPEL)*, *30*(2), 1–12. https://repository.upenn.edu/wpel/vol30/iss2/1

Duggan, M., & Brenner, J. (2013). *The demographics of social media users, 2012* (Vol. 14). Pew Research Center's Internet & American Life Project.

Dürscheid, C. (2004). Netzsprache – ein neuer Mythos. *Osnabrücker Beiträge zur Sprachtheorie, 22*(1), 141–157. https://www.mediensprache.net/archiv/pubs/2409.pdf

Eid, M. (1990). Arabic linguistics: The current scene. In M. Eid (Ed.), *Perspectives on Arabic Linguistics I* (pp. 3–37). John Benjamins. https://doi.org/10.1075/cilt.63.03eid

Eira, C. (1998). Authority and discourse: Towards a model for orthography selection. *Written Language & Literacy*, *1*(2), 171–224. https://doi.org/10.1075/wll.1.2.03eir

El-Essawi, R. (2011). Arabic in Latin script in Egypt: who uses it and why. In A. Al-Issa & L. S. Dahan (Eds.), *Global English and Arabic: Issues of Language, Culture, and Identity,* (pp. 253-284). Peter Lang. https://doi.org/10.3726/978-3-0353-0120-5

Elgibali, A. 1993: Stability and language variation in Arabic: Cairene and Kuwaiti dialects. In M. Eid & C. Holes (Eds.), *Perspectives on Arabic Linguistics V* (pp. 75–96). John Benjamins. https://doi.org/10.1075/cilt.101.06elg

Ennaji, M. (2007). Arabic sociolinguistics and cultural diversity in Morocco. In E. Benmamoun (Ed.), *Perspectives of Arabic linguistics XIX: Papers from the Nineteenth Annual Symposium on Arabic Linguistics* (pp. 267–277). John Benjamins. https://doi.org/10.1075/cilt.289.18enn

Fabian, J. (1990). *History from below*. John Benjamins. https://doi.org/10.1075/cll.7

Fabian, J. (1993). Keep listening: Ethnography and reading. In J. Boyarin (Ed.), *The ethnography of reading* (pp. 80–97). University of California Press. https://doi.org/10.1525/9780520913431-006

Ferguson, C. (1959a). The Arabic koine. *Language, 35*(4), 616–630. http://doi.org/10.2307/410601

Ferguson, C. (1959b). Diglossia. *Word*, *15*(2), 325–340. http://doi.org/10.1080/00437956.1959.11659702

Ferguson, C. (1968). Myths about Arabic. In J. A. Fishman (Ed.), *Readings in the Sociology of Language* (pp. 375– 381). De Gruyter Mouton. https://doi.org/10.1515/9783110805376.375

Ferguson, C. (1996). *Sociolinguistic perspectives: Papers on language in society, 1959–1994*. Oxford University Press.

Fishman, J. A. (1967). Bilingualism with and without diglossia; diglossia with and without bilingualism, *JSI 23*(2), 29–38. https://doi.org/10.1111/j.1540-4560.1967.tb00573.x

Fishman, J. A. (1971). *Sociolinguistics*. Newbury House.

Fishman, J.A. (Ed.) (1977). *Advances in the creation and revision of writing systems*. De Gruyter Mouton. https://doi.org/10.1515/9783110807097

Gade, T. (2015). Sunni Islamists in Tripoli and the Asad regime 1966-2014, *Syria studies 7*(2), 20–65. http://hdl.handle.net/10023/7174

Gal, S. (2018). Visions and revisions of minority languages. In P. Lane, J. Costa & H. De Korne, H. (Eds.). *Standardizing minority languages: competing ideologies of authority and authenticity in the global periphery* (pp. 222–242). Routledge. https://doi.org/10.4324/9781315647722

Gelb, I. J. (1963). *A study of writing*. University of Chicago Press.

Germanos, M. A. (2007). Greetings in Beirut. In C. Miller et al (Eds.), *Arabic in the city: Issues in dialect contact and language variation* (pp. 147-165). Routledge.

Goody, J. & Watt, I. (1968). The consequences of literacy. In J. Goody (Ed.), *Literacy in traditional societies* (pp. 27–68). Cambridge University Press.

Grace, G. (1981). Indirect inheritance and the aberrant Melanesian languages. In J. Hollyman and A. Pawley (Eds.), *Studies in Pacific Languages and Cultures in Honour of Bruce Biggs* (pp. 255–268). Linguistic Society of New Zealand.

Grace, G. W. (1990). The "aberrant" (vs."exemplary") Melanesian languages. In P. Baldi (Ed.), *Linguistic change and reconstruction methodology* (pp. 155–173). Walter de Gruyter.

Grace, G. W. (1992). How do languages change? (More on "aberrant" languages). *Oceanic Linguistics, 31*(1), 115–130. http://doi.org/10.2307/3622968

Grace, G. W. (1993). What are languages? *Ethnolinguistic Notes, 3*(45), 1–17.

Graff, H. J. (1987). *The legacies of literacy: Continuities and contradictions in western culture and society* (Vol. 598). Indiana University Press.

Haeri, N. (1991). *Sociolinguistic variation in Cairene Arabic: Palatalization and the Qaf in the speech of men and women*. [Unpublished doctoral thesis]. University of Pennsylvania. https://repository.upenn.edu/dissertations/AAI9125661

Hannas, W. C. (2003). *The writing on the wall: How Asian orthography curbs creativity*. University of Pennsylvania Press. https://www.jstor.org/stable/j.ctt3fhw8m

Haugen, E. (1966). *Language conflict and language planning: The case of modern Norwegian*. Harvard University Press. https://doi.org/10.4159/harvard.9780674498709

Hawkins, P. (1983). Diglossia revisited. *Language Sciences, 5*(1), 1–20. https://doi.org/10.1016/S0388-0001(83)80010-2

Heath, S. B. (1983). *Ways with words: Language, life and work in communities and classrooms*. Cambridge University Press.

Heilbroner, R. L. (1999). *The Worldly Philosophers*. Touchstone.

Herring, S. C. (1993). Gender and democracy in computer-mediated communication. *The Electronic Journal of Communication*, *3*(2), 1–17. http://www.cios.org/EJCPUBLIC/003/2/00328.HTML

Herring, S. C. (2000). Gender differences in CMC: Findings and implications. *Computer Professionals for Social Responsibility Journal,* 18(1). http://cpsr.org/issues/womenintech/herring/

Herring, S. C. (2003). Gender and power in online communication. In J. Holmes and M. Meyerhoff (Eds.), *The handbook of language and gender* (pp. 202–228). Blackwell. https://doi.org/10.1002/9780470756942.ch9

Herring, S. C. (2004). Computer-mediated discourse analysis: An approach to researching online communities. In S. A. Barab, R. Kling & J. H. Gray (Eds.), *Designing for Virtual Communities in the Service of Learning* (pp. 338–376). Cambridge University Press. https://doi.org/10.1017/CBO9780511805080.016

Herring, S. C., & Paolillo, J. C. (2006). Gender and genre variation in weblogs. *Journal of Sociolinguistics*, *10*(4), 439–459. https://doi.org/10.1111/j.1467-9841.2006.00287.x

Heryanto, A. (1990). The making of language: developmentalism in Indonesia. *Prisma*, *50*, 40–53. https://arielheryanto.files.wordpress.com/2016/03/1990_50_sept_prisma-the-making-of-language-c.pdf

Hillewaert, S. (2015). Writing with an accent: Orthographic practice, emblems, and traces on Facebook. *Journal of Linguistic Anthropology*, *25*(2), 195–214. https://doi.org/10.1111/jola.12079

Hinrichs, L. (2004). Emerging orthographic conventions in written Creole: computer-mediated communication in Jamaica. *AAA: Arbeiten aus Anglistik und Amerikanistik, 29*(1), 81–109. https://www.jstor.org/stable/43025721

Holes, C. (2004). *Modern Arabic: Structures, functions, and varieties*. Georgetown University Press.

Hopkins, S. (1984). *Studies in the grammar of early Arabic based upon papyri datable to before A.H. 300/A.D. 912*. Oxford University Press.

Horesh, U., & Cotter, W. M. (2016). Current research on linguistic variation in the Arabic-speaking world. *Language and Linguistics Compass*, *10*(8), 370–381. https://doi.org/10.1111/lnc3.12202

Huffaker, D. A. & Calvert, S. L. (2005). c. *Journal of Computer-Mediated Communication, 10*(2). https://doi.org/10.1111/j.1083-6101.2005.tb00238.x

Hussein, R. F., & El-Ali, N. (1989). Subjective reactions of rural university students toward different varieties of Arabic. *Al-'Arabiyya*, *22*(1/2), 37–54. https://www.jstor.org/stable/43208677

Ibrahim, M. (1986). Standard and prestige language: A problem in Arabic sociolinguistics. *Anthropological Linguistics*, *28*(1), 115–126. https://www.jstor.org/stable/30027950

Jaffe, A. (2000). Introduction: Non-standard orthography and non-standard speech. *Journal of sociolinguistics*, *4*(4), 497–513. https://doi.org/10.1111/1467-9481.00127

Jaffe, A., & Walton, S. (2000). The voices people read: Orthography and the representation of non-standard speech. *Journal of Sociolinguistics*, *4*(4), 561–587. https://doi.org/10.1111/1467-9481.00130

Jaffe, A. (2009). Indeterminacy and regularization. *Sociolinguistic Studies*, *3*(2), 229–251. http://doi.org/10.1558/sols.v3.i2.229

Joseph, J. E. (1987). *Eloquence and power: The rise of language standards and standard languages*. Burns & Oates.

Kelly, P. (2018). The art of not being legible. *Terrain*, *70*, 1–24. https://journals.openedition.org/terrain/17103

Knudsen, A. J. (2017). Patrolling a proxy war: citizens, soldiers and Zu'ama in Syria Street, Tripoli. In A. J. Knudsen & T. Gade (Eds.), *Civil-Military Relations in Lebanon* (pp. 71–99). Palgrave Macmillan, Cham. https://doi.org/10.1007/978-3-319-55167-8_4

Kress, G. R. (2000). *Early spelling: Between convention and creativity*. Psychology Press.

Krumbacher, K. (1902). *Das Problem der neugriechischen Schriftsprache*. Königliche Bayerische Akademie.

Kuipers, J. C. (1998). *Language, identity, and marginality in Indonesia: The changing nature of ritual speech on the island of Sumba*. Cambridge University Press. https://doi.org/10.1017/CBO9780511558191

Kurman, G. (1968). *The development of written Estonian*. Indiana University Press/Mouton.

Labov, W. (1990). The intersection of sex and social class in the course of linguistic change. *Language variation and change*, *2*(2), 205–254. https://doi.org/10.1017/S0954394500000338

Lee, C. (2007). Linguistic features of Email and ICQ instant messaging in Hong Kong. In B. Danet &S. Herring (Eds.), *The multilingual internet: Language, culture, and communication online* (pp. 184–208). Oxford University Press. http://doi.org/10.1093/acprof:oso/9780195304794.003.0008

Lindqvist, C. (2003). Sprachideologische Einflüsse auf die färöische Orthographie. *NOWELE North-Western European Language Evolution*, *43*(1), 77–144. https://doi.org/10.1075/nowele.43.06lin

Lüpke, F. (2018). Escaping the tyranny of writing: West African regimes of writing as a model for multilingual literacy. In C. Weth & K. Juffermans (Eds.), *The Tyranny of Writing: Ideologies of the Written Word* (pp. 129–147). Bloomsbury. https://doi.org/10.5040/9781474292436.0013

Maggard, L., Sangma, M., & Ahmad, S. (2007). *A sociolinguistic survey among the Chakma and Tanchangya communities*. Dhaka, Bangladesh, ms.

Marçais, W. (1930). La diglossie arabe. *L'Enseignement public, 97,* 401–409.

Matras, Y. (2005). The future of Romani: Toward a policy of linguistic pluralism. *Roma Rights Quarterly*, *1*, 31–44. https://romani.humanities.manchester.ac.uk//downloads/2/Matras_Pluralism.pdf

McLaughlin, F. (2008). The ascent of Wolof as an urban vernacular and national lingua franca in Senegal. In C. B. Vigouroux, S. S. Mufwene & J. Blommaert (Eds.), *Globalization and language vitality: Perspectives from Africa* (pp. 142-170). Continuum.

Mejdell, G. (2006). *Mixed styles in spoken Arabic in Egypt*. Brill. https://doi.org/10.1163/9789047408987

Miller, C. (2007). Arabic urban vernaculars: Development and change. In C. Miller, E. Al-Wer, D. Caubet, &. J. Watson (Eds.), Arabic in the city: Issues in dialect contact and language variation (pp. 1– 31). London: Routledge.

Milroy, J. (2001). Language ideologies and the consequences of standardization. *Journal of sociolinguistics*, *5*(4), 530-555. https://doi.org/10.1111/1467-9481.00163

Mimouna, B. (2012). *Is English there? Investigating language use among young Algerian users of internet*. [Unpublished doctoral dissertation]. University of Oran, Algeria.

Mitchell, T. (1986). What is educated spoken Arabic? *International Journal of the Sociology of Language, 61*(1), 7–32. https://doi.org/10.1515/ijsl.1986.61.7

Mitchell, T., & El-Hassan, S. (1994). *Modality, mood, and aspect in spoken Arabic, with special reference to Egypt and the Levant*. Kegan Paul International.

Mokroborodova, L. (2008). The “New Spelling” of Russian on the internet in relation to phonetics and orthography. *Scando-Slavica*, *54*(1), 62–78. https://doi.org/10.1080/00806760802494190

Oenbring, R. (2013). Bey or bouy: Orthographic patterns in Bahamian Creole English on the web. *English world-wide*, *34*(3), 341–364. https://doi.org/10.1075/eww.34.3.04one

Official Website of the Municipality of Tripoli. (n.d.). *Tarikh al-Madinah* [History of the City]. http://www.tripoli.gov.lb/ar/node/1

Olson, D. R. (1988). Mind and media: the epistemic functions of literacy. *Journal of communication*, *38*(3), 27–36. https://psycnet.apa.org/doi/10.1111/j.1460-2466.1988.tb02058.x

Ong, W. J. (2013). *Orality and literacy*. Routledge.

Owens, J. (2006). *A linguistic history of Arabic*. Oxford University Press. http://doi.org/10.1093/acprof:oso/9780199290826.001.0001

Owens, J. (2011). Arabic Sociolinguistics. In S. Weninger (Ed.), *The Semitic languages: An international handbook*, (pp. 970–981). Walter de Gruyter. https://doi.org/10.1515/9783110251586.970

Palfreyman, D., & Khalil, M. A. (2003). "A Funky Language for Teenzz to Use:" Representing Gulf Arabic in Instant Messaging. *Journal of Computer-Mediated Communication, 9*(1). https://doi.org/10.1111/j.1083-6101.2003.tb00355.x

Panović, I. (2018). 'You don't have enough letters to make this noise': Arabic speakers' creative engagements with the Roman script. *Language Sciences*, *65*, 70–81. https://doi.org/10.1016/j.langsci.2017.03.010

Papacharissi, Z. (2009). The virtual geographies of social networks: A comparative analysis of Facebook, LinkedIn and ASmallWorld. *New Media and Society*, *11*(1-2), 199–220. https://doi.org/10.1177%2F1461444808099577

Perfettini, F., & Agostini, P. M. (1994). Comprendre ce qu'on écrit. *Arritti, June 6*(4).

Phillipson, R. (1992). *Linguistic imperialism*. Oxford University Press. https://doi.org/10.1002/9781405198431.wbeal0718.pub2

Pietrini, D. (2001). «X' 6 :-(?»: Gli sms e il trionfo dell'informalit`a e della scrittura ludica. *Italienisch, 46*, 92–101.

Preece, J., Maloney-Krichmar, D. and Abras, C. (2003). History of online communities. In K. Christiansen & D. Levinson (Eds.), *Encyclopedia of Community* (pp. 1023–1027). Sage. http://doi.org/10.4135/9781412952583.n365

Preston, D. R. (1982). 'Ritin' Fowklower Daun' Rong: Folklorists' Failures in Phonology. *The Journal of American Folklore*, *95*(377), 304–326. http://doi.org/10.2307/539912

Pullum, G. K. (1999). African American Vernacular English is not standard English with mistakes. In R. S. Wheeler (Ed.), *The workings of language: From prescriptions to perspectives* (pp. 59–66). ABC-CLIO.

Rajah-Carrim, A. (2008). Choosing a spelling system for Mauritian Creole. *Journal of Pidgin and Creole Languages*, *23*(2), 193–226. https://doi.org/10.1075/jpcl.23.2.02raj

Riegert, K., & Ramsay, G. (2013). Activists, individualists, and comics: the counter-publicness of Lebanese blogs. *Television & New Media*, *14*(4), 286–303. https://doi.org/10.1177%2F1527476412463447

Roberts, C., Street, B. (1997). Spoken and written language. In F. Coulmas (Ed.), *The Handbook of Sociolinguistics* (pp. 168–186). Blackwell. https://doi.org/10.1002/9781405166256.ch10

Rothstein, R. A. (1977). Spelling and society: the Polish orthographic controversy of the 1930's. In B. A. Stolz (Ed.), *Papers in Slavic Philology I.* (pp. 225–36). University of Michigan.

Sallam, A. M. (1980). Phonological variation in educated spoken Arabic: a study of the uvular and related plosive types. *Bulletin of the School of Oriental and African Studies, 43*, 77–100. https://www.jstor.org/stable/616128

Schieffelin, B. B., & Doucet, R. C. (1994). The "real" Haitian Creole: Ideology, metalinguistics, and orthographic choice. *American ethnologist*, *21*(1), 176–200. https://doi.org/10.1525/ae.1994.21.1.02a00090

Schulthies, B. (2014). Scripted ideologies: orthographic heterogeneity in online Arabics. *Al-ʿArabiyya*, *47*, 41-56. http://www.jstor.org/stable/24635372

Scribner, S., & Cole, M. (1981). *The Psychology of Literacy*. Harvard University Press.

Sebba, M. (2000). Orthography and ideology: Issues in Sranan spelling. *Linguistics*, *38*(5), 925–948. https://doi.org/10.1515/ling.2000.016

Sebba, M. (2003). Spelling rebellion. In J. Androutsopoulos & A. Georgakopoulou (Eds.), *Discourse constructions of youth identities* (pp. 151–172). Benjamins. https://doi.org/10.1075/pbns.110.09seb

Sebba, M. (2007). *Spelling and society: The culture and politics of orthography around the world*. Cambridge University Press. https://doi.org/10.1017/CBO9780511486739

Sebba, M. (2009). Sociolinguistic approaches to writing systems research. *Writing Systems Research*, *1*(1), 35–49. https://doi.org/10.1093/wsr/wsp002

Seifart, F. (2006). Orthography development. In J. Ulrike (Ed.), *Essentials of language documentation* (pp. 275–299). Walter de Gruyter. https://doi.org/10.1515/9783110197730.275

Seurat, M. (1985). *Le Quartier de Bab Tebbané à Tripoli*. Presses de l'Ifpo. https://doi.org/10.4000/books.ifpo.3415

Seuren, P.A.M. (1982). De spellingproblematiek in Suriname: een inleiding. *Oso, 1*(1), 71–79. https://www.dbnl.org/tekst/_oso001198201_01/_oso001198201_01.pdf

Shaaban, K., & Ghaith, G. (2002). University students' perceptions of the ethnolinguistic vitality of Arabic, French and English in Lebanon. *Journal of Sociolinguistics*, *6*(4), 557–574. https://doi.org/10.1111/1467-9481.00201

Shaw, P. (2008). Spelling, accent and identity in computer-mediated communication. *English Today*, *24*(2), 42–49. https://doi.org/10.1017/S0266078408000199

Siebenhaar, B. (2006). Code choice and code-switching in Swiss-German Internet Relay Chatrooms. *Journal of Sociolinguistics*, *10*(4), 481–506. https://doi.org/10.1111/j.1467-9841.2006.00289.x

Soltan, U. (2007). *On formal feature licensing in minimalism: Aspects of standard Arabic morphosyntax.* [Unpublished doctoral dissertation]. University of Maryland. http://hdl.handle.net/1903/7581

Srage, N. (1997) *Etude sociolinguistique du parler arabe de Moussaytbé (Beyrouth)*. Publications de l'Université Libanaise. https://books.openedition.org/ifpo/121

Stebbins, T. (2001). Emergent spelling patterns in Sm'algyax (Tsimshian, British Columbia). *Written Language & Literacy*, *4*(2), 163–194. https://doi.org/10.1075/wll.4.2.03ste

Street, B. (1984). *Literacy in Theory and practice*. Cambridge University Press.

Su, H. Y. (2007). The multilingual and multiorthographic Taiwan-based Internet: creative uses of writing systems on college-affiliated BBSS. *Journal of Computer-Mediated Communication, 9*(1). https://doi.org/10.1111/j.1083-6101.2003.tb00357.x

Tagg, C. (2015). *Exploring digital communication: Language in action*. Routledge.

Taki, M. (2010). *Bloggers and the blogosphere in Lebanon & Syria: Meanings and activities*. [Unpublished doctoral dissertation]. University of Westminster. https://westminsterresearch.westminster.ac.uk/item/9082q/bloggers-and-the-blogosphere-in-lebanon-syria-meanings-and-activities

Themistocleous, C. (2009, August). Written Cypriot Greek in online chat: Usage and attitudes. In M. Rosetto et al (Eds.), *Proceedings of the 8th International Conference on Greek Linguistics* (Vol. 30, pp. 473–488). University of Ioannina.

Themistocleous, C. (2010a). Writing in a non-standard Greek variety: Romanized Cypriot Greek in online chat. *Writing Systems Research*, *2*(2), 155–168. https://doi.org/10.1093/wsr/wsq008

Themistocleous, C. (2010b). Online orthographies. In R. Taiwo (Ed.), *Handbook of Research on Discourse Behavior and Digital Communication: Language Structures and Social Interaction* (pp. 318–334). IGI Global. http://doi.org/10.4018/978-1-61520-773-2.ch020

Thomas, W. P., & Collier, V. P. (2002). *A national study of school effectiveness for language minority students' long-term academic achievement*. Center for Research on Education, Diversity and Excellence. https://eric.ed.gov/?id=ed475048

Thomason, S. G. & Kaufman, T. (1988) *Language contact, creolization and genetic linguistics*. University of California Press. https://doi.org/10.1525/9780520912793

Tseliga, T. (2007). It's all Greeklish to me!": Linguistic and sociocultural perspectives on Roman-alphabeted Greek in asynchronous computer-mediated communication. In B. Danet & S. C. Herring (Eds.), *The multilingual internet: Language, culture, and communication online* (pp. 116–141). Oxford University Press. http://doi.org/10.1093/acprof:oso/9780195304794.003.0005

Versteegh, K. (1984). *Pidginization and creolization: The case of Arabic*. John Benjamins. https://doi.org/10.1075/cilt.33

Versteegh, K. (2014). *Arabic language*. Edinburgh University Press. https://www.jstor.org/stable/10.3366/j.ctt1g0b09q

Vikør, L. S. (1988). *Perfecting Spelling*. Foris. https://brill.com/view/title/23456

Volk, L. (2009). Martyrs at the margins: the politics of neglect in Lebanon's borderlands. *Middle Eastern Studies, 45*(2), 263–282. https://doi.org/10.1080/00263200802697365

Warschauer, M., El Said, G., & Zohry, A. (2002). Language choice online: Globalization and identity in Egypt. *Journal of Computer-Mediated Communication*, *7*(4). https://doi.org/10.1111/j.1083-6101.2002.tb00157.x

Weth, C., & Juffermans, K. (Eds.). (2018). *The tyranny of writing: Ideologies of the written word*. Bloomsbury Publishing. http://dx.doi.org/10.5040/9781474292436

Yaghan, M. A. (2008). "Arabizi": A contemporary style of Arabic Slang. *Design issues*, *24*(2), 39–52. https://doi.org/10.1162/desi.2008.24.2.39

Zaatari, Z. (2015). Desirable masculinity/femininity and nostalgia of the "anti-modern": Bab el-Hara television series as a site of production. *Sexuality & Culture*, *19*(1), 16–36. https://doi.org/10.1007/s12119-014-9242-5

Zu'bi, A. (2019). Aramaic Substrate in the Arabic Dialects of Kufr-Kanna and Mišhad. *Journal of Semitic Studies, 64*(1), 251–277. https://doi.org/10.1093/jss/fgy052

Appendix

Expanded Tables

Chapter 8

Table 8.5X – Full Range of Word-Sets with 25> Total Tokens - ***Dataset 1***

Total Tok.	<7> Form	Tokens	%	%	Tokens	<h> Form	
6	*a7mad	0	0%	100%	6	*ahmad	*"Ahmad"* [*name*]
7	*se7et	2	29%	71%	5	*sehet	*"The square* [*of*] "
16	*mo7amad	5	31%	69%	11	*mohamad	*"Mohammad"* [*name*]
8	*7dar	3	38%	63%	5	*hdar	*"Watch"* [*imp verb.*]
9	*7ayawen	5	56%	44%	4	*hayawen	*"Animal"*
16	*a7san	9	56%	44%	7	*ahsan	*"Better "*
7	*7ezeb	4	57%	43%	3	*hezeb	*"[Political] party"*
20	*soub7an	12	60%	40%	8	*soubhan	*"Hallowed be"*
8	*wadi7	5	63%	38%	3	*wadih	*"Clear, obvious"*
16	*7elou	10	63%	38%	6	*helou	*"Nice, cool* [m.]"
22	*7ali	14	64%	36%	8	*hali	*"Myself"*
9	*7arb	6	67%	33%	3	*harb	*"War"*
19	*mni7	13	68%	32%	6	*mnih	*"Good* [m.]"
13	*7aj	9	69%	31%	4	*haj	*"Enough"*
13	*sa7eb	9	69%	31%	4	*saheb	*"Pulling"*
14	*ta7et	10	71%	29%	4	*tahet	*"Under"*
11	*7a	8	73%	27%	3	*ha	*"I will, shall"*
12	*we7deh	9	75%	25%	3	*wehdeh	*"One, a person* [f.]"
9	*mbere7	7	78%	22%	2	*mbereh	*"Yesterday"*
14	*yseme7	11	79%	21%	3	*ysemeh	*"He forgives"*
15	*ma7al	12	80%	20%	3	*mahal	*"Place, shop"*
16	*7ata	13	81%	19%	3	*hata	*"Even, even this"*
11	*fata7	9	82%	18%	2	*fatah	*"Opened* [m.]"
11	*7emel	9	82%	18%	2	*hemel	*"Carried* [m. *3rd p.*]"
6	*sle7	5	83%	17%	1	*sleh	*"Arms, weapons"*
12	*ra7me	10	83%	17%	2	*rahme	*"Mercy"*
7	*rte7	6	86%	14%	1	*rteh	*"Rested* [m. *3rd p.*]"
16	*7ayet	14	88%	13%	2	*hayet	*"Life [of]"*
18	*7es	16	89%	11%	2	*hes	*"I feel" / "Feel* [2nd p. imp.]
20	*sa7	18	90%	10%	2	*sah	*"Correct, true"*
11	*ro7	10	91%	9%	1	*roh	*"Soul"*
392	**<7>**	**273**	***70%***	***30%***	**119**	**<h>**	

Table 8.6X - Word-Set Breakdown- *"Protect"* /jəħmi/- **With Translations – *Dataset 1***

<7> Form	Tokens	%	%	Tokens	<h> Form	
ye7mi	**32**	**67%**	**33%**	**16**	**yehmi**	*"Protect him" [m. 3rd p.]*
y7mi	**27**	93%	7%	**2**	yhmi	*"Protect him" [m. 3rd p.]*
ye7me	**2**	67%	33%	**1**	yehme	*"Protect him" [m. 3rd p.]*
y7mik	**2**	100%	0%	**0**	yhmik	*"Protect you" [m. 2nd p.]*
yi7mi	**0**	0%	100%	**2**	yihmi	*"Protect him" [m. 3rd p.]*
y7mikon	**1**	50%	50%	**1**	yhmikon	*"Protect you" [pl. 2nd p.]*
ye7meh	**0**	0%	100%	**1**	yehmeh	*"Protect him" [m. 3rd p.]*
y7me	**0**	0%	100%	**1**	yhme	*"Protect him" [m. 3rd p.]*
i7miyon	**1**	100%	0%	**0**	ihmiyon	*"Protect them" [pl. 3rd p.]*
y7miyun	**0**	0%	100%	**1**	yhmiyun	*"Protect them" [pl. 3rd p.]*
ye7mekkk	**0**	0%	100%	**1**	yehmekkk	*"Protect him" [m. 3rd p.]*
y7miha	**0**	0%	100%	**1**	yhmiha	*"Protect her" [f. 3rd p.]*
y7miki	**1**	100%	0%	**0**	yhmiki	*"Protect you" [f. 2nd p.]*
ye7meyoun	**1**	100%	0%	**0**	yehmeyoun	*"Protect them" [pl. 3rd p.]*
wyi7miyon	**1**	100%	0%	**0**	wyihmiyon	*"And protect them" [pl. 3rd p.]*
wye7mikon	**1**	100%	0%	**0**	wyehmikon	*"And protect you" [pl. 2nd p.]*
by7mo	**1**	100%	0%	**0**	byhmo	*"They protect" [pl. 3rd p.]*
yen7amou	**1**	100%	0%	**0**	yenhamou	*"They would be protected" [pl. 3rd p. subjunctive]*
7amina	**1**	100%	0%	**0**	hamina	*"Is protecting us" [pl. 1st p.]*
y7mikkkkkkkk	**0**	0%	100%	**1**	yhmikkkkkkkk	*"Protect you" [m. 2nd p.]*
***ye7mi**	**72**	***72%***	***28%***	**28**	***yehmi**	

Chapter 9

Table 9.3X – Full Breakdown of Word-Sets containing **/ʃ/** but not **/u/** - ***Dataset 2***

***"Something"* - /ʃi:/**

*shi	86	60	*chi
shi	**84**	**59**	chi
shhi	**2**	**0**	chhi
she	**0**	**1**	che

***"Thing* [f.]" - /ʃaɣle/**

*shaghle	57	38	*chaghle
shaghle	**14**	**11**	chaghle
sha8le	**12**	**8**	cha8le
shaghleh	**11**	**7**	chaghleh
sha8leh	**8**	**4**	cha8leh
sh8le	**3**	**2**	ch8le
sh8leh	**0**	**1**	ch8leh
shghle	**2**	**0**	chghle
shghleh	**0**	**0**	chghleh
sha3'leh	**2**	**0**	cha3'leh
shagle	**2**	**2**	chagle
shagleh	**1**	**0**	chagleh
shaglee	**2**	**0**	chaglee
sha8li	**0**	**2**	cha8li
sh8lh	**0**	**1**	ch8lh

***"Why"* - /leːʃ/**

*lesh	52	29	*lech
lesh	**43**	**21**	lech
leish	**7**	**1**	leich
lsh	**2**	**6**	lch
leshe	**0**	**1**	leche

Table 9.5X – All Words containing **/ħ/**, Arranged by Word Position - ***Dataset 2***

Word Initial

Word	*Tokens*	<7>	<h>	<7>%	<h>%
***7ali**	1	32	16	67%	33%
***7alak**	1	32	17	65%	35%
***7ada**	4	122	73	63%	37%
***7ayeti**	1	31	17	65%	35%
***7abibi**	1	20	28	42%	58%
Total	**8**	**237**	**151**	**61%**	**39%**

Word Medial

Word	Tokens	<7>	<h>	<7>%	<h>%
***ba7er**	2	65	33	66%	34%
***we7ed**	2	56	40	58%	42%
***a7san**	2	58	40	59%	41%
***mni7a**	1	30	18	63%	38%
***re7na**	1	32	17	65%	35%
***e7ki**	1	30	19	61%	39%
***te7ki**	1	31	18	63%	37%
Total	**10**	**302**	**185**	**62%**	**38%**

Word Final

Word	Tokens	<7>	<h>	<7>%	<h>%
***mne7**	1	31	17	65%	35%
***mbere7**	3	81	57	59%	41%
Total	**4**	**112**	**74**	**60%**	**40%**

	Tokens	<7>	<h>	<7>%	<h>%
Total	**22**	**651**	**410**	**61%**	**39%**

1,061

Table 9.17X – All Words Showing Word-Final /e/ - ***Dataset 2***

<*shaghleh>	<*shaghle>	Total Tokens	Tokens per Participant
34	57	91	2
37%	**63%**		

<*ahweh>	<*ahwe>	Total Tokens	Tokens per Participant
13	32	45	1
29%	**71%**		

<*3ayleh>	<*3ayle>	Total Tokens	Tokens per Participant
17	28	45	1
38%	**62%**		

<*salemeh>	<*saleme>	Total Tokens	Tokens per Participant
19	27	46	1
41%	**59%**		

Total

<eh>	<e>	Total Tokens
83	144	227
37%	**63%**	

Experimental Transcripts

Part 1

Presented in Arabic Script:

1. طيب بس خليني شوف إذا فيني إيجي معكن عل بحر
2. شو بدي ضل عم إحكي مع حالي؟ ليش ما حدا عم يرد علي؟
3. كل واحد منكن ساكت شو هل شغله؟
4. خير انشالله شو في؟ مبارح رحنا عالقهوه ما كان في شي
5. غير هيك شو ألأخبار؟ انشالله اليوم أحسن من مبارح؟ والعيله مناح؟
6. بالهنا انشالله حبيبي، نشوفك بخير و سلامة انشالله

Transliteration[25]

1. **Tayeb bas khalini shuf iza fini eji ma3kon 3al ba7er**
2. **Shu bedi dal 3am e7ki ma3 7ali? Lesh ma 7ada 3am yred 3layi ?**
3. **Kel we7ed menkon seket shu hal shaghle?**
4. **Kher nshallah shu fi? Mbere7 re7na 3al ahwe ma ken fi shi**
5. **Gher hek shoul akhbar? Nshallah lyom a7san mn mbere7? Wl 3ayle mne7?**
6. **Bl hana nshallah habibi nshufak bkher w saleme nshallah**

Approximate Pronunciation [IPA]

1. tˤaj.jəb bas xal.li:ni ʃu:f iza fi:ni əʒi maʕkon ʕal baħər
2. ʃu: bəd.di dal ʕam əħki maʕ ħali: le:ʃ ma ħada ʕam jrəd ʕlaj.ji
3. kəl we:ħəd mənkon se:kət ʃu: hal ʃaɣle
4. xe:r ənʃa:l.la ʃu: fi mbe:rəħ rəħna ʕal ahwe: ma ke:n fi ʃi
5. ɣe:r he:k ʃul axba:r ənʃa:l.la ljo:m aħsan mən mbe:rəħ wəl ʕajle mne:ħ
6. bəl hana ənʃa:l.la ħabi:bi nʃu:fak bxe:r wsale:me ənʃa:l.la

Translation

1. Alright just let me see if I can come down to the sea with you guys
2. What, am I going to just talk to myself? Why isn't anyone replying to me?
3. Each one of you is quiet, what's this?
4. What's going on, nothing bad God-willing? Yesterday we went to the café and there was nothing going on
5. Other than that, what's the news? God-willing, today's better than yesterday? And are the family well?
6. Enjoy my dear, we'll see you in good health, God-willing

[25] Transliterated using the highest-frequency resolution for each phoneme, as observed within the thesis.

Part 2

Presented as an Oral Recording in My Voice [IPA]

1. bəd.dak təʒi ʕal baħər wal.la laːʔ ʃəfli wrəd.dəl.li xabar
2. ʕam təħki maʕi wal.la maʕ ħaːlak leːʃ ma ħada ʕam jrəd ʕleːk
3. kəl ma bʃuːf weːhəd mən.non bənʔoz ʃuː hal ʃaɣleː
4. ʃəbak mbeːrəħ bətʔəl.li ʃiː wljoːm bətʔəl.li ʃiː teːni
5. ʔe mniːha aħsan mən baleːha xeːr ənʃaːl.la
6. ana bhal ʔəj.jem aːʕəd bəl beːt kəl ħajeːti ma bʃuːf ħada wala ħada biʃuːfni

Transliteration

1. Bedak teji 3al ba7er wala la2? Shefli w redeli khabar
2. 3am te7ki ma3i wala ma3 7alak? Lesh ma 7ada 3am yred 3lek?
3. Kel ma bshuf we7ed menon ben2oz, shu hal shaghle?
4. Shebak mbere7 bet2eli shi w lyom bet2eli shi ghayru
5. Eh mni7a a7san mn baleha kher nshallah
6. Ana bhal iyem a3ed bl bet kl 7ayeti la bshuf hada wala 7ada bishufni

Translation

1. Do you want to come down to the sea or not? Check and let me know
2. Are you talking to me or to yourself? Why's no-one replying to you?
3. Every time I see one of them I get startled, what's up with this!
4. What's wrong with you, yesterday you told me one thing and today you tell me something else
5. Well that's good, better than nothing, nothing bad, God-willing!
6. These days I'm sat at home all my life, I don't see anyone and no-one sees me