Arabic, English, or 3arabizi?

Code and script choice within discussion forums on a Jordanian website

Robert Michael Bianchi
B.A., M.A. (York University, Canada)

PhD Thesis

August 2011
## Table of Contents

### Chapter 1: Introduction

1.1 Overview 17

1.2 Statement of purpose and research questions 17

1.3 A Personal Interest: Comparing and Contrasting Arabic and English 18

1.4 Central concepts 28

#### 1.4.1 Discourse and discourse analysis 28

#### 1.4.2 Identity 29

#### 1.4.3 Discourse analysis of identity 32

#### 1.4.4 Codes and scripts 33

#### 1.4.5 Code-switching vs. Script-switching 38

#### 1.4.6 Code-switching and identity 42

#### 1.4.7 Orthography and spelling 43

#### 1.4.8 Orthography and identity 44

#### 1.4.9 Corpus linguistics and corpus analysis 45

#### 1.4.10 Corpus-driven or corpus-based research? 46

#### 1.4.11 CMC-related terminology 47

1.5 The Present Research Context 48

1.6 Language types explored in this study 52

1.7 Potential value of the research 53

1.8 Outline of the thesis 56
3.3.1 Method for developing the classification system 101
3.3.2 Scripts and orthographies on mahjoob.com 107
3.3.3 The 14 Linguistic codes 109
3.3.3.1 Arabic-scripted Arabic 114
3.3.3.2 Latin-scripted English 115
3.3.3.3 Latin-scripted Arabic 117
3.3.3.4 Arabic-scripted English 119
3.3.3.5 Bисcriptal codes 120
3.3.3.6 Non-scriptal codes 122
3.4 Annotating the corpus 122
3.4.1 Meta-data 122
3.5 Summary 130

Chapter 4: Method of analysis 131
4.1 Overview 131
4.2 Corpus approaches to analysis 131
4.2.1 CMC and corpora 132
4.2.2 Code-switching and corpora 132
4.2.3 Discourse analysis and corpora 133
4.3 WordSmith Tools 5.0 134
4.3.1 Concordances 135
4.3.2 Frequency wordlists 136
4.3.3 Keywords 138
4.4 Statistical Package for the Social Sciences (SPSS) Version 17.0 139
4.5 Investigating code distribution patterns across the corpus

4.5.1 Converting the meta-data into SPSS format
4.5.2 Determining code distribution across text type
4.5.3 Conflating the 14 linguistic codes in the corpus
4.5.4 Determining code distribution across overarching forum topic
4.5.5 Determining code distribution across thread length
4.5.6 Determining code distribution across author posting frequency

4.6 Topical content and stylistic features of the main linguistic codes

4.6.1 Compiling frequency wordlists for the main codes in the corpus
4.6.2 100-line concordance analysis

4.7 Identity construction through strategic use of linguistic codes

4.8 Summary

Chapter 5: The distribution of codes in the corpus

5.1 Overview
5.2. Overall code distribution of the 14 linguistic codes in the corpus

5.2.1 Overall code trends in the corpus across thread titles, seed messages, following messages and quoted messages
5.2.2 Thread title code occurrence across the corpus
5.2.3 Seed message code frequencies across the corpus
5.2.4 Following message code frequencies across the entire corpus
5.2.5 Quoted message code frequencies across the corpus
5.2.6 Conflating biscriptal codes into one code, Code 4 (Biscriptal content-only)

5.3 Code frequency by overarching forum topic
5.3.1 Forums grouped into 8 overarching forum topics

5.4 Code use across threads

5.5 Author-related code use

5.6 Conclusions and summary

Chapter 6: Discursive tendencies of the main codes on Mahjoob.com

6.1 Overview

6.2 Some methodological issues

   6.2.1 Limitations and other considerations

6.3 The Top 10 lexical words in the wordlists of Codes 1, 2, and 3

   6.3.1 Code 1 (Arabic-scripted Arabic)

   6.3.2 Code 2 (BNC English)

   6.3.3 Code 3 (Arithmographemic Latin-scripted Arabic with English)

6.4 Conclusions and summary

Chapter 7: Qualitative results and discussion

7.1 Overview

7.2 Identity via code-switching, script-switching, and orthography

7.3 The relevance of earlier findings in this study

7.4 Structure of the analysis

7.5 Selection of the thread for analysis

7.6 Messages 1-75 in Thread 206940, “masha2a allah, masha2a allah”

   7.6.1 Anti-establishment authors vs. Pro-establishment authors

   7.6.2 Anti-establishment author: Kharooft Tayeh
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.6.3 Pro-establishment author: Snipe_aac</td>
<td>280</td>
</tr>
<tr>
<td>7.6.4 Pro-establishment author: Muslim4</td>
<td>289</td>
</tr>
<tr>
<td>7.6.5 Anti-establishment author: Guillotine</td>
<td>296</td>
</tr>
<tr>
<td>7.7 Conclusions and summary</td>
<td>302</td>
</tr>
</tbody>
</table>

**Chapter 8: Conclusions**

<table>
<thead>
<tr>
<th>Subsection</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>8.1 Overview</td>
<td>308</td>
</tr>
<tr>
<td>8.2 Summary of research findings</td>
<td>308</td>
</tr>
<tr>
<td>8.3 Reflections on the thesis</td>
<td>315</td>
</tr>
<tr>
<td>8.4 Implications of the research and future directions</td>
<td>317</td>
</tr>
<tr>
<td>8.5 Concluding comments</td>
<td>319</td>
</tr>
</tbody>
</table>

**References**

321
List of Figures

Figure 3.1: Screen shot of the mahjoob.com English website main page ................................................................. 89
Figure 3.2: Screenshot of a biscriptal message on mahjoob.com .......................................................... 99
Figure 3.3: Corpus-annotated version of the biscriptal message shown in Figure 3.2 above ...................... 100
Figure 3.4: Thread 210028, “Song Name” from the mahjoob.com corpus ......................................................... 123
Figure 4.1: A concordance of FRIEND in the E7ke wa Fadfed forum showing Lines 1-19 ............................. 135
Figure 4.2: Frequency wordlist of the Code 3 lexis showing Lines 1-27 ......................................................... 137
Figure 4.3: Keywords of Code 2 using BNC as a reference corpus (Lines 1-27) ............................................. 138
Figure 4.4: Mahjoob.com forum messages in SPSS format showing message cases 1-15 ......................... 143
Figure 4.5: All 21 variables used in the SPSS-formatted corpus including value ranges ........................................ 146
Figure 4.6: The Descriptives function in SPSS ......................................................................................... 149
Figure 4.7: Select Only files containing… function in WS 5.0 .............................................................. 158
Figure 4.8: Select Only sections of text files function in WS 5.0 .............................................................. 159
Figure 5.2: Thread title code percentages overall .................................................................................. 177
Figure 5.3: Seed message code overall percentages ................................................................................. 179
Figure 5.4: Following message code overall percentages ........................................................................ 181
Figure 5.5: Quoted content message code overall percentages .............................................................. 183
Figure 5.6: All 41 forums displayed as main forums and sub-forums ...................................................... 187
Figure 5.7: Following message code distribution among longer threads ............................................... 194
Figure 5.8: Following message code distribution among shorter threads ............................................... 194
Figure 5.9: Following message code choice by topic and thread length .................................................. 196
Figure 5.10: Code distribution of following messages across longer and shorter threads .......................... 199
Figure 5.11: Top 10 posters vs. non-Top 10 posters across topic and code .................................................. 201
Figure 5.12: Top 10 poster vs. non-top 10 poster code use in following messages .................................. 203
Figure 7.1: Following message code distribution among longer threads .................................................. 262
List of Tables

Table 3.1: The 16 codes in the corpus.................................................................110
Table 5.1: The 41 topical forums featured on the English site of mahjoob.com..................................................186
Table 5.2: Eight overarching forum topics and their constituent forums...........................................................190
Table 5.3: The six conflated codes across overarching forum topics.................................................................190
Table 6.1: Top 10 lexical words across Codes 1, 2, and 3..........................................................217
Table 6.2: Top 10 lexical words in the Code 1 sub-corpus................................................................................220
Table 6.3: Code 1 Top 10 lexical words showing topical and linguistic features.............................................222
Table 6.4: Top 10 lexical words in the Code 2 sub-corpus.................................................................................232
Table 6.5: Code 2 Top 10 lexical words showing topical and stylistic features ..................................................234
Table 6.6: Top 10 Lexical Nouns in the Code 3 Sub-Corpus ..............................................................................239
Table 6.7: Code 3 Top 10 lexical words showing topical and stylistic features ..................................................241
Table 7.1: Code use among the authors within the first 75 messages of Thread 206940 ......................................270
Table 7.2 Code use among Thread 206940 authors across the entire corpus ....................................................270
Table 7.3: Four main authors in Thread 206940 contrasted.............................................................................304
List of Appendices

(see attached CD-ROM)

Appendix A – permission to use mahjoob.com for data collection

Appendix B – mahjoob.com corpus (forum folders containing discussion thread text files)

Appendix C – Arithmographemic and Non-arithmographemic item wordlist files

Appendix D – list of smileys file

Appendix E – SPSS-related files

Appendix F – Code 1, Code 2, Code 3, and Joke Zone forum wordlist files

Appendix G – Messages 1-75 in Thread 206940

Appendix H – Code 1 top 10 lexical word concordances

Appendix I – Code 2 top 10 lexical word concordances

Appendix J – Code 3 top 10 lexical word concordances
Abstract of Thesis

entitled
Arabic, English, or 3arabizi?
Code and script choice within discussion forums on a Jordanian website

by

Robert Michael Bianchi
B.A., M.A. (York University, Canada)

For the degree of Doctor of Philosophy
At the Department of Linguistics and English Language,
Lancaster University

August 2011

This study investigates code choice, code-switching, and script-switching involving English and Arabic and the Latin and Arabic scripts within discussion forums within the English-language section of a Jordanian website. The research questions sought to: 1) define the various code and script pairings i.e. the distinct linguistic codes available on the website, 2) establish their distribution across text types, forum topics, thread length, text types, and author posting frequency, 3) determine the distinctive topical, stylistic and discursive functions of the most prevalent codes, and 4) analyse specific instances of identity-creation among forum contributors via strategic use of these codes.

A purposive sample of discussion threads, representing roughly 10% of the English website forum content was downloaded and annotated into a WordSmith 5.0-searchable corpus, with a secondary version created for SPSS-analysis. A subset of 75 messages from a linguistically-atypical thread was also investigated using discourse analysis methods to explore identity-related language use.
The findings question the notion that English is the preferred language of the internet since Latin-scripted Arabic intermixed with English and Arabic-scripted Arabic were more prevalent than English-only messages in the corpus derived from the English-language website. They also demonstrate that vernacular Arabic has become a written variety in CMC, challenging the primacy of Modern Standard Arabic as the only form of written Arabic. Topically, Arabic-scripted Arabic occurred most often within Poetry, Humour, and Local Culture forums, Latin-scripted English within professional and academic forums, and Latin-scripted Arabic within General Discussion and Hobby-related forums. Arabic-scripted Arabic was used most for third-person, narrative type discourse, while Latin-scripted Arabic and Latin-scripted English were used most for involved, first-person and second-person expressive and phatic discourse. English was also used for taboo topics such as religious dissent, sex, and sexual orientation. Discourse analysis demonstrated that a purposive sample of authors were able to create distinct identities through strategic code use, reflecting political biases with pro-establishment authors selecting formal language forms especially Modern Standard Arabic and Salafist-style English while anti-establishment authors preferred informal ones such as vernacular Arithmographemic Latin-scripted Arabic with English.
Acknowledgements

Given an undertaking as time-consuming and multi-faceted as a doctoral thesis, it is difficult to adequately express the gratitude properly owed to literally dozens of individuals and numerous institutions. First of all, I would like to thank Mr. Emad Al-Hajjaj for his generous agreement to use his website, mahjoob.com for my research. Next, I need to mention my various doctoral supervisors: Paul Baker, for his painstaking and unstinting effort and valuable support in putting the pieces of a very disjointed thesis together. Sebastian Hoffmann, deserves extra special thanks for his herculanean effort of having built and tested the corpus from scratch with me over several months at various times of day and night, spending literally hundreds of hours, patiently putting up with my incessant demands for perfection in response to which he would often remind me of a saying in his native German: “The impossible is no problem. Miracles, however, take a little longer”. Mark Sebba taught me to become a resourceful and independent researcher. Others at Lancaster deserving of credit are: Costas Gabrielatos for providing moral and technical advice, Jane Sunderland for acting as sounding-board during my many fits of frustration, Paul Rayson for help with Arabic corpora, Elaine Heron, and Anne Stubbs for providing logistical support and flexibility. Then, there are my many Lancaster classmates and friends: Ali AlBulushi, Ken Lee, Li-Chin Chen, Laurel Kamada, Hyeon Oak, and Somchoen Honsa for offering valuable tips and moral support during the process; Kate Power for great encouragement and advice; Rob Taferner for the many chats; and Al Ghamdi and Sarah Mercer for inspiring me. In the field, many thanks go out to David Palfreyman and Buthaina Al Share for their many leads and helpful suggestions. I also thank the folks at the UAE University for supporting and accommodating my studies at the very beginning.
At former workplace, College of the North Atlantic-Qatar, I humbly thank the Executive administration for their generous financial and logistical support of my studies, especially Dr. Hal Jorch and Ms. Enid Strickland. In this regard, my eternal appreciation extends to my former supervisor and mentor, Eileen Bragg. Without her tremendous support and encouragement, I would not have been able to continue this work. Similarly, I thank my research assistants, Ali Hussein Zreik, Sarah Odeh, and Rouba Abou-Amer for their steadfast support and exceptional work. Rolene Pryor provided many hours of SPSS-related training and guidance. Both Cyndy Howman and Roger Kenworthy served as role-models and supporters. The IT support team at CNA-Q also proved unmatchable in terms of their support for my software and hardware-related questions, requests, and demands. At my current workplace, Virginia Commonwealth University in Qatar, I thank my director, Dr. Byrad Yyelland, the IT support team, and especially Molly McHarg for her invaluable proof-reading and moral support.

I also wish to simultaneously thank and apologize to all my family and friends. My prolonged absences may have strained our relationships but they never dimmed my love. I express my gratitude to my family in Italy, Japan, and Canada, especially my mother, my father, my sisters, brothers-in-law, and nieces who had to put up with my forays to libraries while I should have spent more time with them while on vacation. Above all, I wish to thank the dearest and most beloved person in my life, my angel, Machiko who has given so much and deserves so much more than I can ever possibly give her. I can only hope that time will give me the opportunity to repay in however small a measure the immense sacrifices that Machiko made to help see this work through. On that note, I am also indebted to our sweetest daughters, Clarice Akari and Stella Hikari for the many precious and irreplaceable childhood summers and weekends they were forced to endure without Daddy. And finally, though I break with more recent secular
traditions in academia, I thank Bahá’u’lláh, my Higher Power, the guiding spiritual force and motivator in my life to Whom I address these words from a favourite prayer: “Thou art more friend to me than I am to myself, I dedicate myself to Thee, O Lord”. Despite all errors and inconsistencies, for which I am solely responsible, this work is humbly dedicated to Him.
Declaration

I declare that this thesis represents my own work, except where due acknowledgement is made, and that it has not been previously included in a thesis or report submitted to this University or to any institution for a degree, diploma, or other qualification.

Signed

Robert M. Bianchi
Chapter 1: Introduction

1.1 Overview
As the title of this thesis implies, the present study represents an investigation of three forms of language: Arabic, English, and 3arabizi and of the code-switching between them. But what exactly is "3arabizi"? And why does it contain the numeral '3' as its first letter? Essentially, 3arabizi (pronounced [ʕarabizi]) or Arabish as it is commonly known in English-speaking circles, refers to a stylized and mixed code of vernacular Arabic and English written in Latin script, containing numerals used to represent sounds (e.g. the Arabic voiced pharyngeal affricate [ʕ] represented by the ‘3’ in 3arabizi), which have no corresponding graphemes in written English. The focus of this chapter, then, is to outline the purpose of the research and the central research questions in relation to these three main forms of language. My personal interest in Arabic-English code and script-switching will also be discussed. Additionally, some relevant terminology and concepts are presented, including the notions of discourse, identity, orthography, code-switching, script-switching, and corpus. After this, the research context of computer-mediated communication (CMC)-based code-switching and script-switching is outlined and the objects of the study are briefly described. This is followed by a discussion of the potential value of the research to the field of sociolinguistics. At the end of the chapter, an outline of the entire thesis is provided.

1.2 Statement of purpose and research questions
This study focuses on language alternation in computer-mediated communication (CMC) as evidenced in a 37,000,000-word corpus of web forum messages downloaded from a popular Arabic-English website, mahjoob.com. Two distinct types of linguistic behaviour are fore-
grounded in this study: 1) code-switching between Arabic and English and 2) script-switching between the Latin and Arabic scripts.

Both quantitative and qualitative methods of corpus investigation and discourse analysis (see Section 1.5 below) are employed to address the following research questions:

1) What are the various types of code and script-switching featured on the English website of Mahjoob.com?

2) How are script-code pairings distributed across certain types of postings in terms of
   a) text type (title, first message, following messages, quoted content)
   b) overarching forum topic
   c) thread length
   d) author posting frequency

3) What do frequent lexical items in the main script-code pairings in the corpus reveal about the topical content and stylistic and discursive functions of those pairings?

4) How do forum posters use the available script-code pairings to construct identities?

In the remainder of this section, in order to provide further background and context to the study, my choice of Arabic and English as objects of research, the central concepts I refer to, the impact of CMC on language, and the potential value of this research for the field of linguistics are presented. Briefly, though, I will explain the reason for my interest in this topic.

1.3 A Personal Interest: Comparing and Contrasting Arabic and English

First, it is important to explain exactly what is meant by the terms Arabic and English here. In its widest sense, the word “Arabic” is used as an umbrella term referring to several different written and spoken varieties which are mutually intelligible to varying degrees. At one end of the
spectrum, Arabic refers to Classical Arabic and the closely related Modern Standard Arabic which form the basis of written and educated spoken Arabic (see Holes, 2004, p. 5). Classical Arabic is usually described as the language of the Holy Qur’an due to the fact that it was codified according to usage in the Qur’an (Katzner, 2002, p. 155). Holes observes that the Qur’anic variety itself is a 7th century AD form of written Arabic related to the pre-Islamic poetic language of the Hijaz region of Western Arabia (Holes, 2004, p. 16). Modern Standard Arabic is essentially the same language with slight modifications which betray the lexical and phraseological evolution and enrichment of written Arabic over the past 14 centuries (Holes, 2004, pp. 5-6). In addition, certain colloquialisms and grammatical simplifications are noticeable in Modern Standard Arabic. Both Classical and Modern Standard Arabic are closely identified with written Arabic, the former being employed for religious texts and treatises, and the latter for virtually all other written communication. However, both Classical Arabic and Modern Standard Arabic also find spoken expression: Classical Arabic is the language of Qur’anic recitation and Islamic liturgy. Modern Standard Arabic is the normative language of formal speech featured in news broadcasts, university lectures, public debates, and poetry.

---

1 Holes notes that among Arabs there is no distinction made between Classical Arabic and Modern Standard Arabic, both of which are referred to simply as al-‘arabiyyah al-fusHa (eloquent Arabic) and which contrast with al-‘arabiyyah ‘ammiyyah (popular Arabic), the spoken vernacular in its many varieties as found across the Arabic-speaking world. Holes points out that Western linguists have found it convenient to distinguish Classical Arabic from Modern Standard Arabic due to the emergence of the latter as a modernized and standardized version of the former from the mid-19th century onwards (Holes, 2004, pp. 4-5).

2 Holes (2004) devotes an entire chapter to this discussion, pointing out that this tendency cannot be taken too literally and that a body of actual linguistic data points to a much less tidy delineation between standard and vernacular in everyday practice. Indeed, Holes observes that style-shifting between standard and vernacular forms of
At the other end of the spectrum, Arabic encompasses an enormous number of additional varieties of vernacular Arabic which are national, regional, social (city-dweller vs. country-dweller vs. nomad) and communal (Druze vs. Christian vs. Jewish vs. Muslim) (McLoughlin, 1982). Broadly, these vernacular varieties are classified nationally and regionally as Levantine Arabic (Jordanian, Lebanese, Palestinian, Syrian), Egyptian Arabic, North African Arabic (Libyan, Tunisian, Algerian, Moroccan, and Mauritanian), Sudanese Arabic, Gulf/Eastern Arabic (Saudi Arabian, Kuwait, Iraqi, Qatari, Bahraini, Emirati, and Omani), and Yemeni Arabic. In the data used for this thesis, the most commonly encountered variant is Levantine Vernacular Arabic, especially Jordanian Vernacular Arabic (see Al Share, 2005 for a discussion of male and female variants of this). This is due to the fact that the data come from mahjoob.com, a Jordan-based website (see Chapter 2).

In terms of English, the variety of English encountered in this study ranges from formal standard written English to Netspeak (see Crystal, 2001). However, it should be noted that the linguistic distance between written Standard English and informal English is not as wide as that between written Modern Standard Arabic and Vernacular Arabic. Consequently, English will generally be discussed as a single code although salient orthographic and stylistic features when encountered will be examined as well (see BNC English and non-BNC English in Chapters 3 and 6 below).

My decision to investigate these two languages and scripts in contact was motivated by numerous factors. On a personal level, I have been a student of Arabic for twenty years and have Arabic is very common within formal spoken contexts and seems to be increasing within written contexts (Holes, 2004, pp. 341-382).
lived in the Arabian Gulf since 2000. My first exposure to Latin-scripted Arabic occurred in an EFL classroom in the UAE four years ago when I noticed a student composing a text-message during my lesson. The message contained a string of words composed in Latin script with several numerals interspersed among the usual Latin letters. Clearly, this was neither English nor Arabic but something in between. When I was younger, I would compose English messages to myself transliterated into Arabic script as a means of consolidating my own Arabic script fluency. However, I wrote these messages simply as a language learning aide whereas this student wrote Latin-scripted messages for an audience who were able to respond using the same code. Unlike my own isolated literary output, this student’s text represented an actual register of language shared by a real discourse community of Latin-scripted Arabic users.

This encounter with Latin-scripted Arabic exemplifies the long and interesting relationship between Arabic and English. Historically, Arabic has contributed several words to the English language (Katzner, 2002). At present, the trend has reversed as English words such as ‘Internet’ and ‘mobile’ have entered the spoken and written domains of Arabic. The two languages also share historical and sociolinguistic similarities. Both are the languages of historical empires (Katzner, 2002) and have long, well-developed literary traditions which originate around the second half of the first millennium CE. Presently, Arabic and English enjoy prestige status as two of the six UN official languages, each with hundreds of millions of native speakers. Similarly, Arabic and English are official, co-official, or national languages in a number of countries. And both continue to function as languages of wider communication or lingua francas.

---

3 Katzner (2002, p. 155) cites several English words of Arabic derivation such as alchemy, alcohol, alcove, algebra, amber, artichoke, coffee, cotton, elixir, gazelle, giraffe, magazine, mattress, sofa, syrup, tariff, and zero.
with several hundred more millions of second language users (Wardhaugh, 1992:56; Crystal, 1997:9). Finally, like English, the existence of several vernacular Arabics, which contrast in diglossic⁴ fashion with standardized written and formal spoken forms, is well-established (see Brustad, 2004).

There are several notable differences between Arabic and English. First, English and Arabic hail from two unrelated language families: Indo-European and Afro-Asiatic respectively. Consequently, they have widely different structural and phonological properties⁵. In terms of written language, while both Arabic’s current script and English’s current Latin script ultimately descend from the same Phoenician progenitor alphabet, their present forms are now widely divergent. Besides the obvious visual differences between the scripts and the fact that, in direct contrast to English, Arabic is written cursively from right to left, Arabic is also considered to be an Abjad writing system in which consonants are the principal graphemes with vowels generally holding a secondary or optional role (Daniels, 1996)⁶. On the other hand, English features an alphabetic writing system in which vowels play an integral part.

---

⁴ The adjective derived from the noun, diglossia. Ferguson (1959) first introduced the term diglossia to describe settings in which two varieties of the same language co-exist. In the case of the Arab Muslim world, the use of Classical Arabic as a liturgical language, of Modern Standard Arabic as the written code and spoken code in formal contexts, and of vernacular Arabic as the language of everyday informal spoken communication qualifies as diglossia.

⁵ Although syntactically, modern-day spoken varieties of Arabic tend to share Subject Verb Object patterns in common with English (see Brustad, 2004).

⁶ A couple of other points of comparison between the Arabic and Latin scripts: They are both phonographic systems in that they use characters or graphs to represent phonemes, which are combined to form lexical items (Coulmas,
Phonologically, Modern Standard Arabic and Standard English share a majority of common phonemes: [k], [b], [f], [t], [d], [s], [z], [θ], [ð], [ʃ], [ʒ], [dʒ], [m], [n], [l], [w], [j], and [h]. In addition, some English phonemes which do not occur in Modern Standard Arabic are commonly found in several Arabic vernaculars e.g. [g], [tʃ], and [ʒ] (see Palfreyman & Al Khalil, 2003). The phoneme /r/ is articulated differently in both languages, yet speakers of both languages readily recognize it as the same phoneme (Scottish English [r] to Standard English [ɹ]).

Despite these similarities, the two phonologies exhibit considerable differences. The Arabic sound system includes the uvular fricatives /ʁ/ and /χ/, the pharyngeal fricatives /ʕ/ and /ħ/, and the emphatic plosives /sˤ/, /zˤ/, /tˤ/, and /dˤ/ which do not occur in Standard English phonology. Similarly, English phonology includes phonemes such as /p/ and /v/ which are not encountered in either Modern Standard Arabic or the vernaculars. It is this apparent lack of phonological correspondences that leads Latin-scripted Arabic users to employ certain graphs of the ASCII character set such as numerals as graphemes (see 2.4.2 below).

---

1996; Rogers, 2005). The two scripts also contain a similar number of graphs. The Latin script as used in English contains 26 graphs. The Arabic script has a total of 28 graphs. Nevertheless, in addition to letter shape and directionality, i.e. right-to-left vs. left-to-right, the Arabic and Latin scripts are also different with regards to cursivity (connected vs. unconnected graphs), and allography (initial vs. medial vs. final forms of graphs and capitalization) respectively.

7 Note that for the purposes of transliteration, the uvular fricatives /ʁ/ and /χ/ will be written with the digraphs ‘gh’ and ‘kh’ respectively. The pharyngeal fricative /ʕ/ will be written with the reverse apostrophe “ ’ ” while the voiceless fricative /h/ and the emphatic plosives /sˤ/, /zˤ/, /tˤ/, and /dˤ/ will all be written using the capital letters ‘H’, ‘S’, ‘Z’, ‘T’, and ‘D’ respectively.
Concerning script, some Latin and Arabic “cognate” graphs bear slight resemblances owing to their common origins (cf. “q” and “ق”, “L” and “ل”). However, over time, the cursive nature of the Arabic script led to the simplification of many of its graphs. For instance, the Arabic graph Alif “ا” no longer resembles its Latin cognate graph “A”.

As noted above, the Arabic script, like Hebrew, has often been categorized as an abjad writing system as opposed to an alphabet (Daniels & Bright, 1996; Rogers, 2005). This refers to the fact that its orthography places priority on the writing of consonants and semi-consonantal long vowels rather than on short vowels (Wickens, 1980). Indeed, outside of Qur’anic texts, children’s books, and Arabic language learning texts, short vowels are almost never written unless this is deemed necessary in order to resolve ambiguities in meaning. Even when short vowels are employed they are written diacritically either above or below a consonantal or long vowel graph, emphasizing that they do not exist as independent graphs. Thus, when reading Arabic texts collocational and contextual cues are vital to the correct understanding of an unwovelled wordform. For instance, the unwovelled wordform “كتبت”/k-t-b-t/ has at least six potential semantic (and phonological) realizations depending on context: 1) “I read” /katabtu/, 2) “you (masc. sing.) read” /katabta/, 3) “you (fem. sing.) read” /katabti/, 4) “she read” /katabat/, 5) “it (fem.) was read” /kutibat/, and 6) “they (inanimate) were read” /kutibat/. If this principle were followed in English, the simple spelling “bt” could be taken to refer to any of the following: “bat”, “bet”, “bit”, “bot”, or “but”.

---

8 This reflects the lexical tendency of Afro-Asiatic languages to form semantic meanings around tri-consonantal clusters, where vowels are inserted between the consonants in order to create morphological distinctions (see Rogers, 2005, pp. 140-141).
Besides vowel diacritics, Arabic makes extensive use of superscript and subscript dots in order to give the basic graph shapes contrastive graphemic value (Palfreyman & Al Khalil, 2003). These basic grapheme shapes often denote similarly articulated sounds which, nevertheless, represent distinct sounds. Note the following examples:

1. “س” /s/ vs. “ش” /ʃ/ dental vs. post-alveolar
2. “ت” /t/ vs. “ث” /θ/ plosive vs. fricative
3. “د” /d/ vs. “ذ” /ð/ plosive vs. fricative
4. “ص” /sˤ/ vs. “ض” /ḍˤ/ unvoiced fricative vs. voiced plosive
5. “ط” /tˤ/ vs. “ظ” /ẓˤ/ unvoiced plosive vs. voiced fricative
6. “ع” /ʿ/ vs. “غ” /ʁ/ voiced pharyngeal fricative vs. voiced uvular fricative
7. “خ” /ḥ/ vs. “غ” /ʁ/ unvoiced pharyngeal fricative vs. unvoiced uvular fricative

Another kind of diacritic is the superscript shadda “ّ”. This diacritic is used to indicate the phonological doubling of the consonant over which it is written. Similar to the diacritic vowels, the shadda is not normally written and its presence in a word is discerned by contextual cues alone.

Compared with written Arabic, English is noteworthy for its extensive use of digraphs in order to represent phonemes which do not have one-to-one correspondences with single graphs e.g. English “sh” for /ʃ/. In contrast, Modern Standard Arabic has an almost one-to-one graphemic-phonemic correspondence, i.e. a very close graphological-phonological fit. As such, it is often considered to have a relatively shallow orthography (see Sebba, 2007).
At this point, a word is in order regarding the key differences between Modern Standard Arabic phonology and Vernacular Arabic phonology as these appear to have an impact on the orthography of Latin-scripted Arabic. The close graphological-phonological fit between Modern Standard Arabic phonology and the Arabic script does not apply to most varieties of Vernacular Arabic today, which are normally not written anyway. In many cases, the phonologies of these vernaculars differ considerably from Modern Standard Arabic’s phonological system. One such change is the above-mentioned acquisition of phonemes which were not originally present in Modern Standard Arabic e.g. the /t∫/ common in Gulf Vernacular Arabic and Iraqi Vernacular Arabic (Holes, 1984). Another is the merging of certain Modern Standard Arabic phonemes e.g. /q/ and /ʔ/ into Levantine Vernacular Arabic /ʔ/, or their divergence e.g. Modern Standard Arabic /θ/ into either vernacular /s/ or /t/ in Levantine Vernacular Arabic (McLoughlin, 1982). Finally, some phonemes such as intervocalic /h/ and /ʔ/ have been lost altogether in certain vernaculars.

On a cultural note, one other vital area of difference between Arabic and English is their associations with Islam and ‘Western’ modernity respectively. Thus, each language carries symbolic value and is identified with specific domains of use. This is illustrated by the fact that, in countries such as the UAE and Qatar, Islamic sciences such as Shariah law and history are taught in Arabic while all subjects which are considered “modern” i.e., Business, IT, Engineering, etc. are usually taught in English.

Linguistic differences notwithstanding, there are several social domains and geographic areas in which Arabic and English overlap. This helps explain why Arabic-English bilingualism, allowing for varying degrees of proficiency, is quite common, particularly in the Arab world. For instance, English is widely taught throughout the Arabic-speaking world, partly due to former
British colonialism in countries and territories like Egypt, Sudan, Jordan, Palestine, Iraq, and the Gulf States and partly due to strategic economic, educational, and military alliances with English-speaking countries such as the US. But even in former French colonies such as Morocco and Lebanon, English has been displacing French as the preferred foreign language to be studied in schools (Crystal, 1997). Intensification of trade and tourism between Arabic and English-speaking countries has also engendered more Arabic-English code-switching. Add to this the fact that there are sizeable Arabic-speaking minorities present in virtually all the English-speaking “inner circle countries” (Crystal, 2003, p. 60) and it becomes obvious that Arabic speakers have several opportunities to develop fluency in English. Conversely, English speakers who have an interest in or connection to Islam often learn Arabic for its religious value. And since the US has become more intensively involved in Middle Eastern affairs in the wake of 9/11 and continues to monitor Islamist threats, the study of Arabic has been deemed as essential to US policy and security as was the study of Russian in the Cold War era. Accordingly, American Arabic language programs have seen a tremendous growth in enrolment in recent years.

Now that the principal similarities and differences between Arabic and English have been discussed, it will be profitable to consider central concepts encountered in this study.

---

9 Kachru (1985) divides the countries of the world into three distinct categories based on their relationship with the English language: Inner Circle, Outer Circle, and Extending/Expanding Circle countries. Inner circle countries are countries in which English is the primary language such as Australia or the UK. Outer Circle countries, in contrast, are countries in which English has special status as a second language or co-official language such as in India or Malaysia. Finally, Extending/Expanding Circle countries are those which have no colonial ties with Britain but who recognize the importance of English as a global lingua franca. This category comprises countries such as China, Japan, and Portugal (Crystal, 2003, p. 60).
1.4 Central concepts

In light of the above research questions, the key constructs, concepts and terms which are employed in this study will now be defined for the purposes of this thesis. Specifically, these are discourse, discourse analysis, identity, codes, code-switching, script-switching, and orthography. Subsequently, the phenomena of CMC-based code-switching and script-switching and how they impact on the research will be discussed.

1.4.1 Discourse and discourse analysis

Discourse analysis has become increasingly popular in a variety of fields beyond traditional linguistics, especially in the social sciences (Wetherell, Taylor, & Yates, 2001) as evidenced by a relatively recent series of discourse analysis how-to books written by linguists for researchers in other disciplines wanting to look more closely at language in use (Baker, 2006; Fairclough, 2003; Schiffrin, Tannen, & Hamilton, 2001; Wetherell et al., 2001). Baker (2006) points out that the term ‘discourse’ is employed in several different ways across different fields of inquiry. Fairclough (2003, pp. 3-4) observes that like ‘language’, discourse can be used as either a non-count noun in a generic, abstract sense, or as a count-noun to refer to specific discourses. When used in the former abstract sense, Jaworski and Coupland define discourse as “language use relative to social, political and cultural formations – it is language reflecting social order but also language shaping social order, and shaping individuals’ interaction with society” (1999, p. 3). When used in a countable fashion, Fairclough says that “a discourse is a particular way of representing some part of the (physical, social, psychological) world” and that “there are alternative and often competing discourses, associated with different groups of people in different social positions” (2003, p. 17). In this thesis, discourse is employed primarily in the abstract sense and instances of language use among text-producers are examined in order to trace
how distinct identities are constructed through text-based interactions. Consequently, for the present study, discourse analysis involves examining the texts of web forum participants over stretches of interactions in order to locate instances of identity-creation through the strategic use of distinct linguistic forms whether lexical, grammatical, or orthographic (see Chapter 4 for more detail). Since identity is also central to the study, a brief description of this term as used here is warranted, followed by a discussion of its relationship to the other key concepts of the research.

1.4.2 Identity

Despite its recent rise to prominence in the human and social sciences, the notion of identity, like discourse, remains a contested word in the literature (du Gay, Evans, & Redman, 2000), and it has been defined differently by various fields of inquiry (see Fairclough, 2003:160; Meierkord, 2004). To Labov, identity represented a complex of discrete features such as age, gender, social class, ethnicity, etc. which fall generally outside the influence of the language user and could thus be objectively determined (Hinrichs, 2005). Challenging this variable-defined view of identity, LePage and Tabouret-Keller (1985) foregrounded the notion of identity as a kind of performance in which languages can be used to project different social aspects. In parallel fashion, Ochs (1987) introduced the notion of indexicality whereby specific features of language help users to reference distinct social styles. Butler (1990) also views identity as performative, noting that individuals copy behaviours (including language) from others. It is the repetition of these behaviours which gives the appearance of a stable and fixed identity. For most of the time, the uptake of such behaviours is not conscious and governed by what society decrees is acceptable. These more recent concepts have effectively countered the somewhat binary and static categorization of identity in the earlier Labovian approach, highlighting the notion of language users as potentially creative agents who construct their own identities. Indeed, since
their introduction, these concepts have in fact become mainstays of sociolinguistic approaches to identity.

Carrying this focus on user-created identity further, Rampton (2002) describes what he labels as *crossing* where individuals appropriate the language and style of members outside their own social groups to project different social personae which would not normally be considered typical of their own in-group. Bell provides a similar framework in his use of dialect and its role in creating social personae according to his audience theory (Smedley, 2006). Bell’s view parallels, to an extent, Bakhtinian notions of heteroglossia and voice (see Smedley, 2006). Within Bakhtin’s framework, all discourse, even monological discourse in a single code can potentially incorporate different voices which, in turn, index recognizably distinct social personae (Smedley, 2006).

Wodak et al. (1999, p. 11) point out that although the idea of identity as a static and enduring entity has been successfully challenged by several theorists, identity is still often reified, losing its analytic power as a result. Instead, Wodak et al. view identity as part of a complex and dynamic process (1999, p. 11). Fairclough highlights an important distinction between theorists who uphold this dynamic view of identity as a *process* and those who view identity as an *outcome*. Fairclough refers to the process as ‘identification’, and its outcome as ‘identity’, which is constructed discoursally i.e. linguistically (2003, p. 160). Fairclough further distinguishes between social identity and personal identity or personality both of which, he argues, are dialectically related, but not reducible, to each other. For Fairclough, however, identity construction requires another element: self-consciousness (ibid.). In his view, self-consciousness
is itself a pre-requisite to agency which allows individuals to assume distinct social roles which then translate into more complex social identities. A key concept in Fairclough’s approach, taken from MacIntyre (1984), is the notion that each culture possesses a set of ‘characters’ or character types which Fairclough describes as “its culturally most salient identities” (2003, p. 161). He cites some examples of characters in a culture: ‘Manager’, ‘Politician’, ‘Therapist’, etc. Although not exempt from change, Fairclough argues that these character types are “pervasive through social life” as well as relatively durable and recognizable over time (2003, p. 161).

Consequently, individuals as social agents are able to construct social identities through a process of identification with these characters. The process of identification entails some sort of discursive (i.e. linguistic) action on the part of the individual. Central to this linguistic action is the Faircloughian notion of *styles*, which he describes as “ways of being” (2003, p. 159). In linguistic terms, this means that an individual selects (consciously or otherwise) from among a range of linguistic features or subsystems such as lexis, grammar, and phonology in order to construct a distinct identity. It is worth noting that, in the present study, while lexis and grammar are both evident in the data, orthography replaces phonology as a salient linguistic feature since the data consist entirely of written texts.

Thus, for the purposes of the present research, identity is defined as a social construct observable in the form of distinct social personae which users cultivate via strategic use of linguistically-realized styles (see Fairclough, 2003:162). It should be noted, however, that the choice to use a particular code to project a particular social persona may not always be fully conscious to the user. Nevertheless, the impact of this choice will be readily discernible to the audience. Given that this strategic use of codes may well occur without the meta-cognitive awareness of the user,
this research prioritizes the value of the text itself as the primary evidence of the social self-construction and projection work of the user. Consequently, data from questionnaires and surveys administered to actual users, while potentially valuable for yielding the type of demographic data sought in the Labovian approach to identity, may actually obscure the underlying purposes of the strategic deployment of various codes in collected textual data that exhibits code-switching and script-switching. In light of this, it is maintained here that discourse analysis of specific texts without recourse to the original text-producers can still provide sufficient evidence to help delineate the range of locally-relevant stereotypes and styles created and projected by the users in question (see Smedley, 2006; Hinrichs, 2005). Additionally, Baker (2006) notes that, in the field of discourse analysis, recourse to actual text-producers is often not possible nor always desirable when it is possible, and that ethical considerations may preclude this option. In light of such methodological and ethical issues (see Chapter 4), as well as due to issues of space, it was decided not to contact the text producers directly to seek their own understandings of their linguistic behaviour.

1.4.3 Discourse analysis of identity
Pavlenko and Blackledge’s (2004) collection of articles on identities in multilingual contexts provides an overview of potential frameworks for studying identity performance as it relates to linguistic acts. In their introduction to the volume, Pavlenko and Blackledge outline sociopsychological, interactional sociolinguistic, and poststructuralist approaches to researching identity negotiation. They also offer a set of perspectives on potential definitions of identity and its relationships to discourse, ideology, power relations, multiplicity, imagination, and narratives. Of special relevance to this study is their observation that code-switching is a principal means of identity negotiation. The current research also corroborates their position that the “invention and
use of new linguistic varieties” is another way, beyond code-switching in the traditional sense, that identity is negotiated linguistically (Pavlenko & Blackledge, 2004, p. 23).

Putting this theory into practice, Hinrichs (2005) demonstrates the value of discourse analysis of identity-related code-switching involving CMC-based Jamaican Creole and English by identifying the socially recognizable and locally-relevant stereotypes which Jamaican Creole users choose to index through their varied use of the linguistic codes available to them (see also Su, 2007). These stereotypes are akin to Fairclough’s notion of characters (2003:174) such as ‘citizens’ and ‘experts’ (ibid: 184). In common with this study, since these recent studies have focused on discourse analysis of identity-related code-switching, it is now pertinent to define what code-switching means in the present research context, starting with the basic concepts of codes and scripts.

**1.4.4 Codes and scripts**

For the purposes of the present study, the notions *code* and *script* are employed distinctly. *Code*, a term generally applied in studies of spoken code-switching to refer to a distinct linguistic variety or style (see Wardhaugh, 1992, p. 103), is taken to mean the actual lexicogrammatical system in which meaning and content is conveyed (see Halliday, 1985; Morley, 2000). In contrast, *script* refers to the writing system employed to compose messages in a given code (Coulmas, 1996).

The mixed and non-standard character of the linguistic data on mahjoob.com is largely the result of the optional pairing or combination of several distinct yet dialectically-related linguistic subsystems, i.e. lexic grammars, scripts, and orthographies (see Coulmas, 1996, p. 379; 454). Essentially, mahjoob.com authors combine the distinct scripts, orthographies, and
lexicogrmars available to them in unconventional ways to produce texts that are markedly different from the kinds of texts usually encountered in other domains and contexts such as in academia or in print media. Indeed, one of the most intriguing aspects of the present research is that it highlights that the conventional relationships or pairings of subsystems such as lexis, grammar, and orthography are not altogether fixed and cannot be taken for granted. Therefore, it is argued here that it is precisely the ability to combine and adapt these subsystems to produce unconventional linguistic pairings (i.e. new written codes) that enables forum participants to create distinctly recognizable and locally salient online identities (see Chapter 7).

Based on prior observations of the English website of mahjoob.com, the following three separate yet combinable and adaptable linguistic subsystems have been identified on the website:

1. Writing systems (Scripts)
2. Orthographic systems (Orthographies)
3. Lexico-grammatical systems (Lexicogrmars)

Each of these subsystems will now be described briefly. In Chapter 3, a deeper discussion is offered illustrating how these subsystems are combined in both conventional and unconventional ways to produce the discrete linguistic varieties observable on the website.

Rogers defines writing as “the use of graphic marks to represent specific linguistic utterances” (2005, p. 2). In a similar vein, Coulmas describes a writing system as “a set of visible or tactile signs used to represent units of language in a systematic way” (Coulmas, 1996, p. 560). Joshi and Aaron (2006) provide four different categories of writing systems along with examples of languages which use these: 1) Logographic (Chinese), 2) Syllabic (Japanese Kana), 3)
Alphabetic-syllabic (Kannada and Tamil), and 4) Alphabetic (Italian and Spanish) \(^\text{10}\). For analytical purposes in this study, writing systems and \textit{scripts} are considered one and the same and will be used interchangeably, though it should be noted that Coulmas draws a distinction between these two, arguing that scripts are graphic “instantiations” of writing systems (see Coulmas, 1996, p. 454).

In contrast to a writing system, an orthographic system refers to the more or less fixed set of spelling conventions applied to a specific language (Coulmas, 1996, pp. 379-380). Thus, based on Coulmas’ definition of writing systems above, here a writing system is distinguished from an orthographic system in so far as the latter refers to the set of conventional rules by which a writing system is employed systematically to graphically represent a given lexicogrammatical system \(^\text{11}\), or simply a ‘lexicogrammar’ (Halliday, 1985). A lexicogrammar, then, is essentially

\(^{10}\) This four-way classification is not universally agreed upon, however. For instance, some researchers have used the terms \textit{abjad} and \textit{abugida} instead of \textit{alphabet} to describe what Joshi and Aaron call \textit{alphabetic-syllabic} (see also Daniels & Bright, 1996; Rogers, 2005). The former refers to writing systems which favour the use of consonants, making little or no use of vowels. Abugidas, on the other hand, employ both consonants and vowels, but write the vowels diacritically as a kind of subscript either above or below consonants (see the \textit{Brahmi} script-based Kannada and Tamil mentioned above). In light of these concepts, the Arabic and Hebrew scripts are generally defined as examples of abjads because of their users’ preponderant tendencies to use consonants and suppress short vowels in writing Arabic and Hebrew. I argue that such conventional notational practices are a function of orthography as opposed to any inherent limits of the writing systems themselves (see Footnote 12 below).

\(^{11}\) In this research, the term lexico-grammatical system or lexicogrammar encompasses the complex of linguistic subsystems i.e. lexis, morphology, and syntax which characterize a specific language system (Halliday, 1985). In my research, I highlight the difference of lexico-grammatical systems from both the writing systems and orthographic systems which are necessarily employed when text-producers represent lexico-grammatical systems graphically. As
the linguistic lexis and structural system which text producers strive to represent in written form by using a specific script and orthography (see Morley, 2000, pp. 6-8)\(^\text{12}\).

---

the data will show, the relationships between script, orthography, and lexicogrammar are not rigidly fixed on mahjoob.com.

\(^\text{12}\) An example will hopefully suffice to clarify the relationship between scripts, orthographies, and lexicograms. Consider that the Hebrew script, like the Arabic script, has been used to transcribe several different codes over the course of its 3,000-year-old history (Coulmas, 1996; Strolovitch, 2005). Nevertheless, the method of applying the Hebrew script to represent sounds and words has varied considerably from one lexicogrammar to another as well as from one time period to another. In other words, while the writing system i.e. the Hebrew script has remained relatively unchanged, the orthographic systems for spelling these different codes using the Hebrew script have varied considerably. Among these Hebrew-scripted lexicograms are several Semitic, neo-Latin, and Germanic languages and vernaculars. Probably the two most famous examples of these Hebrew-scripted languages in the West are Yiddish and Ladino (also known as Judezmo). Note that in both Yiddish and Ladino, long and short vowels are conventionally written out in full in the Hebrew script. However, in both Ancient and Modern Hebrew, as in Modern Standard Arabic, only long vowels are conventionally written while short vowels are inferred by contextual cues surrounding the written form indicating a marked preference for consonantalism (see Cross & Freedman, 1952). In Arabic, a similar orthographic practice occurs when transcribing foreign words, especially brand names. So, for example, the Japanese brand name ‘Suzuki’ in Arabic script is rendered “سوزوكي” (SuzukiSaudia, 2009). Transliterated into Latin script, this would be equivalent to writing ‘suuzuukii’. However, if the same word were interpreted to be a native Arabic word, based on its popular pronunciation in Arabic, it would most likely be transliterated “سزوكي” (‘szuukii’) or perhaps even “سزكي” (‘szkii’) with all short vowels omitted. The writing of full vowels in the transliteration marks the brand name as foreign. From this example, it is evident that script as a graphical-notational system of graphs differs from the orthographic system. The latter system determines how the graphs of the script are conventionally used to represent a lexico-grammatical system.
Since scripts themselves can readily be distinguished from each other, these are viewed as one of the two most conspicuous types of contrastive linguistic variables on the website, the other being the actual linguistic codes which are the main objects of investigation here (see Chapter 3). In the present study then, the Arabic and Latin scripts are the two main writing system variables which forum participants have at their disposal to choose from. However, the ability to choose linguistic variables does not end with scripts. Forum participants are also able to select between lexicogrmmars as well. For instance, they can choose to write using lexical and grammatical forms from Standard English, Classical Arabic, Modern Standard Arabic, or from various forms of Vernacular Arabic. Consequently, text producers also have to choose between distinct lexicogrammatical system variables. A further choice is apparent in terms of orthography. Text producers must also choose which set of spelling conventions to use when composing a text. For example, they can select a relatively standardized orthography which would concur with ‘dictionary’ forms, or they can employ a less standardized orthography such as one which features truncated Netspeak forms (see Crystal, 2001), arithmographemes (see Bianchi, 2005, see also Chapter 3), or doubled vowels in words such as “Islaam” (see Mujahid, 2009) (see Chapter 7). These separate sets of spelling conventions represent choices between distinct orthographic system variables.

The outcomes of all these linguistic choices are the various linguistic combinations or configurations found on the website i.e. distinct written varieties or codes which form the basis for the linguistic analysis in this study. At a broad level, therefore, a code can be defined as the pairing of a script with a lexicogrammar. Such codes can then be further categorized into sub-varieties in terms of their distinctive orthographic features (see Chapters 3, 6, and 7). Codes, of
course, function as linguistic variables in their own right (see Milroy, 1987a). This is because codes can also be selected and contrasted with one another, enabling text producers to index different sociolinguistic styles and identities (see Bailey, 2007, pp. 257-272).

1.4.5 Code-switching vs. Script-switching

In the present research, the primary focus is on code-switching as opposed to script-switching, that is, the analysis concentrates mainly on the occurrence of the codes themselves as well as on the different stylistic effects which result from switching between them and mixing them. There are a couple of reasons for this. For one, for the sake of analytical clarity, it is useful to limit the number of variables to investigate. Indeed, a focus on codes provides a manageable number of clearly distinct types of language to work with. Also, given my view of scripts, lexicogrmmas, and orthographies as being subordinate to, and constituent of, codes, it makes more sense to analyse these within the broader category of ‘code’. In light of this, I view script-switching as a function of code-switching, and when it occurs in the data, it will be taken to signal an occurrence of the latter. Nevertheless, in the analyses, script, orthography, and lexicogrammatical choices will be considered and discussed separately from code choices where these appear to be particularly marked and salient (see Chapter 7).

13 It should be noted that one systematic form of script-switching evidenced in the data is the use of the Latin script (along with some numerals) to transcribe Arabic lexicogrammatical content. Another type of script-switching in the data is the converse phenomenon of using Arabic script to transcribe English lexicogrammatical content. It may be more appropriate to call the latinization of Arabic and the arabicization of English script borrowing (Tranter, 2001, pp. 180-204) or script adaptation (Strolovitch, 2005, pp. 1-2), or perhaps even script shifting. The widespread latinization of Arabic on Mahjoob.com is taken to reflect the existence of a separate linguistic code, Latin-scripted Arabic which is stylistically distinct from Arabic-scripted Arabic (see Chapter 3 for a discussion, also see Chapters 6 and 7).
At this point, I need to acknowledge that my use of the term ‘code’ here may be somewhat unorthodox from a linguistic point of view. Code has been used as a neutral word to describe a linguistic variety, whether written or spoken, in contrast to the word ‘language’ which is often considered loaded (see Milroy's notion of "linguistic variable", 1987b, p. 11). It is used most obviously in expressions such as code-switching and code-mixing. However, what can be said when a switch in writing systems or scripts occurs? Does the pairing of a script with the lexicogrammar of a specific language with which it is not normally associated entail the creation of a new code (i.e. language), or should this be considered simply the same code written with a different script (see Coulmas’ 1996 definition of digraphia)? Clearly, from a sociolinguistic perspective, changes in scripts have often been used to highlight cultural, ideological, and national differences between groups of people who otherwise might be said to share a common language such as in the case of Hindi and Urdu (see Collin, 2005; Coulmas, 1996, pp. 129-130; Sebba, 2007). For the sake of analytical clarity in this study, I use the term ‘code’ descriptively to refer to the pairing of a lexicogrammar with a script, reserving judgement on whether such script-switching actually signals the existence of a whole new language. Of greater importance, perhaps, are the perceptions of the language users regarding the nature and status of the codes they use (see Auer, 1998, p. 9). For instance, if Latin-scripted Arabic users routinely refer to this code as 3arabizi (lit. ‘Arabish’) (cf. Sakarna's 'Englo-Arabic', 2006, see also Chapter 2) and seem to consider it to be separate from Arabic and English, its putative linguistic forebears, then there may be legitimate grounds to recognize it as socially distinct from say, Arabic-scripted Arabic (see Wei, 2007, pp. 8-11).
The concepts of code choice, code-switching, script-switching, code-mixing, and, to a lesser extent, script-mixing are central to the present study. Each will be briefly defined now. Code choice reflects a decision on the part of a text-producer to use one or more of the numerous linguistic codes in currency on the website in a specific interactional context (see Wardhaugh, 1992, p. 103). In this connection, script choice is considered to be a sub-type of code choice because it ipso facto signals the use of a new code. A case of salient code choice would be an author’s decision to use a code different from his or her interlocutor(s) despite the fact that she or he obviously understands and might even be able to use the same code (see Chapter 7). Another example of salient code choice is when posters use only one code in one thread or forum but use other codes in other threads or forums (also Chapter 7).

Heller (2007) notes that the term ‘code-switching’ to conceptualize all the possible types of alternation between two or more distinguishable linguistic varieties in a given context is not uniformly agreed upon and has consequently been defined variously in the literature (Heller, 2007, p. 7). In this report, I initially align myself with Gumperz’s formulation of code-switching as “the juxtaposition within the same speech exchange of passages of speech belonging to two different grammatical systems or subsystems (1982, p. 59), though I duly note Auer’s reservations that such a view betrays a monolingual bias towards code-switching which tends to see only discrete linguistic codes and ignore the existence of ‘fused lects’ as a by-product of intensive code-switching (see Auer, 1998; Auer, 2008). And in fact, in Chapter 3 the existence of fused lects in the data will be explored further.\footnote{In Chapter 3, it will be seen that one of the key linguistic variables in this study is actually a mixed code (see McLellan, 2005, p. 7). This comprises Latin-scripted Arabic elements along with Latin-scripted English elements.}
Building on Gumperz’s work, Llamas, Mullany, and Stockwell (2007) provide a useful distinction between code-switching, “when speakers switch between codes within a single interaction”, and code-mixing, “when speakers engage in code-switching within sentences, also known as intrasentential code-switching” (2007, p. 208). In light of these notions, in the context of the present research where writing as opposed to speech is the focus, code-switching will be used to indicate one of two conditions: 1) the text producer’s act of switching i.e. deviating from the use of a specific linguistic code within a given message between sentences or 2) the text producer’s switching of codes between messages within a given discussion thread. In contrast, code-mixing is used to refer to a text producer’s act of using more than one code within a given sentence (see Llamas et al., 2007, p. 208; McLellan, 2005, p. 7) though in practice, sentence boundaries in CMC-based utterances can be open to different interpretations. Thus, even borrowings of single lexical items are considered bona fide examples of code-mixing here.

Having discussed code-switching, script-switching, and code-mixing, it is important to note that script-mixing is conceived of in two ways in this study. Whereas code-mixing involves the use of lexical items from different codes within a sentence, at one level, script-mixing involves the use of graphs from different scripts within a single lexical item. In the website data, this is evidenced primarily by distinctive author IDs and signature lines which may feature Cyrillic script graphs alongside Latin script graphs. A secondary type of script-mixing involves the systematic appropriation of graphs or symbols from one script into another script. In the present research, this is exemplified by the use of numeric symbols found on the QWERTY keyboard as

15 Although, in the present study, this phenomenon appears to be extremely rare and highly marked (see Tranter (2007) for a discussion of commonplace script-mixing in Japanese).
graphemes to represent Arabic sounds which have no ready equivalents in English written in the Latin script. An example of such are the use of “7” to represent the Arabic grapheme “ح” or “3” for “ع” (see Chapter 3 for more detail). This process is referred to as arithmographemicization and examples of it are labelled arithmographemes (Bianchi, 2005).

1.4.6 Code-switching and identity
As evidenced above, within sociolinguistics there is a relatively long and rich history of code-switching research. The pivotal theories of Gumperz, Myers-Scotton, and Auer have already been touched upon. At this point, though, it is worth noting that Gumperz’s early conceptualization of code-switching as either situational or metaphorical had placed code-switching firmly within the realm of social interaction and oral discourse, thus overlooking the phenomenon of written code-switching (Hinrichs, 2005). And within this divide between situational and metaphorical code-switching, it is the latter which implies the idea that code choice may be semi-conscious and strategic and that codes can in and of themselves carry differentiated symbolic and representational meanings for their users (see Och’s (1987) notion of ‘indexicality’).

In contrast to Gumperz and informed by the Chomskyan formalist tradition, McLellan observes that Myers-Scotton has sought to uncover putative code-switching universals and common constraints which would help account for where and when code-switching can and cannot take place in any given sociolinguistic situation (McLellan, 2005). But as with both Gumperz and Auer, Myers-Scotton’s focus has been primarily on spoken discourse. What these theorists have in common is that their approaches have emphasized either the structural or contextual elements of code-switching, while code-switching itself is viewed as largely if not wholly unconscious and
autonomous. However, there has been a recent trend in sociolinguistics to redress this research bias focusing on spoken data. This new trend has begun to investigate more conscious types of code-switching, especially as evidenced in written language. Written code-switching is argued to be more ideological in nature than spoken code-switching and underpinned by identity-construction motives (see K.M. Lee 2007; Smedley, 2006; Hinrichs, 2005; McLellan, 2005). It is within this newly emerging area that the current study is situated, and it is hoped that the present findings will further contribute to the relatively small but growing body of literature on such identity-related written code-switching.

At this point, it is important to consider the role that orthographical choices play in signalling code switches in the data and supporting identity construction (see Chapter 7). Thus, the terms orthography and spelling as used here will now be discussed.

1.4.7 Orthography and spelling

Coulmas acknowledges that orthography and spelling are often used interchangeably, and even his own definitions of these terms appear to overlap when he says that orthography is “a normative selection of the possibilities of a script for writing a particular language in a uniform and standardized way” while spelling is “the conventions which determine how the graphemes of a writing system are used to write a language” (Coulmas, 1996, pp. 379, 477). Sebba (2007) provides a more useful distinction between these terms using orthography as a label for the overall system of writing a given language and spelling when referring to specific cases of graphic representations of words. Comparing these two theoretical approaches, Coulmas takes a more descriptive linguistic stance towards orthography and spelling while Sebba presents a case for the sociolinguistic study of orthography and spelling. In the present research, Sebba’s
distinction will be applied and the term orthography is used to describe a relatively fixed system of writing words (standard or otherwise) while spelling is used to refer to specific instances of orthography. Also, in alignment with Sebba (2007:56), it is argued here that orthography and spelling, like other aspects of discourse (see Wodak et al. 1999; Fairclough, 2003) are social practices and that, in acts of spelling, especially systematically unconventional ones, acts of identity can often be discerned. As samples of unconventional spelling are examined here along with unconventional pairings of language and writing systems, a guiding question will be how these linguistic, i.e. orthographic, choices enable web forum contributors to construct different identities (see Pavlenko and Blackledge, 2004:22-24).

1.4.8 Orthography and identity
Sebba (2007) points out that, despite appearances to the contrary, orthographical choices can never really be considered autonomous or neutral (see Ferguson, 1959, p. 235). In other words, spelling is always ideological at some level as it reflects the value systems and biases of text composers. By extension, in this work the view is taken that script choice should also be considered a type of orthographical choice, and as a sociolinguistic behaviour, it is also ideologically-motivated. Regarding orthography itself, this work concurs with Sebba that “[i]t touches on matters of social identity, national identity, cultural politics, representation and voice” (2007:6)

In conjunction with code-switching, therefore, the orthographic repertoires of the mahjoob.com participants are considered resources for identity-negotiation and construction (see Pavlenko and Blackledge, 2004:22-24). As mentioned in Chapter 4 below, this understanding has motivated the decision to distinguish between two principal orthographically-distinct varieties of Latin-
scripted Arabic within the corpus: 1) Arithmographemic Latin-scripted Arabic (labelled “Code 3”) and 2) Non-arithmographemic Latin-scripted Arabic (labelled “Code 10”). As will be seen in Chapter 7, specific authors are noteworthy for their decisions to either use Code 3 or its non-arithmographemic counterpart, Code 10, reflecting decisions which seem to underscore identity-creation processes when code use is contrasted within a selected discussion thread.

Now that these conceptual terms have been defined, it will be useful to discuss some of the key methodological terms, such as corpus analysis, which are employed in this research in order to investigate the kinds of code and script-switching found in the data.

1.4.9 Corpus linguistics and corpus analysis
McEnery and Wilson provide a basic definition of corpus linguistics as “the study of language based on ‘real life’ language use” (2001, pp. 1-2). In corpus linguistics, this real life language data is contained in a corpus (pl. corpora), a collection of texts, generally representative of a specific language variety or register, and stored in machine-readable format in an electronic database (Baker, Hardie, & McEnery, 2006, pp. 48-49). McEnery and Wilson (2001) emphasize that in contrast to branches of linguistics such as syntax and morphology, corpus linguistics is a methodology. Providing further detail, Biber, Conrad, and Reppen (1998) stipulate four basic methodological principles in corpus-based analyses: 1) empirically examining patterns in naturally-occurring language, 2) carrying out large systematic collection of texts i.e. a corpus as data, 3) using computers for both automatic and interactive analyses, and 4) resorting to both quantitative and qualitative methods (1998, pp. 4-5). Since this research is also concerned with patterns of use in naturally-occurring language, a corpus linguistics methodology has been employed for both data collection and analyses in the hope of approaching the topic more
systematically and rigorously (see McEnery and Wilson, 2001:126). A primary corpus technique used in this study is concordance analysis. Briefly, a concordance is a list of all occurrences of a word or cluster of words found in a corpus within their immediate textual contexts. Consequently, a concordance enables a researcher to search for specific words of interest and to examine the contexts in which they appear. Since identity construction among specific web forum posters is examined here, concordances were used extensively in order to locate texts written by these posters within the corpus.

1.4.10 Corpus-driven or corpus-based research?

Tognini-Bonelli (2001) makes a valuable distinction between “corpus-driven” and “corpus-based” research. The first approach focuses on the corpus itself as the main language sample from which hypotheses are then inductively drawn. In this way, corpus-driven research is akin to grounded theory. Corpus-based approaches, in contrast, start with assumptions about language that are then investigated deductively using corpora to either confirm or disprove initial hypotheses. My approach is essentially corpus-based as my first research question illustrates: I begin the study with the concept of code in mind and then seek to establish what codes are distinguishable on the website. Once identified, these codes are then examined using corpus methods in order to determine their occurrence and use patterns in line with the subsequent research questions.

In order to carry out analysis of the textual data, I make use of the corpus analysis software package WordSmith version 5.0. This application allows researchers to load corpora and generate frequency wordlists, concordances, keyword lists, and collocations for these texts (more on these terms in Chapter 4). Of particular importance to this study is the fact that these
applications also allow for simultaneous display of items in several scripts including the Latin and Arabic scripts.

1.4.11 CMC-related terminology

A note is in order about the use of CMC terminology. To date, a number of terms are not fully agreed upon. For the sake of consistency, the virtually synonymous terms message board, discussion board, and web forum will all be referred to as ‘web forum’ or simply ‘forum’ (Claridge, 2007:87, 101; Hoffmann, 2007). Similarly, I simplify the term weblog by using its more commonplace counterpart ‘blog’. As for the units of analysis, the terms message, thread, and author each warrant a brief description. Message refers to a single delimited text posted to a discussion thread within a given forum. The terms message, post, and posting will all be used interchangeably here. I draw a distinction, however, between a seminal message or seed message (i.e. the first message to occur within a given thread) and subsequent messages or following messages (see Wodak & Wright, 2007) (see Chapter 4 for details). Messages are typically not found in isolation. Rather, each message is posted to a specific discussion thread, a series of two or more messages. In principle, these messages are related to the topic of the discussion thread in which they occur as well as to each other. Discussion threads will be referred to simply as ‘threads’ in this study. Finally, the terms author and poster will be used in place of contributor, participant, text-composer, and text-producer to refer to individuals who create messages\(^\text{16}\) and post them to various threads. When considered collectively, the authors on the mahjoob.com website will be referred to as ‘Mahjoobians’ and descriptively as ‘Mahjoobian’, out of deference

\(^{16}\) The term ‘create’ is used loosely here given the fact that several messages in the corpus are clearly the result of copying and pasting written texts from other web sources and have not been composed by the posters who submitted them to the website.
to their own practice since these are epithets which several authors frequently employ to identify themselves and their interlocutors\textsuperscript{17}.

Having explained some of the key terms used in this thesis, I now present information regarding the research context, the use of code-switching and script-switching between Arabic and English in computer-mediated communication.

1.5 The Present Research Context
Modern communications technologies have spread around the globe at a rapid pace. Innovations such as PCs and mobile phones have been introduced to different cultural and geographical areas so quickly that there has not always been sufficient time to adapt them to prevailing local linguistic realities and conventions (Shields, 2000). Yet Herring (1996), Shields (2000), and Crystal (2001) observe that, despite this initial lack of congruence between language of use and language of technology, the linguistic diversity of the planet is, in fact, increasingly being expressed on the Internet and in other electronic media.

One way in which this diversity is being expressed is in the creation of new linguistic norms by users of relatively old and stable languages (Warschauer et al., 2002). For instance, there is a noticeable rise in the script-switching phenomenon of digraphia in CMC environments (Palfreyman and Al Khalil, 2003). Digraphia refers to the sociolinguistic situation in which a single language (or ‘code’) can be written in more than one script (Grivelet, 2001). There are many well-known historical cases of digraphia (e.g. Serbo-Croatian in Cyrillic and Latin scripts called Serbian and Croatian respectively (Katzner, 2002)).

\textsuperscript{17} The website participants also refer to themselves as mahajeeb and mahjoobis, but Mahjoobians appears to be the most frequent of these terms.
In the case of Internet-enabled computers and mobile phones, the most common form of
digraphia is Latinization of a non-Latin-scripted language (see Palfreyman, 2001). Crystal (2001)
observes that the Latin script\(^{18}\) was thrust upon technology-adopters for whom this script was not
their principal orthographical system due to the limitations imposed by earlier character encoding
systems\(^{19}\) in computing. This situation created a dilemma for non-Latin-scripted language users.
They could either forego the new technologies or use a Latin script language such as English for
their CMC even among same language peers.

But for those who were either unable or unwilling to use a Latin script language such as English,
another strategy for using CMC technology became apparent: to appropriate the Latin script for
composing and reading messages in their respective first language or (L1). As a result, several
Latin script-based 'makeshift' orthographies have come into existence such as Latin-scripted
Greek (see Koutsogiannis and Mitsikopoulou, 2003), Latin-scripted Japanese (Nishimura, 2003),
and Latin-scripted Arabic (Warschauer et al., 2002; Palfreyman and Al Khalil, 2003). Thus,

---

\(^{18}\) The Latin script (also called the Roman script) was developed by ancient Romans to write Latin and is currently
used not only to write most of the languages of Europe (i.e. Romance, Celtic, Germanic, Baltic and some Slavic
languages), but also for several non-Indo-European languages as a result of colonization, conquest, and cultural and
economic domination by Western European nations. Consequently, it is arguably the most widely used script in the
world today.

\(^{19}\) Character encoding systems enable computers to process alphabetic data by assigning each alphabetic character a
unique numeric value. Early character encoding systems such as ASCII (see Note 20, below) represented a relatively
limited number of characters, usually only Latin script and therefore it was initially extremely difficult to produce,
send and receive CMC texts in other orthographic systems such as Arabic or Chinese.
CMC-based digraphia seems to have been spear-headed by these non-Latin-scripted language users as a creative way to circumvent technological constraints on script choice.

In electronic contexts, Palfreyman and Al Khalil (2003) refer to this process of Latinization as ‘ASCIIization’ in reference to the formerly widespread use of the ASCII (American Standard Code for Information Interchange) character encoding system20 which offers its users only unmodified Latin characters, numerals, and punctuation marks for composing, sending, and receiving written text. One focus of their research is ASCIIized Arabic which I will refer to as Latin-scripted Arabic throughout the remainder of this study21.

More recently, however, the apparent necessity for Latinization in CMC has been challenged by enhancements in technology (Palfreyman and Al Khalil, 2003; Androutsopoulos, 2006). Crystal (2001) observes that more inclusive character encoding systems such as Unicode22 now exist and are becoming more widely available. Still, there is evidence of the continued use of Latin script

___________________________

20 The ASCII character encoding system was developed in the 1960s. ASCII comprises a total of 128 codes, only 95 of which can be used for graphical characters as compared to 1,000,000 characters for Unicode (see Note 22 below). Because of its graphical limitations, ASCII has obliged CMC users to replace or ignore several of the distinct graphical features of their languages (Danet and Herring, 2007:8-11).

21 I prefer the term “Latin-scripted Arabic” or even “Latinized Arabic” over “ASCIIized Arabic” in order to highlight that, despite its ASCII origins, this form of writing Arabic in Latin script is no longer necessarily conditioned by responses to ASCII-imposed technological constraints. In other words, it appears to be an unconstrained stylistic choice on the part of many of its users at present.

22 According to Unicode.org (2009), “The Unicode Standard is a character coding system designed to support the worldwide interchange, processing, and display of the written texts of the diverse languages and technical disciplines of the modern world. In addition, it supports classical and historical texts of many written languages.”
for non-Latin-scripted languages even though many early technical limitations have now largely been overcome (Palfreyman and Al Khalil, 2003; Al Share, 2005). Indeed, continued Latinization poses some interesting questions about language choice, language use, and identity which inform the present study. In other words, Latinization which was initially a constraint-driven linguistic choice, has now become a new linguistic resource for its users (Lee, 2007; Pavlenko & Blackledge, 2004).

To this point I have discussed CMC as if it were a monolithic whole. However, it should be noted that despite similarities, there are marked differences in the possible uses (both monologic and dialogic) of the various formats of CMC. Consequently, there are also important differences in the typographical complexity of these types of CMC. Consider, for instance, synchronous CMC applications such as web chat and ICQ which are dialogic in nature and are usually carried out in real-time and thus offer simultaneous interaction between users. Largely due to pressures to complete conversational turns in a timely fashion, users employ a range of acronyms, abbreviations, and truncations in their texts. On the other hand, asynchronous forms of CMC such as e-mail, blogs, and web forums usually allow for longer lapses of time between composing one text and the next. Thus, the apparent need for brevity is mitigated (see Herring, 1996). In linguistic terms, these various types of CMC provide their users with different opportunities and challenges both for production and reception of texts. In the context of the present study, this amounts to greater possibilities for different kinds of script-switching, i.e. Latinization and Arabicization, in web forums.

But why should script-switching occur at all anyway? After all, within a given text, switching between the Arabic keyboard set and the Latin keyboard set normally requires more time and
effort on the part of the text composer. This seems to explain why it is rarely if ever evidenced in synchronous formats such as IRC where time is of the essence (Al Share, 2007; Tseliga, 2007, p. 125). That script-switching happens at all in light of the extra time required to achieve it has led to the present investigation of bilingual and biscriptal web forum postings where script-switching is both a viable linguistic option and a commonly observed phenomenon. Among all the possible types of code-switching and script-switching which can occur in CMC (see Lee, 2007 for CMC-based script-switching between Chinese and English), I narrow my research focus to investigate script-switching involving Arabic and Latin scripts in asynchronous CMC texts. In this study script-switching is viewed as a function of code-switching (see Chapter 3), so the analyses will also necessarily include an investigation of the more general patterns of CMC-based code-switching between Arabic and English. In brief, Arabic and English code-switching and script-switching in CMC texts represent a rich research context of striking linguistic and cultural contrasts between two markedly different codes and scripts.

1.6 Language types explored in this study
My interests in Arabic-English bilingualism, in Latin-scripted Arabic, and in CMC code-switching and script-switching have led me to conduct several pilot studies (see Bianchi, 2005). In accord with grounded theory (Cohen et al. 2000; Silverman 2004; Dey 2004), these previous studies have helped me to formulate and test initial hypotheses regarding Latin-scripted Arabic and script-switching while making it possible for me to discover potential sources of data and to trial methods of data collection and analysis. For instance, my small-scale surveys and interviews focusing on CMC language and script choice have indicated to me that technical constraints, habit, and audience appear to be factors in CMC language and script choice. My corpus analyses of Latin-scripted Arabic features in CMC texts have suggested to me processes of both phonetic
and orthographical/visual borrowing from the English and Arabic writing systems as well as orthographical simplification (see Tseliga, 2007). Frequency counts of Latin-scripted Arabic strings on Arabic free-to-air Satellite TV channels and on Internet web pages indicate that, while relatively rare as phenomenon compared to the CMC options to use either English or Arabic, the use of Latin-scripted Arabic is indeed a popular linguistic option for some Arabic-English bilinguals and biscriptals. Through these studies, I have discovered that although a code may be paired with a specific script, these relationships are not necessarily fixed. As a result, I have found that four main language types (i.e. unique script and code pairings) can and do occur in the English website of mahjoob.com corpus. These are: 1) Arabic-scripted Arabic, 2) Arabic-scripted English, 3) Latin-scripted English, and 4) Latin-scripted Arabic. Examples of each will be given and discussed in Chapter 3.

1.7 Potential value of the research

At this juncture, a few words are in order regarding the potential value of the present research especially to the fields of sociolinguistics and descriptive linguistics. First of all, in sociolinguistic literature, there have only been a few studies carried out on script-switching in CMC, and virtually no studies to date on script-switching between Arabic and Latin scripts. Similarly, there is a conspicuous absence in the literature of studies related to identity creation in writing among Arabic-English bilinguals and biscriptals23 (but see Al Share, 2005). Indeed, such identity-related studies to date have been primarily carried out on users of East Asian logographic languages such as Cantonese (see K-M. Lee, 2007), Taiwanese (see Su, 2003), and Japanese (see Nishimura, 2003).

---

23 “Biscriptals” refers to individuals who have knowledge of at least two different writing systems (see Park and Vaid in Taylor and Olson 1995).
With regards to the field of descriptive linguistics, the actual linguistic types under investigation in this study are of particular research value, especially Latin-scripted Arabic also known as ASCIIized Arabic (Palfreyman and Al Khalil, 2003; Al Share, 2005). Latin-scripted Arabic’s value to descriptive linguistics lies in the fact that it appears not to have existed as a functional written register until the advent of the internet. Thus it may be considered a whole new register of Arabic.

A comprehensive descriptive study of Latin-scripted Arabic is particularly compelling because of several of its novel features such as its common use of numerals as graphemes. Indeed, Latin-scripted Arabic has yet to be systematically classified as a writing system. Warschauer et al.’s (2002) seminal work on Egyptian Arabic online, followed by Palfreyman and Al Khalil’s (2003) work on ASCIIized Arabic helped to bring the topic into the literature while raising some important questions about the nature of this mixed orthography. Positing the potential formative elements of ASCIIized Arabic orthography, such as its borrowing of numerals which resemble Arabic graphemes to represent sounds not found in English, Palfreyman and Al Khalil (2003) provide an intriguing introduction into ASCIIized Arabic grammar. However, my own investigations into Latin-scripted Arabic suggest that it has developed features which are dissimilar to its putative precursor orthographies, namely English (and/or French) and Arabic (see Palfreyman, 2001a; Palfreyman & Al Khalil, 2003). It should be mentioned, though, that such a contrastive analysis remains beyond the scope of the present study.

From a historical point of view, Latin-scripted Arabic’s rise is also somewhat curious in that, unlike many of the spelling reform movements and script battles of the twentieth century (see
Collin, 2005; Sebba, 2007), the Latin-scripted Arabic phenomenon appears to be relatively unplanned, stemming from a popular grassroots response to technological constraints. As mentioned above, whether independently, or due to exposure to Latinization of other non-Latin-scripted languages, Latin-scripted Arabic’s development has mimicked certain orthographic patterns in Latin-scripted Greek (Tseliga, 2007) and Latin-scripted Russian (Palfreyman, 2001a) such as the use of keyboard numerals as graphemes to transcribe sounds not conventionally (or conveniently) represented by the Latin alphabet. Nor is Latin-scripted Arabic overtly politically-motivated in that Latin-scripted Arabic seems to have developed in the absence of political support or mention (Warschauer et al., 2002). And unlike the fierce public debate in Greece engendered by the use of Latin-scripted Greek (Androutsopoulos, 2007), Latin-scripted Arabic seems to have received little press within the Arabic-speaking world.

Another reason to focus on Latin-scripted Arabic is that, as a linguistic phenomenon, Latin-scripted Arabic may yet become extinct as increasing numbers of Arabic users can avail themselves of Arabic script-enabled technology. If Latin-scripted Arabic does indeed die out, valuable insights may well be lost regarding the historical process of “script borrowing” i.e. when a graphological system is passed from one language to another.

This research also foregrounds the broader issues of code-switching, which surround the Latin-scripted Arabic phenomenon, and of script-switching which make Latin-scripted Arabic possible. Inspired by the work of Smedley (2006) on Tagalog-English code-switching in blogs and of McLellan (2005) on Bruneian Malay-English code-switching in web forums, I seek to categorize mixed language written texts in an attempt to discover where and when code-switching occurs.
However, unlike either Smedley or McLellan’s research, where script-switching does not apply, I investigate script-switching as a linguistic affordance in its own right.

Finally, I hope to help remedy the relative lack of research into code-switching and script-switching in online CMC discourse communities as identified by Danet and Herring (2007). In this manner, I would like to contribute in some small way to the wider, but still emerging branch of linguistics which Crystal (2001) labels Applied Internet Linguistics.

1.8 Outline of the thesis
Chapter 2 provides an overview of the literature on diglossia and bilingualism including a discussion of the contexts of the study: Jordan and the Internet. Chapter 3 is a response to the first research question (see section 1.2), providing both a description and classification of the data used in the study. Chapter 4 presents the methods of analysis adopted in order to address the remaining research questions. In Chapter 5, research question 2 is addressed: the distribution of the languages and scripts on the website is presented and discussed in terms of forum and topic. Chapter 6 addresses the third research question, employing corpus methods to detect the dominant topics associated with each of the main linguistic codes (i.e. languages) in the corpus. The final research question is broached in Chapter 7, addressing the issue of whether the languages featured on the website are used strategically to construct different identities. Chapter 8 concludes the thesis and offers further directions for research in the field of bilingual CMC.

1.9 Summary
In this introductory chapter, I discussed the purpose of the research and my three central research questions which imply both quantitative and qualitative investigations of Arabic and English code-switching and script-switching. I then overviewed my interest in the field of Arabic-English bilingualism. Next, I defined the basic terminology and concepts of identity, orthography,
discourse, corpus linguistics, code-switching, and script-switching which underpin the study. I also highlighted the role of CMC in fostering code-switching and script-switching in the present research context. Following this, I identified Arabic-scripted Arabic, Latin-scripted Arabic, and Latin-scripted English as the main linguistic varieties found on the English website of mahjoob.com. Lastly, I outlined the potential value to the field of sociolinguistics, especially of Latin-scripted Arabic as a novel, but potentially fleeting, example of technology-constrained linguistic innovation. The next chapter will review the literature which informs the present research.
Chapter 2: The sociolinguistic context

2.1 Introduction
This research presents a dilemma: what can be said to be the true sociolinguistic context of the study? On the one hand, the research involves language users who inhabit physical bodies and specific geographical locales. On the other hand, these same individuals meet virtually in cyberspace and interact linguistically in that medium as opposed to the “real world” (Hutchby, 2001). So, when considering the question of context, in order to be as thorough as possible, a discussion of both the face-to-face context of the Arabic-speaking world and the virtual context of Arabic CMC is desirable. This will be the focus of the current chapter, which acts as the main literature review for the thesis. First, the notions of diglossia and bilingualism are introduced in general terms. Next, these concepts are considered in relation to the Arabic-speaking world and to Jordan specifically, the hosting country of mahjoob.com, which largely informs the sociolinguistic and cultural practices of the website. Finally, the field of multilingual, multiscriptal CMC is presented in both broad terms as well as in terms of its relationship to Arabic and Arabic-English bilingualism and diglossia.

2.2 Diglossia and bilingualism
This section is broken into two sub-sections. First, early research is presented on defining the concepts of diglossia and bilingualism. Next, these concepts are related to research in and on the Arabic-speaking world.

2.2.1 Defining diglossia and bilingualism
Ferguson first introduced the term diglossia into English as a loan word modelled on the French *diglossie*, which, in turn, was coined from the Ancient Greek “di” meaning two and “glossa” meaning language. By this term, Ferguson described “a particular kind of standardization where
two varieties of a language exist side by side, with each having a definite role to play” (Ferguson, 1959, p. 232).

While Ferguson’s original definition focused on related languages which occurred in a given society such as Classical Arabic alongside vernacular Arabic, Fishman (1967) applied the concept of diglossia to societies in which even unrelated languages were functionally distributed. Central to both versions of diglossia was the idea that languages in a given society can be categorized socially and functionally as either High languages (‘H languages’) or Low languages (‘L languages’). Typically, H languages are languages of culture and education which are highly standardized and represent a relatively long and celebrated literary tradition. They are usually also restricted to more formal communicative contexts such as university lectures and official, e.g. governmental and legal, settings. Ferguson notes that they are also often associated with religion or scripture and may be used liturgically and/or in sermons. In this light, the so-called classical languages of Latin, Ancient Greek, Biblical Hebrew, Ancient Chinese, Sanskrit, and Classical Arabic could all be considered examples of H languages, though more modern languages such as Standard French, Standard English, or Modern Standard Arabic can and often do serve as H languages in their own right. At the other end of the sociolinguistic spectrum, the L languages which are related, however distantly, to their respective H languages serve the communicative needs of everyday life in informal and intimate contexts. So, Standard French and Haitian Creole French would be considered H and L languages respectively in a setting such as Haiti (Ferguson, 1959, p. 233). Similarly, in the current research, Vernacular Arabic would be deemed to be an L language in opposition to Classical Arabic and Modern Standard Arabic as H languages.
It is telling that Ferguson’s seminal work on the topic of diglossia actually cites the Arabic-speaking world as a classical example of diglossia with Classical Arabic as an H language and various forms of Vernacular Arabic as L languages. In this same work, Ferguson also alludes to the “analogous situation” where two distinct languages, related or otherwise (as opposed to varieties of the same language), co-exist in a speech community with specific roles (Ferguson, 1959, p. 233). Nevertheless, Ferguson does not provide any concrete examples of this kind of bilingual diglossia here. As will be seen in 2.3.2 below, the functional distribution of English and Vernacular Arabic in Jordan can also be considered to be an example of diglossia involving unrelated languages (see Fishman, 1967, p. 47).

As a result of this theoretical vagueness, it was left to Fishman (1967) to further broaden the original concept of diglossia. Fishman expanded diglossia to cover four distinct sociolinguistic situations pertaining to the overlap of bilingualism and diglossia: 1) bilingualism with diglossia, 2) bilingualism without diglossia, 3) diglossia without bilingualism, and 4) lack of both bilingualism and diglossia (Fishman, 1967; see also Romaine, 1995). Fishman explains that the first type of diglossic situation occurs when two or more languages serve distinct purposes in a given speech community and where the social values attached to each are different yet complementary. In the second type, he notes that there is little consensus on which language should or should not be used in a given interaction, hence it is more likely to be an unstable sociolinguistic situation resulting from either “rapid social change, … great social unrest, … [or] widespread abandonment of prior norms before consolidation of new ones” (1967, p. 51).
The third case refers to a situation in which there is typically a class-divide between the elites and the masses where neither group could claim to be properly bilingual, and where the two groups do not belong to the same speech community. The last type of situation is the rarest according to Fishman, and might describe the sociolinguistic situation of a small clan or tribe in which there is no real diversity in language among its members. Based on these descriptions, Fishman offers the Arab world as an example of bilingualism with diglossia (ibid., p. 50). By way of illustration, he makes reference to the co-occurrence within the Arab region of three languages: 1) a religiously-oriented Classical Arabic, 2) a Vernacular Arabic such as Lebanese, Egyptian or Iraqi, and 3) a scientific cum technical language such as French or English. It is this configuration of diglossia (or even triglossia) in the Arab world that we will explore further in section 2.3.

Wei (2007) describes the phenomenon of bilingualism, providing 22 common labels for different types of individuals who possess varying degrees of competency in more than one language (2007, pp. 6-7). Conspicuously, the bulk of his descriptions pertain to bilingual speakers as opposed to bilingual writers, underscoring the bias current in bilingualism studies which has tended to focus on spoken language as opposed to written language (see also Callahan, 2004; McLellan, 2005).

2.2.2 Diglossia and bilingualism in the Arabic-speaking world

Miller describes the Arabic-speaking world as “a wide and heterogeneous geographical area and includes very different types of urban settings, national constructions, social organizations and language situations…” (Miller, 2007, p. 1). In addition to this, she also notes “the spread of the
Arabic-speaking diaspora in many parts of the world” (ibid., p. 1). Consequently, we have two notions of an Arabic-speaking world: one which is geographical, the other, diasporic.

In light of this, in the present research, there is a strong suggestion from the data that authors on the website come from a variety of geographical and social backgrounds, or at least wish to project this to their audiences on the website. This is evidenced through their choice of distinctive author IDs and author locations which reflect a variety of geographical and ethnic references both from within the geographical Arabic-speaking world as well as from the Arabic-speaking diaspora. Examples of author IDs such as “Palestinian girl” with a location such as “Canada” are very common. If taken at face value, the implications of this diversity of authors for the present research are that what appear to be numerous diaspora-based contributors to the English website of mahjoob.com forums may well bring their face-to-face linguistic competencies and practices into the online context of mahjoob.com (see Al Share, 2005; Hinrichs, 2005 on the Post-Creole Continuum recreated in CMC).

Nevertheless, there are also clear indications that a large number of authors wish to highlight local identities from within the geographical Arabic-speaking world, especially from Jordan, and to a lesser extent from Palestine. Again, a perusal of the author IDs and locations in the data reveals a very large number of references to Jordanian and Palestinian towns such as the attested author ID “Saltyeh” (meaning ‘girl from Salt’, a Jordanian town). This seems to be reflected by the relative popularity of specific forums on the topics of Jordan and Palestine (see Chapter 5).

Holes’ (2008) research indicates that several key works on Arabic bilingualism and code-switching have tended to focus on North Africans both in and outside of North Africa (cf.
Bentahila & Davies) or on Egypt (cf. Ferguson, 1959). Even within the field of sociolinguistics itself, Rouchdy (2002, p. xiii) notes that the Arabic language has suffered from a similar narrowness of focus in that researchers have largely produced variation and dialectology studies (see Brustad, 2000; Eisele, 2002, p. 12; Rouchdy, 2002). Significantly for the present study, there appears to be a relative dearth of English-language research specific to the Jordanian context (but see Harrison, Prator, & Tucker, 1975). As such, this research offers yet another opportunity to remedy the apparent lack of sociolinguistic research on Jordan.24

Despite the relative paucity of such research, there may still be important lessons to be learned from research done in other Arabic-speaking contexts. For instance, while Bentahila’s study of bilingualism in Morocco (1983) differs in certain important aspects from the Jordanian context, his focus on Morocco’s diglossic situation where Classical Arabic, Post-colonial French, and Moroccan Vernacular Arabic all co-occur, still affords certain valuable observations with potential parallels for the language choice investigated in the present research. In this connection, it might be helpful to consider how English and Jordanian Vernacular Arabic in Jordan seem to occupy the same positions that Post-colonial French, and Moroccan Vernacular Arabic do in Morocco. In his findings, Bentahila notes that his subjects’ evaluations of the languages in question seemed to correlate with where, when, and how they used these languages, with clear distinctions in terms of the functions and situations appropriate to these languages (Bentahila,

24 Although the CMC nature of my data means that the ‘Arabic-speaking World’ I am studying is more of a social rather than geographical construct, it is also the case that the website I am examining originates in Jordan and is mainly used by a Jordanian audience. Therefore, paradoxically, my data simultaneously appears to represent both a diaspora and a local region.
He also notes that perceptions of the users by others differed markedly depending on which language was used in a given interaction. Speaking of the bilinguals in his study, Bentahila concludes “that the use of one language rather than the other may influence the kinds of idea which spring to his mind and the type of personality which he projects” (Bentahila, 1983, p. 163).

Bentahila summarizes these perceived differences in language roles and their importance as follows:

It can be seen, then, that the three varieties each have quite different roles to fulfill; Moroccan Arabic serves as the language of intimacy and the home, Classical Arabic as the language of religion and Arabic culture and identity, and French as the language of education and modernity. Accordingly, they are each valued for different reasons, Moroccan Arabic having a purely practical value, while the value of Classical Arabic is largely aesthetic and that of French is clearly instrumental. Each is recognized as useful in Moroccan society, and bilinguals evidently feel that each is necessary to the Moroccan. The contrasting images possessed by the three suggest that they are complementary to one another, no one being surplus to the requirements of daily life (Bentahila, 1983, p. 165).

Bentahila’s findings in the Moroccan context are potentially illuminating in helping to uncover sociolinguistic parallels for the types of language interaction, and the values attached to them, by posters to the mahjoob.com website where written forms of Classical Arabic, Modern Standard Arabic, Jordanian Vernacular Arabic (see Section 1.3), and English are all encountered to varying degrees. In other words, while the linguistic resources between the contexts of Morocco
and mahjoob.com may differ on the surface, the linguistic responses which make use of these resources may actually be quite similar.

Broadening the focus to consider Arabic in a general sense, Eisele reviews the dominant discourses among traditional Arabic scholars pertaining to Arabic and its diglossic relationship with other languages both past and present (2002, pp. 7-9). He observes that historically, Classical Arabic came to be increasingly sidelined by other cultural ethnicities and their languages, especially, Persian and Turkish. As a result, Classical Arabic became confined to religious and legal domains, a process driven by non-Arab elites such as Ottomans within the Middle East. It is worth noting that as a linguistic reflex to this deference to Ottoman rule and Turkish language use, some 200 years ago Turkish-sounding Arabic accents became prestigious especially in cities, a legacy which continues (see Holes, 1995, p. 279). Eisele also explains how European colonial powers impacted on the diglossic situation in the Middle East by introducing languages such as French and English as languages of administration, education, and (Western) culture in place of Classical Arabic. Eisele claims that this development appears to have further limited and weakened the position of Classical Arabic. Within the present research, it is interesting to investigate whether there is any evidence of additional supplanting of Classical Arabic and if so, by which other codes and in which functional domains. A corollary to this is to examine the status of the Arabic script vis-à-vis the Latin script and whether or not any kind of change in its distribution, functions, and prestige may have occurred (see Al Share, 2005; Sakarna, 2006).
2.3 The case of Jordan and its importance to this thesis

Despite having both English and Arabic websites, mahjoob.com, as noted earlier, is ostensibly a website for Jordanians maintained by Jordanians. Although it features forums with names like Palestine and Iraq under occupation, by far its largest forum in terms of the number of messages is the nationality-related forum, Kuluna al Ordun (‘We Are All Jordan’). The title of this forum itself suggests that Jordanian identity is assumed to be the overriding badge of membership for many of the forum participants. In support of this view of the essential ‘Jordanian-ness’ of the website is the fact that the website owner, political cartoonist Mr. Emad Al-Hajjaj is something of a Jordanian celebrity for the creation of his Jordanian everyman cartoon character Abu Mahjoob (the namesake of the website), as well as for his popular social critiques of the Middle East, Israel, and the West. In fact, Mr. Al-Hajjaj has reached such renown as to be featured on Daryl Cagle’s influential Political Cartoonist Index (Cagle, 2009).

Another of mahjoob.com’s website features, which underscores the target audience of the website as Jordanians and Palestinians, is its online surveys which focus principally on Jordanian and Palestinian topics. In addition, the unmistakably Jordanian Vernacular Arabic names of a number of forums further highlight the website’s links to Jordanian society and culture. With

---

25 Daryl Cagle is the editorial cartoonist for the American television news channel MSNBC, as well as being a widely syndicated newspaper cartoonist.

26 It should be borne in mind, however, that it is virtually impossible to prove that a majority of website posters are in fact ethnically Jordanian or that they are even physically based in Jordan. Nevertheless, the circumstantial evidence mentioned above as well as the linguistic evidence available in the form of the frequency of Jordanian vernacular texts in the corpus all seem to indicate that the website is predominantly frequented by individuals with a clear connection to Jordan.
this in mind, the following sociolinguistic background to the country of Jordan will help to provide much of the context for interpreting the language behaviour observed in mahjoob.com.

2.3.1 The Jordanian context
Since the country of Jordan (officially, the ‘Hashemite Kingdom of Jordan’) features prominently in the research, a brief description of its history and demographics is warranted here. Jordan was part of the Ottoman province of Syria until the end of World War I when it was mandated to the UK as the combined territory of Transjordan (East bank) and the West bank territory of Palestine. Subsequently, Jordan gained its independence from the UK in 1946. The population at this time was primarily composed of rural Arabic-speaking Sunnite Muslim Bedouin tribes in the Transjordan area and city-dwelling Muslim and Christian Palestinians in the West Bank area. A small percentage of the Transjordanian population was made up of nuclei of Arabic and Armenian-speaking Christian town-dwellers as well as Muslim Circassians and Chechens. However, the establishment of the Jewish state of Israel in 1948 and the subsequent upheaval in its Palestinian Arab districts precipitated a massive influx of Palestinian Arabs from West Bank areas into the Transjordanian part of the Kingdom. This population increase in the Transjordan was further bolstered by the loss of the West Bank to Israeli occupation as result of the 1967 Arab-Israeli War. In the early 1970s an intense conflict erupted in Jordan in which pro-Palestinian groups attempted to overthrow the monarchy, but ultimately failed (Holes, 1995). Despite this, the growth of the Palestinian Jordanian community has continued to such an extent that, at present, ethnic Palestinians constitute a slight majority of Jordan’s citizenry. The current population is approximately 6.3 million, 31% of which is under the age of 15.
Ethnically, Arabs currently account for 98% of the population with Circassians, Chechens, and Armenians making up the balance (CIA, 2009). Linguistically, however, besides Arabic, the national language, the CIA report observes that English is widely understood and used among the middle and upper classes. This reflects the popular perception among Jordanians of English as a marker of higher social status (see Khuwaileh & Al-Shoumali, 2001, pp. 31-35). Indeed, several prestigious private schools in the country at the primary, secondary, and post-secondary levels are either partly or wholly English-medium (Hussein & Zughoul, 1993).

According to the CIA World Factbook, in contrast to its rural past, some 78% of the population is now urbanized. In terms of education, Jordan enjoys an average literacy rate of 90% overall. Regarding Internet usage, Jordan is estimated to have about 1.5 million users i.e. just under one quarter of its population (CIA, 2009).

2.3.2 English-Arabic diglossia and bilingualism in Jordan

Hussein and Zughoul (1993) emphasize that the Hashemite Kingdom of Jordan’s historical relationship with Arabic and English is not atypical of other countries in the Middle East. Accordingly, Jordanians witnessed their first real exposure to English during the British colonial period spanning most of the first half of the 20th century and ending just after World War II. This period was characterized by enhanced use of English by Arab elites who played a liaison role between their fellow Jordanians and their British overlords. As such, English appears to have been recognized as a new language of prestige in place of the Ottoman Turkish whose administrative functions it had only recently usurped. Despite this, Hussein and Zughoul point out that for the average Jordanian there was probably little real contact with English in this
period. Exceptions to this were the limited number of English-medium schools opened by Western Christian missionaries and servicing Christians and elites.

This limited contact with English was destined to change, however. According to Hussein and Zughoul (1993), paradoxically, it was after Jordan achieved full independence from Great Britain in 1946 that English saw its role expand as it came to be increasingly identified with modernization and development. As a result, as in other Arabic-speaking contexts where French played a similar role, the ex-colonial language of English was inserted into the education system alongside Arabic in a systematic fashion. In a similar way, Hussein and Zughoul point out that the fields of technology, business, medicine, aviation, tourism, and higher education have all come to be dominated by English in Jordan.

Hussein and Zughoul also highlight the increased exposure to English which Jordanians now have. For instance, they report that Jordan has its own English language daily newspaper and a weekly news magazine (1993, p. 240). In addition, their 1993 study observed that Jordanian TV English language programming was already available for five hours daily and that the Jordanian Broadcasting Service hosted its own English radio station with 14 hours of daily broadcasting.

Harrison et al. (1975) note that English had become a sociolinguistic marker of higher class Jordanians in the wake of independence. As noted above, some 30 years later, the CIA World Factbook for Jordan (2009) corroborates this finding.

In terms of covert prestige, Holes’ description of linguistic tendencies within Amman Jordanian Arabic is also of particular relevance to the present research (1995). Essentially, Holes identifies three broad categories of Arabic vernacular represented in Jordan. It should be noted at the outset
that these vernaculars reflect the divergent social and geographic origins of the Arabic-speaking population of Jordan and do not include the significant non-Arab minorities in Jordan such as Circassians, Roma, Chechens, and Armenians although members of these ethnic groups are also generally fluent in one or more of these Arabic vernaculars in addition to their ethnic languages (Suleiman, 2008, pp. 131-133).

Holes (1995) examines the sociolinguistic impact of societal attitudes toward different varieties of vernacular Arabic in terms of the frequency of phonological realizations of the Classical Arabic voiceless uvular plosive /q/ among Jordanians living in Amman. This Classical Arabic and Modern Standard Arabic phoneme is realized in three different ways in these vernaculars: In the urban vernacular, it is most frequently realized as the glottal stop /ʔ/. In the Fellaheen vernacular, it is realized as /k/. In the Bedouin variety, its normal realization is /g/. In a sociolinguistic survey, Holes relates that the /k/ realization is the most stigmatized and consequently least frequently encountered among urbanites (see Suleiman, 2008, pp. 128-129). On the other hand, /ʔ/ is widely encountered especially among females. The phoneme /g/, however, is the second most encountered realization, particularly frequent among young males, though it appears to compete with /ʔ/ in this regard. Holes found that among this demographic grouping, the two variants were evenly split. Interestingly, this phonological distinction also appears to be replicated graphologically in CMC environments. As proof of this, Al Share (2005) relates that males in chat rooms actually vary their use of the Latin script grapheme ‘g’ for /g/ and ‘2’ for /ʔ/ according to their addressees. These males favour the former when addressing same sex peers and the latter for use with females. Al Share concludes that this phonological-graphological shifting behaviour betrays a kind of linguistic accommodation reflex whereby
males want to appear more masculine among other males and more gentle, sophisticated, and educated among females (also Suleiman, 2008).

Sakarna (2006) draws the link between urbanization mentioned earlier, social class, technology, and language use in Jordan. He observes that upwardly mobile young Jordanians in urban settings such as Amman routinely code-mix between Arabic and English and carry this type of language behaviour to online contexts. Dubbing this new variety “Englo-Arabic” in light of its hybrid lexical nature and predilection for Latin script in online contexts, Sakarna supports the idea that a new urban youth ‘lect’ (see Sebba 1997) has emerged in both face-to-face and CMC contexts among young Jordanians. For the present research, Sakarna’s study offers some potentially useful insights for tracing the ideological, demographic, and social characteristics of the forum posters at mahjoob.com.

Suleiman (2008) provides an overview of the language ecology of the Middle East using Jordan as a case study. His study confirms much of the earlier findings of Holes, Cadora, and Al Share regarding the relative valuations of nomadic, rural, and urban vernaculars in Jordan. However, Suleiman rejects essentialist views which posit that certain phonological realizations always index specific identities as prestigious, stigmatized, or otherwise. Instead, Suleiman argues that the potential sociolinguistic reception of a given phonological form must be viewed within the context in which it is used. As proof of this, he cites the interesting example of linguistic accommodation among male Jordanian Palestinians who use the /g/ phoneme with indigenous male Jordanian interlocutors in urban settings such as Amman and then switch to /ʔ/ in their interactions with Syrian urbanities along the Syrian-Jordanian border (Suleiman, 2008, pp. 130-131). Suleiman also highlights the rise in /g/ among Jordanians as emblematic of a resurgent
indigenous Jordanian national identity. This appears to have occurred in the wake of the political
and cultural victory in the early 1970s of the ‘ethnically Jordanian’ Hashemites over the
‘immigrant’ Palestinian leadership when this latter group seemed capable for a while of eclipsing
the former (see Holes, 1995, pp. 277-278). Based on these observations, my analyses of the
identity work performed on mahjoob.com will take into account as much as possible the context
of interactions and whether or not specific identities can in fact be traced through selective use of
orthography in an attempt to mimic face-to-face phonology (Palfreyman & Al Khalil, 2003; Su,
2003) (see Chapter 7).

So far, most of the research I have discussed on diglossia and bilingualism in Arabic (and
particularly Jordanian) contexts, has focused on more traditional (e.g. spoken and written) modes
of interaction (with the exception of Al Share (2005) and Sakarna (2006)). In the next section, I
will discuss the phenomenon of computer-mediated communication (CMC) and further relate
this to use of Arabic, to Jordan, and to Arabic-English bilingualism and diglossia.

2.4 CMC and language use

Herring’s (1996) reference work on CMC delineates some of the possible wide-ranging impacts
of CMC on language forms (see Baron, 2008). In similar fashion, Crystal (2001) presents a broad
scope of potential research areas in the burgeoning field of what he terms “Internet linguistics”.

Examining the emergence of the Internet and its impact on language use and form, Crystal
proceeds to detail the salient features of “Netspeak” as a dialect of the Internet. Although his
treatment is admittedly English language-centric, he does raise the important question of how
other languages are making the transition to online domains.
Danet and Herring (2007) provide an introduction to the emergent phenomenon of CMC in languages other than English. They identify technical constraints such as the ASCII-based interface which obliged early CMC adopters to compose local languages in the Latin script. They also raise the issues of patterns of code-switching and code-mixing as well as the influence of the conventions of “Netspeak” on CMC in different languages. Furthermore, the authors allude to the possibility that CMC texts might reflect a third genre of language which blurs the traditional lines between conventionally spoken and written forms of language. While this last assertion appears to apply most aptly to synchronous forms of CMC such as web chat, in the present study, initial analyses of asynchronous web forum posts and blogs indicate that Vernacular Arabic provides the basis of CMC-based Written Arabic. This is especially true of Latin-scripted Arabic as opposed to either Classical Arabic or Modern Standard Arabic (see 1.3 for a description) which are the normative written codes for non-CMC Arabic (see Chapter 3).

2.4.1 CMC-based bilingualism, code-switching and script-switching
Androutsopoulos observes that “bilingual interaction is still a neglected issue in the study of the multilingual Internet” (2007, p. 340). To help remedy this situation, he explores code-switching in three diasporic web forums among ethnic Persians, Indians, and Greeks living in Germany. His analysis of a Persian-German website takes into account how forum topics may serve as potential cues for differentiated language use of German and Farsi. In this regard, Androutsopoulos’ findings indicate that certain forums do in fact correlate with different codes. For instance, Persian is used most frequently and consistently in forums related to joke-telling and those featuring erotic pictures.
In one notable code-switching interaction, Androutsopoulos demonstrates how two male ethnic
Greek forum participants employ Greek and German selectively in order to tease a female ethnic
German forum participant: The Greek participants direct Greek messages towards the German
participant who obviously does not know Greek while using German with each other to talk
about her. This highlights the fact that code-switching can be used strategically to both exclude
and otherize interlocutors.

Nishimura’s (2003) study presents numerous novelties of written Japanese CMC vis-à-vis
standard written Japanese. Her findings, indicating that many features of Japanese CMC appear
to be common in English CMC, suggest the existence of broader CMC tendencies or even CMC
universals (see Crystal, 2001). For instance, frequently encountered stylistic features such as
abbreviations and emoticons (also known as smileys)27 appear to characterize both the Arabic-
scripted Arabic and the Latin-scripted Arabic texts in my own data set as well.

Su’s (2003) study of CMC practices of Taiwanese college students raises several important
issues which also relate to code and script choice. First of all, evidence that the subjects in the
study violated standard conventions of transliteration reflects the creative nature of their
language use. Also, acts of code-switching and code-mixing between English, Putonghua, and
Taiwanese Mandarin suggest the diglossic and identity-formative aspects of these online
interactions. This theme of identity-related code-switching will be taken up again in Chapter 7.

27 Emoticons or smileys refer to sets of generally small graphic items which can be inserted into written texts.

Usually, a range of faces with different expressions are available to CMC users. In this research, several examples of
the use of emoticons will be examined as they add meaning to the multimodal discourse of web forum postings.
Wodak and Wright (2007) offers a look at online language choice on the EU government-sponsored multilingual web discussion forum *Futurum* which allows popular debate on language policies in the EU. The researchers employ a mixed quantitative and qualitative approach by first determining language usage on the entire forum and then selecting a specific thread for detailed discourse analysis. For their quantitative analyses, Wodak and Wright examined language usage in each thread, paying particular attention to English seed vs. non-English seed posts\(^{28}\). Their findings indicate that language of seed post was in fact a significant indicator of the subsequent posts in a thread. This finding seems to support Gumperz’s situational code-switching theory that the language used in an initial frame will invite replies in that same language. Nevertheless, they also found that non-English seed posts still received a high proportion of subsequent replies in English though French was the most common language in such threads. Together, these results seem to confirm the primacy of English in multilingual CMC contexts (see Bianchi, 2005; Paolillo, 2007, p. 425).

Beyond these findings, the qualitative approaches employed by Wodak and Wright are relevant to the present investigation. As a first step, the researchers delineate the major code choice patterns across forum messages. Next, they carry out a micro-analysis of a salient thread which deals precisely with the topic of language use in the EU. In this research similar methodologies are used: patterns are traced across the whole corpus and then specific threads are selected for detailed discourse analysis based on their salient linguistic characteristics (see Chapters 4 and 7).

\(^{28}\) A *seed post* refers to an opening post i.e. the initial post that starts off a given thread (see Section 1.5.8).
Hinrichs’ (2005) doctoral dissertation explores the uses of Jamaican Creole in CMC. He probes the question of whether code-switching between Jamaican Creole, Jamaican English and Standard English, which is well-attested in face-to-face environments, also occurs in e-mail and if such code-switching parallels the functions of face-to-face code-switching. In his work, Hinrich refers to two dominant theories to describe when two varieties of a language co-occur within a given society: diglossia and the basilect-acrolect continuum\(^{29}\) (2005, p. 66). Hinrich’s treatment of the subject provides a useful point of departure for exploring whether or not a similar basilect-acrolect continuum might also be present in Arabic CMC, especially given that Ferguson (1959) describes Arabic as a classical example of diglossia. Other important issues raised for the present study are to what extent Latin-scripted Arabic may be considered a new variety of Arabic rather than a simple “transliteration” of the vernacular into a written register, and the related question of what happens to a primarily spoken code when it takes up written domains.

\(^{29}\) Stewart (1965) coined the terms *basilect* and *acrolect* to refer to the co-occurrence of two lexically-related languages in a single society where, at one end of the spectrum, one of these languages functions sociolinguistically as an acrolect (high language) used in formal settings and domains while the other language functions as a basilect (low language) used in informal settings and domains. The notion of spectrum foregrounds the observation that clean boundaries do not exist between the acrolect and basilect. Bickerton (1975) then introduced the concept of mesolects (middle varieties) to describe the linguistic gradations that exist between the basilect and acrolect which exhibit features of both. In some ways, Modern Standard Arabic itself occupies a somewhat mesolectal position between Classical Arabic and Vernacular Arabic. Nevertheless, there are also mesolectal forms of language between both Modern Standard Arabic and the Vernacular Arabic (see Holes, 2004, pp. 341-382 for a relevant discussion).
McLellan (2005) examines Bruneian Malay-English code-switching in web forum data in his doctoral thesis. His study employs a mixed quantitative-qualitative approach to language alternation. McLellan begins with a quantitative frequency analysis of Bruneian Malay vs. English occurrence in his corpus data. He then applies a quantitative grammatical analysis which is triangulated with a questionnaire on motivations for code choice. Using the Myers-Scotton’s Matrix Language Frame Theory, McLellan determines that in his own data several cases of equal language distribution in mixed-code texts cast doubt on the explanatory power of Myers-Scotton’s theory\(^\text{30}\). McLellan points out that his investigation of CMC-based Malay-English code-switching is novel given the relative genetic distance between these two languages and the fact that little research has been done on CMC in languages other than English. The present work in Arabic-English CMC is motivated by similar reasons with the added dimension that not only do the two codes vary greatly, their writing systems are also vastly different.


\(^{30}\) According to Myers-Scotton’s Matrix Language Frame (MLF) Theory (1993a, 1993b), in each code-switching situation, one language functions as the dominant or matrix language (ML) of the interaction. The other language is termed the embedded language (EL) which is borrowed from and incorporated into the code-switcher’s ML utterances to achieve code-switching. In this theory, the ML accounts for most of the lexis and grammar in code-switched utterances so that there can never really be equal distribution of two languages in a code-switched utterance i.e. there is always a primary or dominant language which provides the frame for the EL to be incorporated into. McLellan (2005) challenges this claim, pointing to several instances in his data where Bruneian Malay and English appear to be equally used in texts in a way that makes it impossible to tell to which language is the ML and which is the EL.
code-mixed blog texts. Smedley observes that code-switching in blogs seems to challenge earlier notions of code-switching being essentially conversational and dialogic. Applying Bakhtinian concepts of dialogism, heteroglossia and voice, Smedley reports that code-switching in blogs helps text-producers to construct complex identities despite the apparent lack of specified addressees. Although the present study does not include blogs, Smedley’s emphasis on complex identities has been suggestive of ways in which identity construction might be achieved linguistically among mahjoob.com forum posters.

In the past few paragraphs, the recent literature on CMC-based code-switching involving various languages has been surveyed. The next section deals with studies of CMC that have focused primarily on Arabic-English code-switching.

2.4.2 CMC involving Arabic-English code-switching and script-switching

In her introduction to a contrastive analyses of Arabic urban vernaculars, Miller (2007, p. 1) observes that “cities are 'par excellence' places of contact and heterogeneity”. I argue that the same holds true of web forums. Like cities, these virtual meeting places offer participants from different social and linguistic backgrounds the opportunity to communicate with each other. The following studies indicate that heterogeneity is a hallmark of Arabic-English CMC-based interaction.

Palfreyman (2001a) surveys existing informal orthographies that employ Latinization due to technical constraints. He highlights the widespread use of numerals as graphemes in Latin-scripted versions of languages such as Greek, Russian, and Arabic. Curiously, this novel phenomenon, which is observable in other non-Latin script languages when Latinized in CMC such as Greek and Russian (see Palfreyman, 2001a; Tseliga, 2007), has not received much
attention in the literature, nor has it been labelled. Consequently, I have coined the term “arithmographemicization” to refer to the process of appropriating numerals for use as graphemes (based on the Greek prefix *arithmo-* meaning ‘number’). The resultant graphemes I have dubbed “arithmographemes” and the systematic use of such arithmographemes within a particular orthographic system I call “arithmographemics”. Arithmographemics aside, Palfreyman’s work represents one of the earliest records of the Latin-scripted Arabic phenomenon in the literature.

Warschauer et al. (2002) examine linguistic pluralism on the Internet taking Egypt and Singapore as cases in point. Focusing on Egyptian Arabic-English bilinguals, the researchers found that approximately half of the 43 subjects in their study reported that they frequently used Latin-scripted Egyptian Arabic in their chat and private e-mails. This work is seminal in bringing the occurrence of Latinization of vernacular Arabic into the literature. In addition, their observations and analyses regarding online code-switching and script-switching point out that

[i]n bilingual messages, Egyptian Arabic was most often found in greetings, humorous or sarcastic expressions, expressions related to food and holidays, and religious expressions…(2002, p. 11)

These observations provide a basis for investigation of language roles in my own selected data sets, especially among ostensibly bilingual and biscriptal Arabic-English CMC users.

Palfreyman and Al Khalil (2003) investigate what they refer to as “ASCIIized Arabic”, namely the Latinized variants of Arabic found in online chat rooms. They compiled a corpus of ASCIIized texts and analysed these for orthographical features. They note the common usage of number graphemes to represent sounds not readily associated with any of the Latin script’s 26
standard characters. This work is also seminal in that it attempts a linguistic analysis of ASCIIized Arabic for salient orthographical features. The authors’ observation that Latinization sometimes occurs even when it is clear that the text producer has access to the normative Arabic script implicitly raises the issue of script choice, which is central to the present research.

Al Share (2005) observes that very few studies to date have been done on what she terms Jordanian Netspeak, the Jordanian Vernacular Arabic found in web chat. Web chat is a synchronous form of CMC and as such is shaped by the communicative exigencies and constraints of simultaneous interaction (Herring, 1996). In contrast, web forums are a form of asynchronous CMC and therefore afford participants more time and reflection in both production and reception of texts (Herring, 1996). It is therefore entirely plausible that differences in text production might be discernible between synchronous and asynchronous forms of CMC. For instance, in a personal communication, Al Share points out that script-switching is virtually absent in the web chat data which she has compiled. On the other hand, my own data confirms that script switching within a single forum message is not only possible, but is actually well attested in several cases (see Chapters 3 and 5). What this means for the present research is that the asynchronous element of web forums is likely to be a determining factor in the ability to script-switch. Thus, asynchronicity can be considered a unique affordance of web forums (also available to e-mail and SMS text message composers), enabling authors to script-switch more readily than in synchronous web chat contexts. As an important aside, it is worth noting that while both e-mail and chat involving Arabic and English have been studied, to the best of my knowledge there have been no studies to date done on Latin-scripted Arabic in web forums.
Al Share (2005) provides an orthographic description of CMC-based Latin-scripted Arabic among Jordanian web chatters similar to Palfreyman and Khalil’s study in the UAE (2003). Al Share also carries out a comparison of orthographical patterns observable in chat room discussions featuring male only and male-to-female discourses. Her findings indicate that text-producers modify their linguistic output to accommodate their audiences, with males adopting different orthography when writing to females compared to other males. This key finding of Al Share is relevant to the present study because it implies that Jordanian Latin-scripted Arabic users are able to create distinct identities in CMC contexts through the use of particular linguistic forms, especially orthographical ones (see Sebba, 2007)

2.4.3 Evidence of diglossia in CMC

Earlier in this chapter, both Ferguson’s and Fishman’s views of diglossia were discussed. While these views seem to apply to a number of sociolinguistic settings, they do nevertheless seem to present an essentialized view of H languages and L languages. I take issue with this dichotomous view of language varieties as either H or L arguing that in the realm of multilingual CMC the differentiation of roles played by each language may not be as clear-cut as Fishman would have us believe. In fact, in the English section of the mahjoob.com website, bilingualism without diglossia may well be the norm precisely for the very reasons given by Fishman above, namely, the relative social upheaval engendered by the new CMC space coupled with an abandonment of old norms before new ones have been firmly established (see Herring, 1996) (see Chapter 8).

Of particular relevance to the present study is the fact that in his illustration of contexts of diglossia, Ferguson cites the Arab world as a prime and longstanding example, contrasting Classical Arabic, the H variety, with Egyptian Vernacular Arabic, the L variety. Ferguson then
outlines ways in which the H and L may differ. In terms of function, H and L are used for different purposes and in different contexts, they are in complementary distribution. For example, in the case of Arabic, Ferguson mentions that Classical Arabic is used for the delivery of university lectures while subsequent discussions will usually be in Vernacular Arabic. The H and L varieties of Arabic also differ in terms of prestige, literary tradition, methods of acquisition, and level of standardization. To illustrate, Ferguson argues that the H, in contrast to the L, is always more highly valued, has a long and considerable literary tradition, is learned at school not at home, and is grammatically, stylistically, and orthographically-standardized (Ferguson, 1959, pp. 237-240).

Although these propositions all seem to hold true for Classical Arabic and Vernacular Arabic, the research here will challenge in particular the purported prestige of Classical Arabic and its status as the written language par excellence of the Arabic world, at least in CMC contexts. Indeed, one of the central claims of this study is that newly evolving CMC-based literacy practices are reshaping much of the diglossic landscape of the Arabic world (see Al-Tamimi & Gorgis, 2007; Sakarna, 2006). Again, I take a line that actually challenges the very notion of H and L, arguing that there are distinct problems in trying to identify the H vs. L language in a given sociolinguistic context. Indeed, perhaps in Arabic-English web forums, a mixed language featuring both Latin-scripted Arabic and English is the normative or unmarked code (see McLellan, 2005) (see Chapter 3 and Chapter 5). In addition, given the graphemic advantages of an ‘out-group’ writing system such as the Latin script which enables users to relatively easily self-identify as /g/ users vs. /k/ or /ʔ/ when compared with Arabic script (see Lee, 2007; Palfreyman & Al Khalil, 2003), a ‘fitness for purpose’ model of language choice rather than a
diglossia model might be more appropriate for describing the code-switching behaviour on mahjoob.com.

In Romaine’s survey of post-Fergusonian views of diglossia, she indicates that concepts such as ‘triglossia’ and ‘broad diglossia’ and ‘polyglossia’ have also been coined to describe various sociolinguistic contexts in which more than one H or L code occur in functional distribution (1995). In this connection, Platt’s (1977) concept of polyglossia might be applicable to the English website of mahjoob.com because of its apparent functional distribution of Classical Arabic, Modern Standard Arabic, and Vernacular Arabic across forums, in conjunction with a range of more or less formal varieties of English.

Having said this, great caution is warranted in trying to compare the web forum domain to other functional domains of language use in face-to-face society such as say, the mosque, to use one of Ferguson’s original examples. This is because there are no direct one-to-one correspondences between online asynchronous discussion board contexts and face-to-face synchronous oral contexts. For one thing, the fact that scripts can be switched has no parallel in the spoken world. Speakers can change their accent, perhaps, but cannot adopt a whole new phonology while speaking a language and still expect to be understood by their audience. Thus, conscious script-switching adds a new stylistic dimension to the written interaction that has no ready equivalent in the domain of speech. Nevertheless, Al Share (2005) finds that at least in synchronous forms of CMC such as Internet Relay Chat (IRC), spoken norms do in fact seem to form an important source of input for chat communication and that interlocutors have spoken models in mind when they compose their synchronous texts in an attempt to approximate spoken discourse (see Hinrichs, 2005; Nishimura, 2003; Palfreyman & Al Khalil, 2003). In the current study which
deals with asynchronous written material, script-switching is viewed as a stylistic device which enables its users to construct their identities in novel and complex ways which are unavailable to them in regular face-to-face interaction.

2.6 Summary

In this chapter, I have briefly reviewed both the face-to-face and the online contexts of research on bilingualism and diglossia, relating this to my own research topic. I also discussed CMC-based code-switching and script-switching. I have also tried to show how the existing literature points to a gap in the research: although Latin-scripted Jordanian Arabic has been examined by a couple of researchers (see Al Share, 2005; Sakarna, 2006) no one has yet examined CMC code-switching and script-switching involving Arabic and English and the Arabic and Latin scripts in the Jordanian context.

In this regard, the English side of mahjoob.com represents a rich context for sociolinguistic research. In terms of diglossia, it features 41 topically-differentiated forums. Each of these forums, in turn, can be classified in terms of whether its main topic is more or less humorous, traditional, patriotic, professional, leisure oriented, etc. (see Chapters 3 and 5 for more detail). Thus, a closer examination of the dominant codes featured in each of these forums promises to shed light on whether or not topic does in fact contribute in part to determining code use. For instance, the Religion forum contains several different threads each with their own sub-topics. Many of these are dedicated to discussing aspects of Islam while a few others are designated as ‘Christian’. In light of previous diglossia research, we might be led to assume that Classical Arabic would be featured more often in the Islamic threads whereas the Christian threads might feature relatively more Vernacular Arabic or even English. But given Suleiman’s research cited
earlier in this chapter, there is a definite risk involved in drawing such simple linear relationships between language use and topic, despite evidence that topic may well correlate with certain codes (see Androutsopoulos, 2007 above). This call for caution is warranted and duly noted. Therefore, I proceed as much as possible with an open mind about how the scripts and codes in the corpus are woven together around specific topics. One of the main advantages of using a corpus-driven approach to the discursive data available from the mahjoob.com corpus is that it gives an opportunity to verify any prior assumptions about language use and topic quantitatively, as opposed to drawing conclusions impressionistically.

The above-mentioned studies on multilingual CMC have tended to employ corpus linguistic methods in order to establish frequencies and usage patterns among their respective subjects. In Chapter 4, I will elaborate on my own use of these methods and describe the corpus in fuller detail. In doing so, I hope to establish the value of such an approach, emphasizing that corpus linguistics offers discourse analysis an empirical method of exploring CMC data and therefore can be used profitably when trying to establish both quantitative and qualitative evidence of code-switching and script-switching patterns.

Now that the context of the research has been described, a description of the actual linguistic codes encountered in the corpus will be useful in providing an understanding of the basic linguistic resources Mahjoobians have at their disposal to do their identity work. Thus, I will turn my focus to the primary linguistic variables which are featured on the website, namely, Arabic-scripted Arabic, Latin-scripted English, and Latin-scripted Arabic. This discussion will

---

31 Arabic-scripted English is also evident on the website but is quite infrequent (see Chapter 3). Consequently, while examples of it will be cited where relevant, it is not deemed to be a primary linguistic variable in the research.
naturally encompass the method used for defining and delineating these variables. This will be the focus of the next chapter.
Chapter 3: Data and classification system

3.1 Overview
The focus of the current chapter is a description of the data used in this study and the classification system used to categorize this data into distinct linguistic codes for further contrastive analysis. As such, this description constitutes a response to the first research question of this study:

1) What are the various types of code and script-switching featured on the English website of Mahjoob.com?

In this chapter, first, a description of the mahjoob.com website is given including the rationale for choosing the English website for data collection. Next, the process of locating the website and selecting data from it is outlined. This is followed by a description of the data itself, the building of the corpus, and the classification system used to delineate the various linguistic codes in the corpus. Finally, the annotation system employed for the corpus is presented.

3.2 The mahjoob.com corpus

3.2.1 A description of mahjoob.com and its English and Arabic websites
Mahjoob.com is owned by Mr. Emad Hajjaj, a popular political cartoonist from Jordan, currently living in London. The website itself is hosted in Jordan and a significant number of its users appear to be Jordanian as well. However, the site also attracts posters from across the Arabic-speaking world and from the Arab diaspora as reflected by its advertising. Mahjoob.com is organized into two parallel websites, an Arabic one, and an English one. As of November 2008 the Arabic site contained 35 forums, 1,330,999 posts, 58,855 threads, and 28,025 members while the English site contained 41 forums and sub-forums, with 982,084 messages (or posts), and 13,724 members. The website is actually composed of several linked web pages, the largest of
which are the forum web pages. On the main portal to the English side of mahjoob.com there is a menu that provides various links. For instance, visitors can select the Arabic link and be taken to the Arabic website of mahjoob.com where they can then enter the Arabic-language forums. Alternatively, the visitor can stay within the English website and visit its blogs, archives, and, of course, its forums.

Given this surface division of the website by language, one might expect the English side of the mahjoob.com website to feature only English-language content and the Arabic side to feature only Arabic-language content given the pre-eminent status of English on the internet noted by researchers such as Crystal (2001). Accordingly, one would expect that the English used on the English website to be written in Latin script and that the Arabic on the Arabic website be written in the Arabic script as they are conventionally written in most offline domains\(^{32}\). However, even the most superficial browsing of both the Arabic and English forums of mahjoob.com makes it clear that forum posters do not follow these well-established conventions. Consider, for instance, the following screenshot (Figure 3.1) taken from the main page of the English website:

\[\text{Holes (2004) does note the existence of a written “mixed” variety of Arabic-scripted Arabic composed of standard and vernacular forms in limited use among Egyptians in popular print media such as magazine editorials. He observes that this kind of colloquial written style helps authors get closer to their readership and appear more “folksy” (2004, pp. 381-382).}\]

32
The English featured here is ostensibly standard with readily understood lexis and normative spelling conventions. However, on the left-hand side of the screen shot we can see the menu entitled, “Discussions”. Underneath this, there are several forums which a website visitor can choose from. Already, from the red, bolded, asymmetrical font for label “Discussions”, we anticipate a more playful, less formal kind of discussion. The punctuation accompanying the subsequent discussion options such as “Men’s Corner”, “Men Only!!” and “Girls Talk”, “No MEN!!” marked as flaming type language (see Herring, 1996) indexes informality. Upon deeper probing, it becomes clear in fact that much of the English used in the forums is actually of an informal, non-standard style. Several of the features of Netspeak (Crystal, 2001) are immediately
observable: abbreviations such as “plz”, neography (see Anis, 2007) such as “u” for “you”, “coz” for “because” and “r” for “are” are all commonplace in the forum texts. Thus, there is a picture of English which is not altogether uniform: different levels of formality are indexed by distinct lexis and orthography.

More surprisingly, perhaps is that, in addition to English, other forms of language are immediately discernible on this web page such as Arabic-scripted Arabic. For instance, the website’s logo and name are given as Abu Mahjoob written in Arabic alongside Mahjoob.com in English. Also, on the top menu, the rightmost link says “Arabic” written in Arabic and thus provides an option for website viewers who want to visit the Arabic side of the website. Again, on the right side menu, there are links to articles with Arabic names. The political cartoon of the website’s mascot, Abu Mahjoob, is also written in Arabic-scripted Arabic albeit in a Jordanian vernacular variety.

Similar to the English observable on the English website, the Arabic-scripted Arabic found on mahjoob.com cannot be said to be of one single variety or style. For instance, discussions are often carried out in written vernacular Arabic as opposed to Modern Standard Arabic which is the normative code for virtually all traditional written discourse in offline contexts (Holes, 2004). In this connection, it is telling that in Figure 3.1 above the cartoon character Abu Mahjoob “speaks” in Jordanian vernacular in the cartoon snippet just as one would expect in a “real” offline spoken context. And yet, in order to create this “real life” feeling in the cartoon, the

---

33 The cartoon approximately translated says, “So, basically you’re (the government) telling us that the global economic crisis has actually increased the size (lit. shoe-size) of the global economy! Well, there’s someone here who would like to direct a question to you (lit. towards your head)!” This is a play on words in Arabic as “your
cartoonist was forced to break the conventions of Modern Standard Arabic writing by employing vernacular lexis and structure and non-standard orthography such as doubled vowels which signal to the reader the oral nature of the interaction (see Holes, 2004, pp. 381-382).  

On this webpage, it is also interesting to note the Latin script phrase “e7ke wa fadfed” (trans. “get it off your chest”) found under the “discussions” menu of forums. Clearly, this is not English, nor is it Arabic-scripted Arabic. Rather it is Latin-scripted Arabic of a vernacular style. While barely noticeable on the main page, the presence of such Latin-scripted Arabic is in fact much greater within the web forum messages themselves as we will see in Chapter 5. Of additional interest, we also find Arabic-scripted Arabic messages posted within the English website and, to a far lesser extent, a smattering of English messages within the Arabic website. At a deeper level, we also observe that numerous postings in the various forums on the English website reveal various examples of code-mixed and script-mixed messages, whereas the messages on the Arabic website appear far more homogeneously Arabic. 

It is this relative linguistic diversity in light of the purported primacy of English on the internet that makes the English website of mahjoob.com so fascinating to explore. Put simply, several of mahjoob.com's English forum posters flout conventional wisdom that English is the language of the internet. They do not follow the rules. And such linguistic non-conformity renders the head” means “you” emphatically (cf. AAVE ‘your ass’ for ‘yourself’). This cartoon, which was published on March 7th, 2009, invokes the shoe-throwing incident which took place on December 14th, 2008 when then US President George W. Bush was the target of an irate Iraqi reporter during an unannounced visit to Iraq.  

34 The concept of borrowing vernacular forms into writing will be discussed in greater detail in Chapter 7.
English website a rich source of data for investigating code-switching and script-switching trends involving both Arabic and English. It should be noted, however, that although it is beyond the remit of this thesis, a study of the Arabic website would also be fascinating insofar as it might reveal why Arabic posters would have reason to make occasional use of English in their postings.

Having discussed the decision to focus on the English website of mahjoob.com, it will now be useful to describe how the corpus data was located and collected.

**3.2.2 Locating the data**

In recent years, several corpus linguists have discussed both the opportunities and the challenges of working with corpus data derived from the Internet. In this connection, Hoffmann (2009b, p. 2) discusses three main approaches to locating data and making use of the Internet for corpus building and analysis: 1) using the entire Internet itself as a “mega-corpus”, 2) downloading a (representative) sample of the Internet to use as a “mini-Internet” corpus, and 3) using the Internet as a “corpus shop” to locate interesting sets of text for corpus analysis. Of all three approaches, his third approach, which he considers most reliable and feasible, has been followed here. Specifically, this has meant creating a specialized corpus built from “a clearly defined set of texts” (Hoffmann, 2009b) that have been automatically downloaded from the Internet.

Hundt et al. (2007) present an edited volume of articles on issues pertaining to the use of the web in corpus linguistics. Several of these articles are very pertinent to this study given that the present data come directly from web sources. For instance, Lüdeling et al. (2007, p. 7) highlight the usefulness of the web to explore new types of language which are not represented in existing standard corpora. Lüdeling et al. then proceed to delineate two main approaches (along with sub-
options) to collecting web data: 1) searching the whole web via a commercial search engine or 2) downloading web pages and searching them locally. In order to locate data used in this study, I initially used Lüdeling et al.’s first method, using the search engine at google.com which helped uncover the mahjoob.com site as a fertile and ample web-based archive of code and script-switching involving Arabic and English. Once the website was found, the second approach was employed in order to obtain a purposive sample of web forum data which was downloaded to a server. In this way, I was able to safeguard the stability of the data which, if I had continued to access remotely, might have become corrupted, changed, or even removed from the web.

For analytic and comparative purposes across languages and varieties, Claridge (2007, p. 100) advocates the creation of a large scale corpus of forum messages, which she considers a new genre of CMC. Her pilot study compares web forum message data from four different English-speaking regions in an attempt to identify systematic discursive differences among these which might help define regional identities. Claridge advises that different graphics-to-text ratios in the data need to be taken into account when building a corpus from different message boards and that similar data must be annotated in a standardized fashion. She adopts a hierarchical structure for her corpus which is forum-thread-message. However, Claridge does acknowledge that perhaps “thread” could be put at the top of the structure hierarchy as it may reveal the effect of topicality on the messages contained within it. In the current research, the thread-message hierarchical structure has been adopted since threads appear to have more specific and focused topics than forums do.

Hoffmann (2007) outlines the creation of a large corpus of over 150 million words from 773,772 Usenet messages. In this study, he notes that since authors can “cut and paste” from across the
web, original authorship is very hard to detect. He also highlights the difficulty of establishing specific author identities since an author may contribute severally to a newsgroup using various IDs and e-mail addresses. Due to this relative uncertainty, he cautions that “on the whole, researchers need to accept the fact that next to no information is available that could form a reliable basis for any sort of sociolinguistic investigation of the data” (2007, p.162). Acknowledging this limitation, the present study is less concerned with determining ultimate authorship than with determining the actual linguistic samples of code-switching in texts regardless of the original source of the text. In fact, taking a Bakhtinian view, in addition to creating original texts, authorship can also be considered to be a process of appropriating pre-existing texts from other sources and incorporating them into one’s own texts (see Bakhtin, 1981). Indeed, in this study it is argued that, for the purpose of constructing specific identities, forum posters often appropriate pre-existing texts as a crucial step in projecting specific linguistic styles (see Fairclough, 2003 also, see Chapter 7), which are themselves modelled on other pre-existing styles.

The work of these theorists has been central in guiding the choice to select a single website for data collection and corpus building. In the next section, the actual data on the website will be described and discussed.

3.2.3 Data from the English website of mahjoob.com

As was noted above, the mahjoob.com website contained both Arabic and English sections. The data collection involved downloading only forum threads from the English language web forums as opposed to the Arabic language site.
Cursory observation of various web pages across several of the mahjoob.com forums had already indicated that language and script use on the website was quite unconventional. Specifically, traditional pairings of certain scripts with certain lexicogrammars were not maintained. Further, in several message texts more than one script was employed by a single author. So, alongside the more conventional varieties such as Arabic lexicogrammar written in Arabic script and English lexicogrammar written in Latin script, there were numerous examples of Arabic lexicogrammar written in Latin script and vice-versa. In order to capture such linguistic diversity within a single corpus, it was decided that, in addition to downloading actual forum messages to populate the corpus, it would be highly desirable to develop a coding system which could differentiate between, and annotate, strings of text written in different scripts and codes. This process will be described in Section 3.3 below.

In terms of topical content, the website was quite diverse. To illustrate, by adapting Bentahila’s (1983) topical coding scheme for investigating topic-related code choice and code-switching, it was possible to divide the 41 forums into eight broad topics: 1) Humour, 2) Poetry, 3) Work/Study-related topics, 4) Friends and Family, 5) Local Culture (nationality, politics, religion), 6) Hobbies, 7) Gender and Age-related topics, and 8) General Discussion (more detail on specific forums is provided below in Section 5.3.4). From a linguistic point of view, it is worth noting that five of the 41 forum names are actually written in Latin-scripted Arabic: E7ke wa Fadfed, Kuluna Al Ordun, A7la al-kalam, Dababees and Hala 3ammi and the San7 w Le3b w Jadd w 7obb forum. The rest are in English. There are no Arabic-scripted Arabic names among the forums. This tendency seems to be in keeping with the so-called “English” content of
the English website. If this is the case, then Latin-scripted Arabic titles are probably considered English due to the Latin script in which they are written (see Sakarna's "Englo-Arabic", 2006).

Now that this overview of the website data has been sketched, it will be useful to discuss the construction of the corpus, but first a consideration of the ethical issues involved in corpus building is in order.

**3.2.4 Ethical issues in building the corpus**

Herring’s (1996) appeal to linguists to exploit the rich research possibilities offered by the web has provided an impetus for this study. Her discussion of ethical issues concerning the distinction between data obtained from restricted access and open access web forums and blogs has informed my decision to focus on the latter source of data because open access web forums allow for unobtrusive observation, thus preventing the observer’s paradox.

Baker’s (2006) discussion of ethics in dealing with electronic texts raises the important issue of permissions, acknowledging that theory and practice often diverge in this regard. While advocating obtaining permissions where possible, he admits that this is not always feasible (Baker, 2006, p. 38). This is especially true when critical discourse analysts cannot realistically expect to obtain permission for their use of texts which, once analysed, might show their text-producers in a negative light. Nevertheless, for this research, written permission was obtained from Mr. Emad Hajjaj, the Mahjoob.com website creator and owner, in order to use the website for data collection (see Appendix A). Having discussed these issues, the actual process of building and annotating the corpus will be explored in the next few sections.
3.2.5 Building the corpus

Baker (2006) discusses various types of corpora and their potential uses. In one sense, the mahjoob.com corpus could be characterized as a specialized corpus since it is composed of texts from a specific genre of discourse i.e. web forum messages. The corpus may be considered specialized, too, in another sense because it represents texts from a single website.

The decision to investigate the various kinds of language and script on a single website has had implications for the construction of the corpus: collecting a purposive sample of texts from all available web forums on the English website was deemed desirable in order to provide an overall understanding of which scripts and languages were actually used on the website. In order to obtain such a sample, I enlisted the help of Sebastian Hoffmann who created and applied the Perl scripts which were used to collect the corpus data and prepare it for further analysis (Hoffmann, 2009b). In this regard, Hoffmann collected a sample of all messages posted in the time period spanning March 2007 until June 2008, roughly 14 months. In numeric terms, this meant downloading 21,626 threads saved as text files, containing a total of 460,220 messages from all 41 forums (see Chapter 5 for a description of these forums). This figure represented roughly 10% of all messages on the English website at the time of data collection.

Procedurally, the downloading of the website data, its conversion, and annotation into a corpus format involved several steps (Hoffmann, 2010). A brief overview of these steps is provided here followed by a more detailed account of each step. First, a Curl Perl module script was employed to download data from the website. Then, a second Perl script was used to convert the data from its html (and javascript) format to text format, removing boiler-plate material from the data in the process. Next, the same Perl script was employed to automatically annotate all stretches of
message text, including both quoted message material and new message material. This script was also used to differentiate between new message material and quoted message material and to annotate these accordingly. A further step involved annotating these stretches of message text with specific tags to indicate their script and linguistic code make-up. This also included the calculation of a ratio between items written in different scripts in biscriptal messages (see 3.3). The end result of this downloading, conversion, and annotation process was a set of files saved in UTF-8 text format so that they could be readily loaded into the WordSmith 5.0 corpus analysis software (see corpus folders and files in Appendix B) (Hoffmann, 2009b).

In terms of size, the individual thread files in the corpus range from as small as 1 or 2KB to several MB. During the process of collecting messages from the website described above, for the sake of inclusiveness, all message-containing threads occurring consecutively between URL 200,000 and 236,000 were accessed and downloaded. This ensured that no message-containing threads in the given time span of the sample of web data were left out, mitigating the risk that the absence of certain threads from the corpus might skew the corpus’ content.

At this point, it will be helpful to consider an actual example of a forum message which has been converted into .txt format (see Figures 3.2 and 3.3 below). Figure 3.2 displays a message from mahjoob.com in its original web browser format. Figure 3.3 shows the same message in

---

35 The programming script was further refined using frequency keyword lists in order to identify and label message content as exclusively one or more of the following linguistic codes: Arabic-scripted Arabic, Latin-scripted English, and Latin-scripted Arabic (see section 3.3).

36 Some threads in this range were found to be messageless. As a result, such threads were not included in the corpus.
converted and annotated format\(^{37}\). As Hoffmann observes, the simplified format of the converted message makes it very easy to analyse the message using corpus software such as Wordsmith, or even to examine it using simple Perl scripts (Hoffmann, 2009b, p. 7).

---

\(^{37}\) To illustrate, note that the message in Figure 3.2 displays “#10” in the upper right-hand corner, indicating that it is the tenth message within the discussion thread in which it was posted. The Figure 3.3 version provides the same information via the tag “<message n=’10’>”.
Despite their obvious differences, the actual message text shown in Figure 3.2 has been captured in Figure 3.3 (this text is enclosed by the tags <new> and </new>). Nevertheless, Figure 3.3 clearly resembles the kind of HTML source code one might expect to underpin the webpage data in Figure 3.2 minus the boilerplate elements. However, additionally, Figure 3.3 features several annotational tags that would not be found in the original HTML source code such as <code_9> which refers to the linguistic code in which the message was composed and <arabic_s> which refers to the type of Arabic found in the message (see Section 3.3 below).

Having discussed the downloading, conversion and annotation process, it will now be useful to consider the classification scheme employed to categorize the website data into various linguistic

---

38 In Figure 3.3, tagged elements such as "<smiley>blushingface</smiley>" indicate graphical smiley emoticons (e.g. 😊).
codes as alluded to by the tags <code_9> and <arabic_s> mentioned in the preceding paragraph. This annotational process and the resultant system will be described in the next few sections.

3.3 Developing the classification system

The mahjoob.com website, the data selected from it, and the building of the corpus based upon this data have just been described. It is now opportune to discuss in detail how the various linguistic codes present in the data were identified. This process was heuristic in nature, a result of effecting several conversions of the downloaded website data, each time with a view to detecting whether all distinctive linguistic codes had been captured as accurately as possible.

3.3.1 Method for developing the classification system

At a basic level, the classification system involved moving from larger distinctions between different language systems to ever finer levels of detail and nuance. To begin, besides the actual downloading of content from the website, it was deemed necessary to program the Perl script to distinguish strings of Latin script text and Arabic script text from one another within each downloaded message. Yet, distinguishing between scripts alone was insufficient in delineating the various linguistic codes featured on the website since not all Latin-scripted items were necessarily English. As a result, it became necessary to refine the corpus further. This process involved modifying the Perl script to assign specific labels to stretches of message text in order to indicate which language was used in the message. These tag labels are found in the header of each converted message. Thus, all messages composed entirely in the Arabic script were given the tag <code_1> while messages written exclusively in Latin script were all given the tag <code_2>.

However, as a result of hand-checking the <code_2> wordlist, it became clear that numerous words were not English at all. A particularly salient feature among many of these non-English
elements was the use of numerals as graphemes i.e. arithmographemes. Thus, the Perl script was further enhanced to automatically detect words that contained numerals in addition to letters. Messages that contained any number of arithmographemic words were tagged as <code_3>. However, through hand-checking the corpus again, it became clear that several examples of these were in fact false arithmographemes e.g. *MP3* (player), *gr8* (‘great). As described above, a wordlist of all ASCII tokens in the corpus was compiled, then all such words occurring eight or more times were hand-checked in order to identify the arithmographemic terms.\(^{39}\) Additionally, any terms that were non-arithmographemic but contained Latin-scripted Arabic were also identified. Thus, a list of non-arithmographemic Latin-scripted Arabic words was compiled that could also be used as a means of identifying messages that contained Latin-scripted Arabic words in the corpus. At the end of this process, the Perl script was able to detect and label messages as containing either exclusively Arabic-scripted words i.e. Code 1, exclusively English words i.e. Code 2, or at least some Latin-scripted Arabic words (both arithmographemic and non-arithmographemic) with some English words i.e. Code 3.

In order to further clarify the lexical content of texts in Arabic and Latin script, it was decided to make use of reference corpora. By using frequency wordlists derived from these reference corpora, it was possible to identify and label message content as Arabic-scripted Arabic, Latin-scripted English, Latin-scripted Arabic, or as various combinations of these. Procedurally, as a first stage, the Perl script was programmed to use a frequency wordlist derived from the British

---

\(^{39}\) The decision to not examine items that occurred less than eight times in the frequency list was deemed reasonable given that such a relatively low occurrence rate of these rare lexical items would only marginally impact the accuracy of the annotation of the corpus if they were indeed found to be false arithmographemic items.
National Corpus (BNC)\textsuperscript{40} in order to cross-check each message to determine if it contained only BNC types (i.e. unique words in a corpus). If so, the Perl script then annotated the message with a specific tag in its header that indicated that it contained only BNC words within it. Similarly, for Arabic, a frequency wordlist based on the Arabic Gigaword Corpus\textsuperscript{41} was used in order to identify and label each Arabic Gigaword-only message. For other linguistic codes such as Latin-scripted Arabic, a frequency wordlist of Latin-scripted Arabic was compiled from the

\textsuperscript{40}The British National Corpus (BNC) is a collection of 100 million words of both written and spoken British English from a variety of sources compiled between 1991 and 1994. While its texts are all from British sources, it does in fact contain items from non-British sources as well as some foreign language items and some written texts from the 1960s-1980s. The decision to make use of the BNC was motivated in part by the fact that British English has traditionally been the type of English taught in Jordan. Indeed, the current sovereign of Jordan, King Abdullah is British on his mother’s side. Also, given that CMC English has often appropriated language forms from spoken and slang models, the use of a corpus which includes such forms was deemed desirable.

\textsuperscript{41}The \textit{Arabic Gigaword 4\textsuperscript{th} Edition} is a corpus of primarily Modern Standard Arabic containing over one billion words compiled from newswire texts from nine different news agencies over a period of several years and including texts as recent as December 2008. While the main language of such texts is Modern Standard Arabic, the corpus authors do acknowledge that regional dialects and varieties will have had an impact on the corpus’ contents. Specifically, they acknowledge the potential presence of Egyptian, Lebanese, and Tunisian vernacular items. They also point out that a number of these news sources are based outside the Arabic-speaking world (Consortium, 2008). However, it is not clear to what extent the transliteration of items from languages such as English may also have had an impact on the corpus’ contents. The decision to use the Arabic Gigaword was pragmatic in that it is one of the largest corpora of Arabic in existence and its texts are relatively international in provenance. Given the Arabic Gigaword’s focus on Modern Standard Arabic it was expected that non-standard forms in the mahjoob.com corpus would be relatively easy to identify using the Arabic Gigaword as a filter. An alternative Arabic corpus such as the International Corpus of Arabic, which does include spoken forms, was not used primarily because of its very small size i.e. roughly 55,000 tokens.
mahjoob.com corpus itself by compiling a list of all types (i.e. unique words in a corpus) that contained numerals. These types were then hand-checked to ensure that they were, in fact, Arabic and not something else (see Appendix C file “arithmographemic_items”). But since not all Latin-scripted items contain arithmographemes, a comprehensive list of all non-arithmographemic non-BNC items was compiled in order to locate other Latin-scripted Arabic items. At this stage, an interesting issue emerged: several types were found that were clearly Latin-scripted Arabic but seemed to be lacking arithmographemes where they might be expected. In other words, several non-arithmographemic allographs (spelling variants) of typically arithmographemic items were discovered. For example, the arithmographemic Latin-scripted Arabic word shei5 /ʃεɪχ/ (meaning religious or political elder) was found in the corpus alongside a corresponding non-arithmographemic Latin-scripted Arabic allograph shaykh. Numerous spelling doublets such as these indicated that Latin-scripted Arabic was not uniform in orthography and suggested that, alongside an arithmographemic variety, a non-arithmographemic variety existed as well. Consequently, a further wordlist was compiled to distinguish frequent\(^{42}\) arithmographemic Latin-scripted Arabic types from non-arithmographemic Latin-scripted Arabic allographs (see Appendix C file “non-arithmographemic_items”). This non-arithmographemic Arabic-containing text was tagged <code_10> (<code_4>, <code_5>, etc. to 9, 12, and 13 were used for various biscriptal codes, see below).

\(^{42}\) By frequent here I mean occurred five or more times in a frequency wordlist of the whole corpus. At Hoffmann’s recommendation, a decision was taken to use five occurrences as a cut-off. This was motivated by a desire to be as inclusive as possible while not spending an inordinate amount of time on several thousand obscure items (Hoffmann, 2009a, personal communication).
A next stage involved identifying and labelling messages that were biscriptal in terms of which linguistic codes they were composed of and in what proportions. At the end of the process, it was imperative to check that the Perl script had consistently and accurately assigned the various code and script tags to different messages. In order to achieve this, a concordance of each code tag was compiled. In this way, all messages from each of the 14 linguistic codes (see Section 3.3.3) were collected into their own unique concordance. Once this was done, for most codes, each 60th message in a concordance of all messages labelled with a specific code tag was read and checked to see whether its linguistic content actually corresponded to the code label it had been assigned. In this way, the accuracy of the annotational system was checked. Based on this process of reviewing samples of messages at specific intervals, it was found that the automated annotational system functioned with great accuracy and consistency.

Admittedly, the classification system used here was heuristic in nature. Alternative methods could have been employed, which would surely have yielded different linguistic codes to investigate. For instance, one plausible classification system might have distinguished simply between Arabic-scripted, Latin-scripted, and mixed script texts, resulting in only three linguistic codes to compare. The advantage in this would have been to make analysis simpler by strictly

---

43 In the case of Codes 12 and 13, which have relatively few messages in the corpus, 5% of their respective messages were checked for consistency. This was achieved by calculating how many messages composed in each of these ‘minor’ codes accounted for 5% of all of its messages. Once this was done, this figure was used to determine the interval at which each code should be checked. For instance, in the case of Code 12, which has approximately only 120 messages, 5% meant six messages in total. Based on this figure, the interval for checking was determined to be every 20th message in the Code 12 message concordance.
limiting the number of linguistic variables. But such a simplification would have restricted the possibility of identifying texts that contained Latin-scripted Arabic texts from those that did not, a crucial feature of the corpus for addressing the second, third and fourth research questions of this study.\footnote{Although, as will be seen in Chapter 5, the relative paucity of biscriptal codes did, in fact, necessitate the conflation of these for statistical purposes.}

Another option would have been to ignore the distinction between BNC English-containing and non-BNC English-containing texts, or between arithmographemic and non-arithmographemic Latin-scripted Arabic. Again, this may have yielded fewer linguistic codes to contrast, but at the expense of being able to investigate more readily whether topical, stylistic, and discursive differences might exist between these seemingly near-identical codes. Ultimately, if for the sake of simplicity I had overlooked key lexical and orthographic differences between the codes as they are currently defined, it would have made it more difficult to explore their distinct orthographical and lexical properties, the very features that turn these codes into contrasting linguistic resources for their users. Indeed, despite the classification system's considerable complexity, maintaining the nuances between the various codes proved highly fruitful when examining the situated identity construction of a group of mahjoob authors (see Chapter 7). Further, such a nuanced system also leaves open the possibility of further detailed exploration of stylistic and lexical aspects of the corpus in future studies.
In brief, as a result of employing the classification system, it was easier to gain an overall picture of the various scripts and codes used on mahjoob.com. The next section will focus on what these scripts and linguistic varieties actually were.

### 3.3.2 Scripts and orthographies on mahjoob.com

The process of classifying messages as belonging to a specific script highlighted the fact that, with few exceptions, the website featured two primary writing systems, the Arabic and the Latin scripts. A decision was taken to concentrate on these two scripts because they are used to encode the vast majority of texts on the website, and because they appear to be used contrastively in non-random ways as parts of codes. But the focus in this research extends beyond script since script use is most meaningful when viewed in conjunction with lexicogrammatical and orthographic choices. This is so because script choice alone, while certainly not neutral, represents only one stylistic aspect of a given text.

In terms of orthographic systems, the website presents a number of descriptive challenges. For instance, on the one hand, there are clear examples of both Modern Standard Arabic and Classical Arabic orthography. Instances of the latter orthography are evident in the relatively numerous quotations from the Holy Qur’an found throughout the web forums. In stark contrast

---

45 There are, in fact, a few examples of other symbol sets and scripts on the website such as Cyrillic, Greek, and Chinese. These appear to be inputted via computer font types which can be selected by text composers. However, these other fonts and symbols are by and large not used consistently for long strings of text. Rather, they are featured in short strings such as user ids and signature lines. As such, they seem to highlight the uniqueness of a poster’s identity and are used by posters to perform a sort of graphic one-upmanship as they vie to create the most distinctive and complex user ids and signatures possible.

107
to this level of orthographic formality, there are several examples of non-standard Arabic orthography such as the doubling of consonants and vowels, practices that normally do not occur in formal written Arabic (see Section 3.3.3.1 below). Another conspicuous example of a non-conventional orthography for Arabic is the use of the Latin script with arithmographemes, which help to represent Arabic phonemes that have no traditional graphemic equivalents in the Latin script.

In light of these observations, I contend that, due to the highly unregulated nature of the website, such distinctly non-standard orthographies are able to flourish. Novel orthographies occur in conjunction with both the Arabic and Latin scripts and with the Arabic and English lexicogrammars. Ostensibly, they can be used strategically to index different styles (see Lee, 2007; Su, 2003; also Chapter 7 below).

Given this relative heterogeneity, one might be led to believe that mahjoob.com is characterized as an orthographic free-for-all. Yet despite this, there are still some broader orthographic patterns of use i.e. ‘unconventional’ conventions followed by forum posters. As Sebba (2007) observes, while non-standard orthography may be markedly deviant from the norm, it is not entirely arbitrary, and pre-existing standardized forms are adapted in more or less logical and comprehensible ways. In fact, one of the aims of this research is to uncover how non-standard orthographies are systematically employed, especially in forum participants’ attempts to construct distinct styles and identities (see Chapter 7).

Now that the system for identifying the various script, lexicogrammatical, and orthographic configurations observable in the corpus has been described, it is time to consider the actual
linguistic codes that were defined as a result of this process. The next two sections will present the characteristics of each of these codes and discuss the basic lexicogrammatical script pairings i.e. Arabic-scripted Arabic, Arabic-scripted English, Latin-scripted English, and Latin-scripted Arabic, which the 14 linguistic codes ultimately consist of.

3.3.3 The 14 Linguistic codes
Having discussed the iterative method that allowed a picture of different code and script configurations to emerge from the data, a total of 16 codes were ultimately identified in the corpus, 14 of which could be labelled linguistic scriptal and two of which could be labelled non-scriptal. The scriptal codes can further be categorized as either mono-scriptal or biscriptal based on whether they contained only one script (Latin or Arabic) or a mixture of both of these scripts respectively. Table 3.1 below provides examples of all 16 codes:
Table 3.1: The 16 codes in the corpus

<table>
<thead>
<tr>
<th>Code No.</th>
<th>Code Name</th>
<th>Example</th>
<th>English Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>mono-scriptal: Arabic script only</td>
<td>من زمان انا هناك بس اللون اختط شوي بتسألي عن</td>
<td>I have been here for a long time but the colour is a little different- are you asking about the signature?</td>
</tr>
<tr>
<td>2</td>
<td>mono-scriptal: British National Corpus (BNC) items only (no Arithmographemic Latin script)</td>
<td>I am not there too, i run away when they were talking about another diversion coming &lt;smiley&gt;eyeswatering&lt;/smiley&gt;</td>
<td>N/A</td>
</tr>
<tr>
<td>3</td>
<td>mono-scriptal: Arithmographemic Latin script (with some non-Arithmographemic Latin script)</td>
<td>ana ba3raf nzelet 3al neyabeh 3endak ta6alo3at 3al 7okomeh mashy ya 5ayen &lt;smiley&gt;eyeswatering&lt;/smiley&gt;</td>
<td>I know you arrived in the governmental department (Niyaba) and you thought you were above your government ok traitor</td>
</tr>
<tr>
<td>4</td>
<td>biscriptal: Arabic script and non-Arithmographemic Latin script (bias Latin script)</td>
<td>Miss Engineer I'm glad you liked it.. take lessons &lt;smiley&gt;offersflower&lt;/smiley&gt; Hatsoo &lt;smiley&gt;huggingfriend&lt;/smiley&gt; Tayobeh I will put up a very big fight ... &lt;smiley&gt;Boxing&lt;/smiley&gt;</td>
<td>Miss Engineer I'm glad you liked it.. take lessons Hatsoo live and remember, Tayobeh, I will put up a very big fight ...</td>
</tr>
</tbody>
</table>

46 In Table 3.1, Column 3, entitled “English Translation”, underlined text indicates that the translation was made from an original Latin-scripted Arabic stretch of text. In contrast, bold text indicates that the translation was made from original Arabic-scripted Arabic stretch of text. However, user IDs such as Shafalla7 and Tayyeb Lesh remain untranslated. Instead, these user IDs are italicized in the translation column.
<table>
<thead>
<tr>
<th>Code No.</th>
<th>Code Name</th>
<th>Example</th>
<th>English Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>biscriptal: Arabic script and Arithmographemic Latin script (bias Latin script)</td>
<td>mbrook emi! yetraba be3ezzik :) محبي..</td>
<td>Congratulations, Mom! May he be raised well with love</td>
</tr>
<tr>
<td>6</td>
<td>biscriptal: Arabic script and non-Arithmographemic Latin script (bias Arabic script)</td>
<td>تشلسي vs نيفربول</td>
<td>Chelsey vs. Liverpool</td>
</tr>
<tr>
<td>7</td>
<td>biscriptal: Arabic script and Arithmographemic Latin script, bias Arabic</td>
<td>لا يا أخت لوليتا صباح فخري خليلك اياه منورة Shafalla7 ان شاء الله دايم مبسوطة وبتضحكي</td>
<td>No, sister Lolita sabah fahkry made it glow God willing Shafalla7 will always be happy and laughing in the presence of all of you.</td>
</tr>
<tr>
<td>8</td>
<td>biscriptal: Arabic script and non-Arithmographemic Latin script, equal</td>
<td>هلوسات شيطانية...فجرتها صورة Amazing! Simply amazing! I can't wait to see who wrote it!</td>
<td>the devilish hallucination… shocking pictures Amazing! Simply amazing! I can't wait to see who wrote it!</td>
</tr>
<tr>
<td>9</td>
<td>biscriptal: Arabic script and Arithmographemic Latin script, equal</td>
<td>Lolita تسلمي .. وبسموا لي دعواتك أحمد &lt;smiley&gt;sidewaysblushingsmile&lt;/smiley&gt; شفرا shafalla7 I like your name :D &lt;smiley&gt;offersflower&lt;/smiley&gt;</td>
<td>Lolita, thank you… and thank them all for your prayers Ahmed. &lt;smiley&gt;sidewaysblushingsmile&lt;/smiley&gt; thanks Shafalla7 I like your name :D.</td>
</tr>
<tr>
<td>Code No.</td>
<td>Code Name</td>
<td>Example</td>
<td>English Translation</td>
</tr>
<tr>
<td>---------</td>
<td>-----------</td>
<td>---------</td>
<td>---------------------</td>
</tr>
<tr>
<td>10</td>
<td>mono-scriptal: Non-BNC/Non-Arithmographemic Latin script only (with non-Arithmographemic Latin-scripted Arabic frequency list items)</td>
<td>heey metal fever thnx alot dear i’m really happy that i started to write in arabic its great feeling wallah hehehe nice title thnx alot &lt;smiley&gt;happyfacesmallsmile&lt;/smiley&gt;</td>
<td>heey metal fever thnx alot dear i’m really happy that i started to write in arabic its great feeling really hehehe nice title thnx alot &lt;smiley&gt;happyfacesmallsmile&lt;/smiley&gt;</td>
</tr>
<tr>
<td>11</td>
<td>biscriptal: Arabic script and Non-BNC/Non-Arithmographemic Latin script, bias Arabic</td>
<td>Teyeb lesh, i’m here &lt;smiley&gt;nananananafacet&lt;/smiley&gt; أحيانا ، يجعلنا الحزن نكتب... أحيانا يكون أكبر من أن يكتب .. فتصمت لكن هذا لا يعني إبدا ، انا لستنا هذا انتم في القلب، وإن لم يكتب القلم شيئا لكي تحاتي &lt;smiley&gt;offersflower&lt;/smiley&gt;</td>
<td>Teyeb lesh, i’m here &lt;smiley&gt;nananananafacet&lt;/smiley&gt; sometimes the sadness makes us write....sometimes it’s bigger than our writing...it’s quiet but not for ever, we are not here and you are in the heart, and if the pen did not write anything to live&lt;smiley&gt;offersflower&lt;/smiley&gt;</td>
</tr>
<tr>
<td>12</td>
<td>biscriptal: Arabic script and Non-BNC/Non-Arithmographemic Latin script, bias Latin</td>
<td>yes IF sadeg walla mm I go for Happiness, this standard you can create it whenever wherever you live, التي يطلع لوق ينكسر قفته any one can manage his own life and let it be happy in the way he wants it to be.</td>
<td>yes IF he is honest or not I go for Happiness, this standard you can create it whenever wherever you live. The one who climbs up (in life) will break his neck. anyone can manage his own life and let it be happy in the way he wants it to be.</td>
</tr>
<tr>
<td>13</td>
<td>biscriptal: Arabic script and Non-BNC/Non-Arithmographemic Latin script, equal</td>
<td>لماذا نفتش عن أغنيات البكاء بديوان شعر قديم؟ ونسأل: يا حبي! هل تدوم؟ أحبك حب الأموات واحة obesity وحرب الفقير الغيف i dont mind inno yseer nizareyyat, bas hay mahmoud darweesh lal taghyeer &lt;smiley&gt;greentoothyismile&lt;/smiley&gt;</td>
<td>Why do we search for sad songs without old poetry? And we ask: our love! Will it last? I love you like the caravan beasts love their grass and water and the poor loves bread! i dont mind Nizareyyat (songs) but that Mahmoud Darweesh (stuff) has to go &lt;smiley&gt;greentoothyismile&lt;/smiley&gt;</td>
</tr>
<tr>
<td>Code No.</td>
<td>Code Name</td>
<td>Example</td>
<td>English Translation</td>
</tr>
<tr>
<td>---------</td>
<td>-----------------------------------------------</td>
<td>----------------------------------</td>
<td>---------------------</td>
</tr>
<tr>
<td>14</td>
<td>mono-scriptal: Non-BNC/Non-Arithmographemic</td>
<td>I just loved this one! amazing</td>
<td>I just loved this</td>
</tr>
<tr>
<td></td>
<td>Latin script only (contains BNC English plus</td>
<td>really.. thank you JT &lt;smiley&gt;</td>
<td>one! amazing really</td>
</tr>
<tr>
<td></td>
<td>non-BNC items such as abbreviations,</td>
<td>huggingfriend&lt;/smiley&gt;</td>
<td>.. thank you JT</td>
</tr>
<tr>
<td></td>
<td>misspellings, neologisms, Latin-scriptsed</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Chechen, Circassian, Hebrew, etc.)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1557</td>
<td>Non-alphabetic content i.e. numbers, or</td>
<td>17% &lt;smiley&gt;happyfacesmallsmile</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td>punctuation with or without smileys (e.g.</td>
<td>&lt;/smiley&gt;</td>
<td></td>
</tr>
<tr>
<td></td>
<td>“$300” or “???”)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1648</td>
<td>non-text elements only such as graphics and</td>
<td>&lt;smiley&gt;pointinglaughing&lt;/smiley&gt;</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td>smileys</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Now that example messages of each of the 16 codes have been provided, it will be useful to discuss the basic linguistic variables which underpin these codes. These can be divided into two broad categories: mono-scriptal codes and bисcriptal codes, where the latter are derived from

---

47 This was originally a subdivision of Code 2 (BNC English).

48 This was annotated as <code_0> in the corpus.
combinations in various amounts of the former. The key mono-scriptal codes discussed here are Arabic-scripted Arabic, Latin-scripted English, Latin-scripted Arabic, and Arabic-scripted English.

3.3.3.1 Arabic-scripted Arabic

The Perl script mentioned above was used to distinguish all Arabic-scripted content in the corpus from Latin-scripted content. A unique tag was used to refer to such Arabic-scripted content, <code_1>. Essentially, any message was labelled as Code 1 if it was found to contain only Arabic-script with the exception of Latin script used in hyperlink strings. Thus, a message which contained a file name such as *php.arabic_news.html* would still be classified as Code 1. Code 1 messages were then further annotated to indicate what types of Arabic they contained. In order to achieve this, the Arabic Gigaword frequency word list was used. This meant that each Code 1 message was automatically scanned with the annotating Perl script and cross-checked against the Arabic Gigaword wordlist to see whether it contained only Arabic Gigaword content or not. If a message was found to contain only Arabic Gigaword lexis, it was assigned the tag <arabic_s>. If an Arabic-scripted message contained no Arabic Gigaword lexis, it was tagged as <arabic_v>. If it contained both Arabic Gigaword and non-Arabic Gigaword items, it received the tag <arabic_x>. Finally, a fourth tag, <arabic_0> was applied to all messages which contained no Arabic script whatsoever.

As discussed earlier in this chapter, the Arabic-scripted Arabic featured on mahjoob.com could not be said to be of one single variety or style. For instance, discussions are often carried out in written Vernacular Arabic as opposed to Modern Standard Arabic which is the normative code for virtually all traditional written discourse in offline contexts (Holes, 2004). In order to
differentiate such diglossic varieties in my analyses, I employ the term Arabic-scripted Vernacular Arabic for the written vernacular where necessary to distinguish this from the more conventional Arabic-scripted Modern Standard Arabic (see Al Share, 2005). On the surface, Arabic-scripted Vernacular Arabic appears relatively close to written Modern Standard Arabic\(^49\). Yet in CMC contexts, Arabic-scripted Vernacular Arabic is quite different from Modern Standard Arabic especially in terms of lexicogrammar, though overlapping forms are abundant. Indeed, because of the considerable lexical overlap between Modern Standard Arabic and Vernacular Arabic, it can sometimes be difficult to determine whether a string is actually Vernacular Arabic or Modern Standard Arabic, though examination of the larger context usually makes this clear. This stylistic ambiguity is further compounded by the ability of text-producers to code-switch between these two varieties. Nevertheless, as a generally observable trend in the data, the occurrence of Arabic-scripted Arabic often indicates the use of Arabic-scripted Vernacular Arabic. And where Modern Standard Arabic forms contrast stylistically with Vernacular Arabic forms in the data, they will be noted (see Chapter 7).

3.3.3.2 Latin-scripted English

Code 2 (BNC English) was used to identify messages which contained only Latin script elements which were also found in the 100 million-word British National Corpus of English (BNC). The BNC was discussed earlier but at this point it is worth adding that the BNC contains 90% written and 10% spoken data. The majority of the written data is from previously published material, although about 4 and a half million words of the BNC consist of unpublished written data, of

\(^49\) Previous studies undertaken by Warschauer et al. (2002), Palfreyman and Al Khalil (2003), and Al Share (2005) indicate that, similar to other standardized written languages, Arabic CMC texts exhibit highly unconventional orthography (see also Sebba, 2007; Hinrichs, 2005; Nishimura, 2003).
which 214,000 words are from emails. So although most of the data in the BNC represents standard written British English, not all of it does. However, as the BNC was compiled in the early 1990s, during the early days of Internet usage, the electronic mailing data included in the corpus do not make use of smileys or other conventions that are currently popular in CMC, such as \textit{gr8} (which does not occur in the BNC). Consequently, the early CMC texts in the BNC tend to conform to written English. Thus, to a large extent, Code 2 represents general written and spoken British English, as opposed to current CMC English\textsuperscript{50}.

In this regard, it is important to mention the existence of another Latin script-based code namely, Code 14 (Non-BNC English). Code 14 was used to label text that contained any number of Latin-scripted items extraneous to both the BNC and the Code 10 (Non-arithmographemic Latin-scripted Arabic with English) wordlists. These items typically consisted of neologisms and post-BNC Netspeak such as \textit{rapidshare} and \textit{gr8} and proper names like \textit{Sarkozy} that would have been absent from 1990’s British English as well as very obscure Latin-scripted Arabic such as \textit{sadaaqaat} (see Chapter 7) and words from other languages such as Circassian e.g. \textit{adiga} 'Circassian', Turkish \textit{domna} 'Ottoman crypto-Jewish converts to Islam', Hebrew e.g. \textit{ma’ale edumim}, a Jewish settlement outside of Jericho, etc. Based on the slight differences between these two codes, I initially considered conflating them into a single code. However, I decided to keep Code 14 separate from Code 2 in case its texts might reveal distributional and stylistic patterns different from those of Code 2, points borne out in Chapter 5 and Chapter 7 respectively. Indeed, a Chi-squared test of significance revealed that Code 14's distribution patterns are

\textsuperscript{50} Hoffmann (2010), however, notes that the texts of the BNC exhibit a wide range of formality and standardness.
significantly different to those of Code 2 (see Chapter 5). Nevertheless, due to Code 14’s relative obscurity as well as space considerations, it is not treated in-depth in this study.

### 3.3.3.3 Latin-scripted Arabic

Auer (1998, 2008) deprecates a monolingual bias in code-switching research which makes a priori assumptions about the existence of distinct and discrete linguistic systems which are then mixed in the speech of bilinguals to produce code-switching. Instead, Auer posits the possible existence of mixed codes or “fused lects” as the normative code of interaction among certain groups (Auer, 1998). Such a categorization naturally blurs the lines between discrete linguistic varieties. Prior research into Latin-scripted Arabic seems to point toward the existence of such a hybridized form of language which, while using the Latin-script, incorporates lexicogrammatical elements from both Arabic and English (Abdallah, 2008; Al-Tamimi & Gorgis, 2007; Al Share, 2005; Palfreyman & Al Khalil, 2003; Sakarna, 2006; Warschauer et al., 2002). Further, most of these studies also point to a unique feature of this hybrid language, namely the use of numerals as graphemes in order to represent Arabic sounds which have no ready or widely agreed upon equivalents in the Latin alphabet (Abdallah, 2008; Al-Tamimi & Gorgis, 2007; Al Share, 2005; Palfreyman, 2001a; Palfreyman & Al Khalil, 2003; Warschauer et al., 2002). Thus, the arithmographemics of CMC-based Latin-scripted Arabic typically include the arithmographemes: 2, 3, 5, 6, 7, 8, and 9 and digraphs such as ‘3’, ‘6’, ‘7’, and ‘9’

---

51 Variants of these digraphs also exist: some Latin-script Arabic users place the apostrophe after the arithmographeme i.e. 3’, 6’, 7’, 9’, or use periods instead of apostrophes either before or after the arithmographeme i.e. .3, .6, .7, .9, and 3., 6., 7., and 9.
Bearing in mind this widespread characteristic of Latin-scripted Arabic as attested in the literature, it was decided to detect stretches of Latin-scripted text, containing this arithmographemic lexis as a means of identifying Latin-scripted Arabic. However, as mentioned above, not all Latin-scripted Arabic contains such arithmographemics. Consequently, in order to capture these details, a wordlist of non-arithmographemic Latin-scripted Arabic was compiled based on a Perl script-generated wordlist of all non-BNC lexical items (see Appendix C file “non-arithmographemic_items”). In this way, it was possible to scan the entire corpus and annotate messages which featured non-arithmographemic Latin-scripted Arabic.

During this process another peculiarity was discovered. There were several cases of items which did not contain arithmographemics despite the presence of phonemes in these items that would normally require the use of arithmographemes. These were words such as Shaykh and ‘Ali which would normally be rendered shei5 and 3ali in normal arithmographemic Latin-scripted Arabic. Another trait of this non-arithmographemic variety of Latin-scripted Arabic was the extensive use of digraphs to indicate long vowels which occur in Arabic: allaah instead of allah, deen instead of din. Given this stylistic difference, as mentioned in the preceding paragraph, a frequency wordlist was compiled of all such non-arithmographemic variants of Latin-scripted arithmographemic words. Using this wordlist it was then possible to annotate message texts in the corpus which contained this non-arithmographemic lexis with their own tag. This special tag was <code_10>. So, in brief, unlike Code 3 discussed above, Code 10 refers to Latin-scripted text which, although featuring a mixture of Latin-scripted English and Latin-scripted Arabic lexemes, contains no arithmographemics whatsoever and is also characterized by extensive use of doubled vowels to indicate long vowels. Such non-arithmographemic Latin-scripted Arabic is
not without precedent: several Islam-related websites such as ummah.com feature a similar variety of Latin-scripted Arabic (see Commins, 2006; Mujahid, 2009).

3.3.3.4 Arabic-scripted English

Arabic-scripted English describes the English lexicogrammar paired with Arabic script. There is a longstanding tradition of borrowing and transliterating foreign names and terms into written Arabic. Such transliterated forms are widely attested outside CMC contexts. Script-borrowing is also widely attested for the Arabic script: Farsi, Urdu, Pashto, Turkish, Malay, Swahili, Somali, Hausa, and Medieval Spanish have all been or are currently being written using Arabic script. At present, the widespread influence of English has occasioned several borrowings into Arabic. However, the use of the Arabic script to transliterate English i.e. Arabic-scripted English appears to be a very recent development and small-scale phenomenon. With reference to the present research, Arabic-scripted English’s occurrence seems to be very limited and marginal when compared with the other codes found on mahjoob.com. In this regard, it serves mainly for nonce borrowings of terms which have no ready equivalents in the Arabic language, especially from the realm of computers and CMC. Consider these examples taken from the forums: اون لاين [āūn lāīn] ‘online’, اوف لاين [āūf lāīn], ديليت [dīlīt] ‘delete’, and رووم [rūūm] ‘room’ as in chatroom.

The last example is particularly interesting from an orthographic point of view. The form “رووم” transliterated [rūūm] even mimics English orthography in its use of two consecutive rounded back vowels ‘و’ modelled on the double ‘o’ in the English source word “room”. As such, it seems to be an example of visual/orthographic orthography, in contrast to the first three Arabic-scripted English samples which present a more phonetic orthography in an attempt to accurately convey how these borrowings should be pronounced by readers of Arabic (see Tseliga, 2007, pp. 118-119). Besides such domain-related nonce borrowings, the forum also contains referential
and vocative forms of Arabic-scripted English e.g. [‘āūld wən] ‘old one’, [mān] ‘man’.

Finally, apparently for humorous effect, the website also contains short Arabic-scripted English strings such as [āī dūnt rīməmbər] ‘I don’t remember’. And although such forms are relatively infrequent, the relevance of Arabic-scripted English to the present study should not be overlooked as it may be found to play a unique stylistic role (see Su, 2003). Given this relative paucity, Arabic-scripted English was not identified with its own code designation. Rather it was subsumed under Code 1, the tag for Arabic-scripted Arabic. Nevertheless, using a frequency wordlist derived from the Arabic Gigaword corpus, it was possible to annotate Code 1 texts to indicate any items which were written in Arabic script but were not found in the Arabic Gigaword corpus. As pointed out above in section 3.3.3.1, these texts were annotated with a secondary tag <arabic_v>. However, as will be seen in Chapter 5, these elements were found to be extremely rare in the corpus.

Having discussed Codes 1, 2, 3, 10, and 14, the principal mono-scriptal codes, it will be beneficial to consider the remaining codes used in the classification system of the corpus. These are the biscriptal codes which were defined as various combinations, in different proportions of, the mono-scriptal codes.

3.3.3.5 Biscriptal codes

Codes 4, 5, 6, 7, 8, 9, 11, 12, and 13 are unique in that they all contain varying proportions of both Arabic-scripted and Latin-scripted items. Code 4 was defined as text containing more than twice as many Latin-scripted items, such as Code 2 (BNC-English) items, than Arabic-scripted items, i.e. items typical of Code 1. As such, Code 4 texts tend to resemble English with small amounts of Arabic text interspersed. Code 5 is similar to Code 4 in composition except that its
Latin-scripted items also contained arithmographemic Latin script i.e. the Code 3 lexis. As such, Code 5 texts tend to resemble Code 3 texts, i.e. Latin-scripted Arabic mixed with English, but with the addition of small amounts of Arabic-scripted items.

Code 6 texts, on the other hand, look much more like Code 1 (Arabic-scripted Arabic) texts as they are composed of more than twice as many Arabic-scripted items than Latin-scripted items (bias=Arabic). So, in addition to Arabic-scripted elements, Code 6 texts also feature smaller amounts of Code 2 (BNC English). In contrast to Code 6, Code 7 texts also resemble Arabic-scripted Arabic but with small amounts of Code 3 (Arithmographemic Latin-scripted Arabic) items included.

Code 8 is similar to both Code 4 and Code 6 insofar as it contains both Arabic-scripted items and Non-arithmographemic Latin-scripted items from Code 2. However, Code 8 differs from both Code 4 and Code 6 due to the fact that it contains a mixture of Arabic-scripted and Latin-scripted items where neither of these scripts accounts for more than twice as many items as the other script (bias=balanced). Code 9 is similar to Code 8 in its ratio of script composition except that Code 9 contains arithmographemic Latin-scripted arithmographemic items typical of Code 3.

Codes 11, 12, and 13 are composed of Arabic-scripted items (typical of Code 1) interspersed with non-arithmographemic Latin-scripted items (typical of Codes 10 and 14). The difference between Codes 11, 12, and 13, however, is in their ratios of Latin-scripted to Arabic-scripted items. Code 11 contains more than twice as many Arabic-scripted items as Latin-scripted items (bias=Arabic), while Code 12 contains more than twice as many Latin-scripted items as Arabic-scripted items (bias=Latin). In contrast, Code 13 contains a mixture of Arabic-scripted and Latin-
scripted items where neither of these scripts accounts for more than twice as many items as the other script (bias=balanced).

3.3.3.6 Non-scriptal codes
In addition to the mono-scriptal and biscriptal codes, two other types of texts were discovered in the corpus. These consisted of texts which contained no specifically Arabic or Latin-scripted items. The first of these, initially tagged as <code_0> refers to stretches of text which are devoid of any linguistic content other than hyperlinks or graphic content such as smileys. Eventually, however, Code 0 was renamed Code 16 in order to avoid giving the impression that this type of text had no content whatsoever. The second type of non-scripted text was labelled Code 15. Code 15 texts featured content which was composed of punctuation and/or numerals exclusively. It should be mentioned that these tags were applied to thread titles, new message content, and quoted message content.

Now that all 16 linguistic codes have been described, it is appropriate to discuss the other tags that were developed and applied to the corpus in order to facilitate its analysis. This will be the focus of the next few sections.

3.4 Annotating the corpus
3.4.1 Meta-data
The classification system which identified the 16 distinct codes into which each message in the corpus could be categorized has just been described. Yet, identifying the various linguistic codes in the corpus was not sufficient to allow for a comprehensive investigation of all the potential patterns of code-switching and script-switching which occur on the website. For instance, code and script-switching are not only observable within and between messages, but these phenomena can also be investigated across a wide variety of variables: forums, threads, authors, author
locations, and even dates and times of posting. Nor are message texts the only loci for code-switching and script-switching on the website. Indeed, forum titles, discussion thread titles, author IDs, author signature lines, and quoted message material all provide opportunities for forum contributors to both code-switch and script-switch. Capturing such meta-data was considered highly desirable because of the possibilities it afforded for deeper analyses of code-switching and script-switching. While not all such variables can be examined in equal depth in this study, at the very least it was deemed useful to record as much of this circumstantial data as possible in order to provide a more complete context for analysing specific instances of code and script-switching in the corpus. Consequently, further enhancements to the annotation system had to be made, resulting in the creation of several different tags.52

Although these categories often overlap, broadly, the meta-data tags applied to the corpus can be grouped into three main types: 1) structural, 2) linguistic, and 3) contextual. In order to illustrate the function of these types of tags, an example of an entire thread from the corpus featuring virtually all meta-data tag types used in the corpus is provided below (Figure 3.4). Continual reference will be made to these same tags in the description of their role and functions which follows.

**Figure 3.4: Thread 210028, “Song Name” from the mahjoob.com corpus**

52 Due to the highly technical nature of the tagging used in the original mahjoob.com forum web page source files, it was preferable to create a whole new set of tags in order to annotate the corpus in such a way as to facilitate subsequent analysis of the data. These new tags were all designed according to the conventions of the Extensible Markup Language (XML), the most commonly used language for encoding programs for source files of web pages.
there is a song for Asala i dunno what's the name <smiley>wideeyedperplexed</smiley>

what's the name of this song <smiley>wideeyedperplexed</smiley> Thanks moa8adaman <smiley>greentoothysmile</smiley>

---

there is this part ألبتي إنجرح ، ما عاد شي يجرحو ، صاير حجر حطيت ألبك مطرحو

what's the name of this song <smiley>wideeyedperplexed</smiley> Thanks moa8adaman <smiley>greentoothysmile</smiley>

---

there is the 3 songs with the same name , not the new one (u will find 2 songs live and solo with this name this is the one) enjoy <smiley>rollingeyes</smiley>
For the sake of space, Thread 210028, a very short thread containing only two messages, has been selected for discussion here. Despite its small size, this thread demonstrates the ability of authors to quote earlier messages in the mahjoob.com forums. In this case, solio, the author of Message 2 actually quotes Message 1 within his own message. This thread also exhibits another phenomenon not witnessed in all messages: the first message, as it is composed in Code 5, provides an authentic example of script-switching.

Returning to the three different categories of tags mentioned above, structural tags refer to tags created in order to structure the data so as to facilitate corpus-based searches and to identify and distinguish different sets of data within the corpus. This category includes the consistent use of header and end tags in order to create tag pairs for all data in the corpus. An example of such a tag pair from the corpus is the ubiquitous <content>…</content> tag pair which encloses

---

53 Although in reality, all three categories of tags were indispensable in enabling corpus-based searches using WordSmith 5.0 software.
message content data in each message in the corpus, setting this textual content apart from other message-related data such as author ID and code (i.e. linguistic composition) of the message.

Some structural tag pairs were designed in order to distinguish new message content from quoted message content e.g. <new>…</new> versus <quote id="Flower^Angel">…</quote id="Flower^Angel">^54.

Examples of other types of structural tags were the identifying tags used at the beginning of each thread such as

```
<thread n="210028">
<forum n="27" name="Entertainment">
<thread_title name="song name"/>
```

These four tags provide information about the thread file number, the original URL from which the thread was downloaded, the forum in which the thread occurs, and the thread’s title respectively. Another important structural annotation is the <message n=…> tag which indicates the order of the message in the original discussion thread. Thus, in the above thread, it is seen that the two messages are actually Message 1 and Message 2 in this thread i.e. <message n="1"> and <message n="2">. These message order tags were particularly helpful in tracing the sequence of discourse in a thread and the development of a discussion.

In contrast to structural tags, linguistic tags shed light on the actual linguistic properties of the textual data captured in the corpus. Such tags indicated details like the code and script

^54 It should be mentioned that quoting prior messages is not in fact mandatory in posting subsequent messages and that only one third of all messages in the mahjoob corpus actually quote directly from other messages within the corpus (see Chapter 5).
composition of message data e.g. `<code_5>` and `<script id="mixed" ratio="0.477777777777778" bias="latin">` in Message 1 above. The latter tag indicates that Message 1 contains more Latin script than Arabic at a ratio of slightly more than two Latin-scripted items to one Arabic-scripted item i.e. the Arabic-scripted items are less than half as frequent as the Latin-scripted items in this message. Indeed, smileys aside\textsuperscript{55}, there are 23 Latin-scripted items to 11 Arabic-scripted items in Message 1. Message 2 does not display such script ratio information simply because it is not a biscriptal message as indicated by its script tag, `<script id="latin">` above.

Some linguistic tags indicated features such as the presence and type of Arabic used in a message e.g. `<arabic_x>` (i.e. Arabic Gigaword Arabic-scripted lexis mixed with non-Arabic Gigaword Arabic-scripted lexis) in Message 1 and `<arabic_0>` (i.e. absolutely no Arabic-scripted items present) in Message 2 above. The presence of arithmographemic elements within each message was also indicated using the following tags: `<AG id="yes">` and `<AG id="no">`. As can be seen above, both Message 1 and Message 2 contain arithmographemes as evidenced by the tag `<AG id="yes">` observable in each of these annotated messages. Such tags also offered the added benefit of providing a very convenient means of verifying the accuracy of the Perl script’s automated linguistic code detection and annotation process. For instance, by definition, `<arabic_0>` tagged messages could not also be tagged `<code_1>`, `<code_4>`, `<code_5>`, `<code_6>`, `<code_7>`, `<code_8>`, `<code_9>`, `<code_11>`, `<code_12>`, or `<code_13>` because these tags refer to texts which must contain at least some Arabic-scripted content. Similarly, `<AG id="yes">` could not be assigned to messages which were identified as belonging to Codes

\textsuperscript{55} Although annotated and renamed using Latin script, smileys are omitted from the ratio calculation of biscriptal texts since they do not count as a form of conscious script-switching on the part of text composers.
1, 2, 4, 6, 8, 10, 14, because, by definition, such codes were devoid of arithmographemic elements. Indeed, as much as possible, such potential annotational conflicts were looked for, and whenever found they were addressed by modifying the Perl script until the inconsistencies were rectified. This process of double-checking seemingly redundant tags ultimately resulted in a more accurate set of meta-data and a more robust corpus.

It was mentioned earlier in this section that messages were not the only loci of code and script-switching phenomena in the corpus. Since earlier research into code-switching in web forums seems to indicate that prior texts (especially seed messages i.e. initial messages in discussion threads) can in fact have an impact on code choice in subsequent messages (see Wodak & Wright, 2007) it was decided that such situational framing texts (see Gumperz, 1982) should also be annotated to indicate their own linguistic composition. Consequently, tags for thread title code e.g. <title_code id="2">, seed message code e.g. <seedMessage_5>, and quoted message code <codeQuote_5> were also designed and applied to the corpus data. Thread titles were further annotated for script and arithmographemic content with the tags <title_script id="latin"> and <title_AG id="no"> respectively. Quoted message content received the same set of annotation tags as new message content, however, quoted message content was annotated separately from the non-quoted message content in a given message and the proportions of codes found in these two different parts of the message were calculated independently of each other. To illustrate, consider that Message 2 above contains quoted message content from Message 1. Notice that Message 2 is annotated with the following set of tags:

```
<scriptQuote id="mixed" ratio="0.477777777777778" bias="latin">
<AG_Quote id="yes">
<codeQuote_5>
<arabicQuote_x>
```
These four tags mirror the following four tags found in Message 1:

```
<script id="mixed" ratio="0.477777777777778" bias="latin">
<AG id="yes">
<code_5>
<arabic_x>
```

This is due to the fact that the Perl script annotated the quoted material in Message 2 (which is actually derived from the new message material in Message 1) in the exact same way. The only difference between the two sets of tags is that the ones which apply to the quoted material have the word “Quote” inserted in them. Such parallel tagging helped at a glance to indicate whether the linguistic code of new message material was similar to the linguistic code of quoted material within the same message, shedding possible light on another proposition of situational code-switching, namely, that bilinguals will tend to respond to an earlier utterance using the same language (see Blom & Gumperz, 1972).

Beyond purely linguistic annotational tags, there was also a need to create a pair of paralinguistic tags, namely, the `<smiley>…</smiley>` tag pair. These tags were used to indicate the presence of smileys embedded within message texts. For instance, in Message 1 above, the author uses two different smileys a total of three times: `<smiley>wideeyedperplexed</smiley>` is used twice, while `<smiley>greentoothysmile</smiley>` is used only once. Message 2’s author, however, has employed only one type of smiley: `<smiley>rollingeyes</smiley>`, which he has used only once.\footnote{\textsuperscript{56} It should be noted that while extremely commonplace in the corpus’ messages, especially in Code 3 texts (see Chapter 6), the use of smileys is by no means compulsory for text composers and thousands of messages in the corpus are altogether devoid of smileys. A list of the smileys encountered in the corpus and their graphic equivalents is found in Appendix D.}
The final group of tags used to annotate the corpus was contextual tags. These tags were conceived of out of a desire to provide as much contextual detail as possible about the posted messages. Using Messages 1 and 2 respectively to illustrate once again, data such as author IDs e.g. `<author id="Flower^Angel">` and `<author id="solio">`, author locations e.g. `<location id="UAE">` and `<location id="Unexpected places">`, and date and time of the posting of the message e.g. `<date id="11th July 2007, 06:29:30 AM">` and `<date id="11th July 2007, 08:16:12 AM">` were all collected and recorded from the actual web pages of the original messages throughout the corpus.

3.5 Summary

In this chapter, the first research question of the study was addressed: *What are the various linguistic codes found on the English website of the mahjoob.com website?* As a matter of course, the source of the data from which the corpus was built was discussed and the reasons for its selection were given. Next, the actual process of downloading and converting the website data into a searchable corpus format was explored. Then, the classification system devised and employed to annotate the corpus data into distinct linguistic codes was described. As a result it was discovered that there are four main code-script pairings on the website that account for virtually all of the linguistic data in the corpus i.e. Arabic-scripted Arabic, Latin-scripted English, Latin-scripted Arabic, and Arabic-scripted English. Each of these was described in turn. Finally, the meta-data and annotation scheme deemed necessary to make further sense of the corpus data were discussed. This also included the method by which the annotation system was tested. In the next chapter, the methods used to exploit the fully annotated mahjoob.com corpus in order to explore the remaining research questions of this study will be presented.
4.1 Overview
In the previous chapter, the corpus data used in this study was described along with the classification system which was developed in order to categorize the data into 16 distinct linguistic codes. Building on this foundation, the current chapter reviews the quantitative and qualitative methodological approaches employed to examine these linguistic codes in order to address the second, third, and fourth research questions of this study. The chapter begins with a brief discussion of corpus approaches to analysis of CMC, code-switching, and discourse-based data. Next, the computer software applications WordSmith 5.0 and SPSS 17.0, which were employed extensively in the study, are outlined and discussed. The chapter then describes the application of these methods in order to address the remaining research questions.

4.2 Corpus approaches to analysis
One of the benefits of using corpus methods in order to investigate linguistic phenomena is that it allows for findings which have an empirical and quantifiable basis. Indeed, the original impetus to develop corpus methodologies within linguistics came from a desire to counter a perceived over-reliance on introspection and native-speaker intuition as championed by Chomsky. In fact, corpus studies with their deliberate focus on linguistic performance, i.e. output as opposed to competence, have revealed that naturally occurring language is never as polished as some grammarians would have us to believe. With this in mind, the use of corpus methods in the present research is deemed profitable insofar as they can help capture the range of linguistic heterogeneity possible when multilingual beings interact in an online environment.
4.2.1 CMC and corpora

A number of studies over the past decade have focused on analysing CMC data using corpora. Given the electronic format of CMC texts, CMC data is particularly suited to corpus building (see Baker, 2006). Indeed, several studies have used corpus methods to investigate CMC-based language behaviour. For instance, Herring and Danet (2007) is an entire edited volume dedicated to studies involving CMC use of languages other than English. Within this emerging CMC literature, several researchers demonstrate the profitability of developing corpora out of CMC texts. Some earlier corpus-based code-switching CMC studies have investigated e-mail (Hinrichs, 2005; Warschauer et al., 2002). Some have focused on IRC (Al Share, 2005; Paolillo, 1999) or SMS text-messaging (Palfreyman & Al Khalil, 2003). Others have looked at blogs (Smedley, 2006). Still others have focused on code-switched web forum postings similar to the ones examined in the present research (Androutsopoulos, 2007; McLellan, 2005). In each case, researchers have exploited the fact that CMC texts are generally easily accessible and readily convertible to other file formats in order to build specialized corpora. In this study, as noted in the last chapter, the same ease of access has been relied upon to build a relatively large corpus of web forum messages.

4.2.2 Code-switching and corpora

As with other fields within linguistics which have benefited from the analytical rigour afforded by corpus methods, several previous code-switching studies have also resorted to corpus data in order to examine the contexts in which code-switches occur. Typically, these studies have involved spoken data that has been carefully transcribed and analysed (see Myers-Scotton, 1998). Callahan (2004), however, offers a break with this tradition, focusing as she does on written code-switching involving Spanish and English in literary texts. For the sake of analysis,
Callahan compiled a specialized corpus of fictional works which exhibited code-switching. Interestingly, Callahan’s study is one of the first of its kind because of its focus on written, as opposed to spoken, code-switching within published texts.

The emergence and increasing popularity of the Internet has occasioned a similarly increasing interest in written code-switching in online contexts as researchers come to view these loci as fertile ground for code-switching. A common trend in most of these more recent code-switching studies has been to compile relatively small specialized corpora (see Palfreyman & Al Khalil, 2003; Warschauer et al., 2002).

Researchers such as Hoffmann (2007) have pioneered automated approaches to downloading large numbers of online texts for corpus building. Generally, though, these corpora have looked at a specific linguistic variety or register (see Hoffmann’s (2007) study of the English-language CNN tape-scripts). Nevertheless, a string of recent studies have begun to employ corpora to investigate CMC-based code-switching (see Hinrichs, 2005; McLellan, 2005; Smedley, 2006). Such studies are promising in that they offer linguists the opportunity to examine patterns of code usage over larger amounts of texts than previously possible or feasible, potentially highlighting more general code-switching tendencies than have been observed to date. Another area which has benefitted from an increase in the use of corpus methodologies is the field of discourse analysis. This will be explored in the next section.

**4.2.3 Discourse analysis and corpora**

As with code-switching studies, the field of discourse analysis has been impacted by the recent fervour for corpus methods. In practical terms, this has meant that discourse analysts have made much more use of corpora as sources of data in which to investigate meaning production than
previously (Wetherell et al., 2001). Typically, discourse analysts are able to identify lexical patterns within a body of texts which suggest systematic ways of representing “the world” linguistically (Fairclough, 2003, pp. 134-135). Naturally, the larger the body of texts, the more substantial the claims that can be made based on a thorough analysis of its discursive contents. In this regard, given its size, one of the benefits of the mahjoob.com corpus is that it provides the opportunity to examine specific cases of identity-creation by certain text-producers across a large number and variety of texts (see Chapter 7).

Having considered how corpus methodologies have come more recently to be valued and employed by researchers in the fields of CMC, code-switching, and discourse analysis alike, it will be useful to describe the actual corpus analysis software packages that were utilized in this study in order to address the research questions. Consequently, a brief description of the key functions of WordSmith 5.0 will be given in the next section, 4.3, followed by a cursory description of SPSS 17.0 in section 4.4. Then, in the remainder of the chapter, the methodologies employed to respond to the second, third and fourth research questions will be examined in detail.

**4.3 WordSmith Tools 5.0**

WordSmith 5.0 is a suite of software applications which help linguists to analyse the contents of corpora (Scott, 2010). For the present research, a specific advantage of WordSmith 5.0 is that it can process texts in Unicode format (see Section 1.5) which means that multi-scriptal data can be searched and displayed quite easily. The main applications used in this study were the concordance and wordlist features. These applications will be presented briefly in the next two sections before turning our attention to how they were employed methodologically in this study.
4.3.1 Concordances

A concordance is a list of all occurrences of a search word (or words) displayed within their immediate textual contexts within a corpus (see Figure 4.1 below):

Figure 4.1: A concordance of FRIEND in the E7ke wa Fadfed forum showing Lines 1-19

As Figure 4.1 reveals, the value of a concordance lies in the opportunity it affords to examine several occurrences of an interesting word simultaneously and in context. In an annotated corpus such as the one used in this study, beyond the actual word items which occur in the forum messages, annotational tags themselves can also be searched using the WordSmith 5.0 concordance application. In this study, concordances were conducted for a number of purposes. First, they were used to count actual occurrences of specific words and tags of interest (see Section 3.4). In the case of tags, this was particularly helpful in tabulating the number of messages which were written in a certain code or script. Secondly, concordances were used to help identify and investigate specific topics (see Section 4.6; also Chapter 6). It should be noted
that, for the sake of manageability, these concordances, which were several thousand lines long, were reduced to a random sample of 100 lines for further in-depth analysis of lexical patterns (see Chapter 6). Thirdly, concordances were used to locate specific authors and their texts across the corpus in order to facilitate a more comprehensive analysis of identity-related code use among these authors (see Chapter 7).

It is worth noting that the concordance program also enables users to examine lexical clusters of two or more items as well as collocations. But since these features were not employed they will not be discussed further here. Now that the basic functions of the WordSmith 5.0 concordance software pertaining to the study have been described it will be beneficial to consider the remaining commonly used application here i.e. frequency wordlists and keyword lists.

4.3.2 Frequency wordlists
Unlike a concordance, a frequency wordlist (or wordlist) is a list of all lexical items in a given text or group of texts ranked in terms of frequency or alphabetic order (see Figure 4.2 below). In the present study, wordlists were employed for several purposes. For one, they were used to reveal the distinct lexical content of each of the three main linguistic codes i.e. Code 1 (Arabic-scripted Arabic), Code 2 (BNC English), and Code 3 (Arithmographemic Latin-scripted Arabic with English) (see Section 4.5). They were also used to explore the lexical content of specific forums, threads, and messages (see Section 4.6).
A wordlist is useful in giving a general picture of the most and least common lexical items within a text or group of texts. Consequently, wordlists can suggest topical focus and stylistic tendencies within a group of texts. For instance, in Figure 4.2 above, it becomes clear that Code 3 is in fact a mixed code due to the presence of English lexical items such as THE, I, AND, etc. alongside Latin-scripted Arabic lexical items like EL (the), MA (not), W (and), ANA (I), etc. Stylistically, Code 3 also exhibits a number of highly frequent smileys (e.g. GREENTOOHYSMILE, OFFERSFLOWER, CHINESECARICATURE in Lines 20, 25, and 26 respectively), which are markers of humour and informality in CMC (Crystal, 2001), suggesting that Code 3 may be used for informal discussions and humorous exchanges (see Chapter 6).
4.3.3 Keywords

Another useful application of WordSmith 5.0 is the keyword function. By comparing two distinct wordlists, it is possible to obtain a list of lexical items which occur more frequently in one wordlist than in the other list which is known as the ‘reference corpus wordlist’. These comparatively more frequent words in the first wordlist are called ‘keywords’. If a set of items is conspicuously rare in the first wordlist when compared to the reference corpus wordlist, these infrequent items are labelled ‘negative keywords’ (see Figure 4.3):

Figure 4.3: Keywords of Code 2 using BNC as a reference corpus (Lines 1-27)

![Figure 4.3: Keywords of Code 2 using BNC as a reference corpus (Lines 1-27)](image)

The keyword list example in Figure 4.3 is derived from a comparison of the Code 2 (BNC English) wordlist to the actual wordlist of the entire British National Corpus (BNC). This keyword list reveals that the black font items I, YOU, IS, IT, etc. are more frequent in the Code 2
wordlist than in the wordlist derived from the BNC reference corpus. For instance, in Line 2, I is shown to occur 2.78 times per 100 words in Code 2 messages whereas I occurs only 0.74 times per 100 words in the BNC. Clearly, Code 2 texts are characterized by more use of I than in the BNC corpus. In contrast, the red font items reveal words which are relatively infrequent in the Code 2 messages when compared to the BNC wordlist. For instance, Line 1 shows that THE occurs only 3.54 times in every 100 words in the Code 2 messages while THE occurs 6.09 times per 100 words in the BNC. In other words, THE is much more frequent in the BNC corpus and is actually a negative keyword in Code 2 messages. As this example shows, keyword analyses are valuable for highlighting the salient lexical differences between sets of texts (see Section 5.3.2).

Now that some of the applications of WordSmith 5.0 pertaining to this research have been discussed, it is pertinent to discuss the other main software package employed in this study, SPSS 17.0. This will be the focus of the next section.

4.4 Statistical Package for the Social Sciences (SPSS) Version 17.0

SPSS is a suite of software applications designed to process large data sets of individual cases. Typically, SPSS is employed for statistical analyses of ethnographical data such as survey responses and population characteristics. In each instance, the data is organized into a set of unique cases whose features can be explored as discrete variables. To illustrate, a set (i.e. population) of completed surveys (cases) can be entered into an SPSS file where each survey question item becomes its own variable. Each variable can have a range of acceptable values. For instance, two different survey items such as “How old are you?” and “Where do you live?” would be coded as two distinct variables in an SPSS data set such as “AGE” and “RESIDENCE” respectively. Case 1 (i.e. the first completed survey) might feature “28” as a response to the first
survey item and “London” as a response to the second. By investigating the AGE and RESIDENCE variables for this data set using SPSS, statistical data such as average age of survey respondents and most common place of residence can then be determined across all cases i.e. completed surveys.

While SPSS is ideally suited to the processing and analysis of such survey type data, to date, no corpus studies appear to make use of an entire SPSS-formatted version of a corpus. Yet, given the structure of the annotated mahjoob.com corpus with its inclusion of several linguistic and contextual details for each message in the form of distinctly tagged bits of information (see Section 3.4), it was deemed worthwhile to consider whether such meta-data could in fact be captured in an SPSS format in order to facilitate the calculation of frequencies and distributions of this data across the corpus. If this conversion could in fact be effected, it was reasoned that the process of addressing the second research question of this study would become both relatively straightforward and highly reliable. With this goal in mind, a series of iterative steps were taken which ultimately resulted in an SPSS-formatted corpus parallel to the original .txt version of the corpus (see Appendix E file “whole_data_set”). These steps are outlined in section 4.5 below.

Now that the key functions and uses of both WordSmith 5.0 and SPSS 17.0 in this study have been described it will now be helpful to consider how these software packages were put to use in order to address the remaining research questions of this study. In this regard, sections 4.5, 4.6 and 4.7 deal with the methodologies employed for the second, third and fourth research questions respectively.
4.5 Investigating code distribution patterns across the corpus

Having addressed the first research question of this study in the previous chapter, namely, the linguistic make-up of the corpus in terms of different linguistic codes, it then became possible to quantify the actual occurrence and distribution of these codes across the corpus. As a result, the second research question could now be answered:

2) How are script-code pairings distributed across certain types of postings in terms of

   a) text type (title, first message, following messages, quoted content) 57
   b) overarching forum topic
   c) thread length
   d) author posting frequency

Below, and in the next few sections, the specific methodologies used to address each of these sub-parts of the second research question will be discussed and illustrated.

In order to investigate and quantify the distributions of codes across the above-mentioned variables, recourse was had to the annotational system described in Chapter 3. To make efficient use of the annotational system, several distinct steps were taken. For instance, in order to examine overall code distribution across text type, it was initially decided to run concordance searches in WordSmith 5.0 of all 14 code tags across all 460,220 messages. However, while WordSmith 5.0 enables its users to compute total numbers of a search word using a concordance, it does not

57 There were four distinct text types investigated in this study: thread titles, seed messages (i.e. the first message posted in a thread), following messages (any messages posted after the seed message of a thread), and quoted message content. Author signature lines constitute another type of text type but due to their use of imported graphic content they were excluded from the analysis.
automatically allow for the comparison of the quantities of search words from different concordances. Therefore, in order to facilitate the tabulation of percentages of this data it was decided that an SPSS-friendly version of the corpus should be created. This conversion process will be discussed in the next section.

4.5.1 Converting the meta-data into SPSS format

In Chapter 3, the mahjoob.com corpus was described as a set of discussion thread files composed, in turn, of varying quantities of messages. It was also mentioned that for the sake of facilitating the investigation of specific questions such as language and script distribution and use across the corpus it was deemed necessary to develop an annotational system. This system could provide meta-data about the corpus such as author ID, linguistic code of message, time of posting, and so on, and this data was captured using various tags. In order to be able to fully exploit such meta-data, I decided that an SPSS version of the corpus might provide a more convenient basis for quantifying the corpus meta-data i.e. the different types of structural, linguistic, and contextual tags created for the corpus. This necessitated another conversion of the corpus data, this time from its WordSmith 5.0-friendly text file format into SPSS format. Again, I solicited Sebastian Hoffmann’s assistance to effect this second conversion. This process took several steps to ensure both accuracy and completeness as described below.

First, all 21,626 of the txt.-formatted corpus files were combined into a single text file. Since SPSS requires that data be organized into one or more sets of basic units of analysis (i.e. single unique cases), it was decided that the message and not the thread should be the basic unit of analysis of the corpus in SPSS format. There were several advantages to this change in hierarchical structure. For one, the discrete linguistic data of each message could now be
compared across all other messages. Another advantage was that the number of messages composed in a certain code could be determined very quickly and contrasted to other messages within the same table of percentages or frequencies. Nevertheless, thread-level analysis could still be carried out quite efficiently: messages could be resorted ordinally into their original threads since each message was annotated to display its original thread number.

Once Hoffmann had created a single text file of the entire corpus out of the original 21,626 corpus text files, he then employed his own purpose-made Perl script in order to assign each message its own unique number (from 1-460,220). This unique number was a feature which did not exist in either the original webpage version of the message or in the message’s WordSmith 5.0-friendly format. Now, each unique message represented its own case (see Figure 4.4):

<table>
<thead>
<tr>
<th>msg</th>
<th>thread_id</th>
<th>url_id</th>
<th>forum_id</th>
<th>thread_title_id</th>
<th>title_script_id</th>
<th>title_AE_id</th>
<th>title_code_id</th>
<th>seed_message_code</th>
<th>msg_order</th>
<th>date</th>
<th>author</th>
<th>location</th>
<th>script_id</th>
<th>script_code</th>
<th>scriptQuote_id</th>
<th>AG_Quote_id</th>
<th>codeQuote_id</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>231417</td>
<td>http...</td>
<td>24</td>
<td>Who Kne...</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>114</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>335</td>
<td>333</td>
<td>333</td>
<td>333</td>
</tr>
<tr>
<td>2</td>
<td>231417</td>
<td>http...</td>
<td>24</td>
<td>Who Kne...</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>214</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>333</td>
<td>333</td>
<td>333</td>
</tr>
<tr>
<td>3</td>
<td>231417</td>
<td>http...</td>
<td>24</td>
<td>Who Kne...</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>314</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>339</td>
<td>333</td>
<td>333</td>
<td>333</td>
</tr>
<tr>
<td>4</td>
<td>215269</td>
<td>http...</td>
<td>24</td>
<td>Great time</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>12</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>14</td>
<td>339</td>
<td>333</td>
<td>333</td>
<td>333</td>
</tr>
<tr>
<td>5</td>
<td>215269</td>
<td>http...</td>
<td>24</td>
<td>Great time</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>22</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>333</td>
<td>333</td>
<td>333</td>
<td>333</td>
</tr>
<tr>
<td>6</td>
<td>215269</td>
<td>http...</td>
<td>24</td>
<td>Great time</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>14</td>
<td>4</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>14</td>
<td>333</td>
<td>2</td>
<td>2</td>
<td>14</td>
</tr>
<tr>
<td>7</td>
<td>215269</td>
<td>http...</td>
<td>24</td>
<td>Great time</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>22</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>333</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>8</td>
<td>215269</td>
<td>http...</td>
<td>24</td>
<td>Great time</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>14</td>
<td>4</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>14</td>
<td>333</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>9</td>
<td>215269</td>
<td>http...</td>
<td>24</td>
<td>Great time</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>22</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>333</td>
<td>333</td>
<td>333</td>
<td>333</td>
</tr>
<tr>
<td>10</td>
<td>215269</td>
<td>http...</td>
<td>24</td>
<td>Great time</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>22</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>333</td>
<td>333</td>
<td>333</td>
<td>333</td>
</tr>
<tr>
<td>11</td>
<td>215269</td>
<td>http...</td>
<td>24</td>
<td>Great time</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>22</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>333</td>
<td>333</td>
<td>333</td>
<td>333</td>
</tr>
<tr>
<td>12</td>
<td>212389</td>
<td>http...</td>
<td>24</td>
<td>public heal...</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>14</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>333</td>
<td>333</td>
<td>333</td>
<td>333</td>
</tr>
<tr>
<td>13</td>
<td>212389</td>
<td>http...</td>
<td>24</td>
<td>public heal...</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>21</td>
<td>9</td>
<td>9</td>
<td>2</td>
<td>2</td>
<td>12</td>
<td>333</td>
<td>333</td>
<td>333</td>
<td>333</td>
</tr>
<tr>
<td>14</td>
<td>212389</td>
<td>http...</td>
<td>24</td>
<td>public heal...</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>9</td>
<td>9</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>333</td>
<td>333</td>
<td>333</td>
<td>333</td>
</tr>
<tr>
<td>15</td>
<td>212389</td>
<td>http...</td>
<td>24</td>
<td>public heal...</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>9</td>
<td>9</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>333</td>
<td>333</td>
<td>333</td>
<td>333</td>
</tr>
</tbody>
</table>

In Figure 4.4, each message in the forum is displayed with its own unique number. The leftmost column gives the case number (each message being its own case). The next column entitled “msg_n” gives the unique number assigned to each message. The third column from the left

---

58 This Perl script is copyright property of Sebastian Hoffmann and is unavailable for publication or distribution.
entitled “thread_n” displays the thread number in which the message was found. The significance of the other columns will be discussed shortly.

Although a crucial step in the conversion process, assigning unique numbers to each message so that messages could be treated as distinct cases of data in SPSS was not sufficient to allow for the processing of the data in each message. The next required step was to be able to define each of these ‘message’ cases by a common set of variables (see the labelled columns in Figure 4.4. above). In so doing, it would be possible to quantify and contrast messages based on their distinguishing properties. In order to achieve this, it was vital to develop a set of SPSS variables based on the various structural, linguistic, and contextual tags which were applied to the WordSmith 5.0-friendly version of the corpus. Conveniently, these meta-data tags from the original corpus provided a set of ready-made labels to be reused in SPSS. Accordingly, Hoffmann reformatted the single corpus file so that each message appeared on its own line in the file and that each tag within a message was separated by a tab space. In this way, the SPSS program would read each line of data as a distinct case and each tab-spaced entry within a line as a distinct data point belonging to that case.

Due to difficulties in importing such a large single text file into SPSS, the file had to be broken into ten smaller text files of 50,000 cases each to facilitate the direct importation of these message cases into SPSS. Another complication resulted from the fact that string (i.e. 59 File 1 contained message cases 1-50,000, File 2 contained message cases 50,001-100,000, File 3 contained message cases 100,001-150,000, and so on. Thus, each of the ten text files contained approximately 50,000 messages. However, File 10, as the last file, was relatively small as it contained the last few message cases of the corpus i.e. cases 450,001-460,220.
alphanumeric) data require more space within SPSS files than simple numeric data do. Consequently, with the goal of reducing file size and thereby reducing data upload and processing times in SPSS, a decision was taken to delete this material from each message case\textsuperscript{60}. This step was considered acceptable since message content was no longer deemed necessary in the SPSS formatted corpus. In addition to these deletions, Hoffmann converted all the remaining string data to numeric data. In this connection, string data such as code names, forum names, author IDs, and location IDs were all converted to numerals. For instance, an original string such as the linguistic tag “\texttt{<code_3>” was simplified to the numeral “3” (see Appendix E file “Numeric Variable Scheme”). Exceptions to this were the URL details, the thread names, and the date and time of posting data. Once these modifications were carried out, all data was successfully imported into an SPSS file which was then checked rigorously for accuracy and consistency\textsuperscript{61}.

The resultant SPSS file could now be enhanced to highlight the desired variables for further analysis. In this regard, a total of 21 variables were borrowed from the WordSmith 5.0 version of the corpus, each corresponding to a separate discrete point of data which characterized the messages in the corpus. As mentioned above, for the most part, these variables were based on the same tags which were used to annotate the original .txt format of the corpus (see 3.4 above). Nevertheless, a brief definition of each of these variables in the SPSS data set will help to clarify their functions (see Figure 4.5 below).

\textsuperscript{60} In any event, such message content was still easily accessed via the WordSmith 5.0-friendly version of the corpus.

\textsuperscript{61} An interval sample consisting of every 10,000th converted message case (i.e. a total of 45 messages) was examined and checked against the original messages in the WordSmith 5.0-friendly corpus.
The first of group of variables provided identifying information about the message: Variable 1 (msg_n), Variable 2 (thread_n), Variable 3 (url_id), Variable 4 (forum_n), and Variable 5 (thread_title_name) referred to the number of the message (between 1 and 460,220), the number of the thread of the message (between 200,000 and 236,000), the url of the message, the number of the forum of the message (between 3 and 78), and the name/title of the thread of the message respectively. In contrast, Variables 6 through 9 provided specific linguistic data about the thread title of the discussion thread in which the message occurred: Variable 6 (title_script id), Variable 7 (title_AG id), and Variable 8 (title_code id) each referred to the script content, arithmographemic content, and linguistic code content of the thread title respectively while Variable 9 (seedMessage_code) contained information regarding the linguistic code of the first message in the thread in which the message was found. Variables 10 to 13 furnished further contextual detail about the message: Variable 10 (message n), Variable 11 (date id), Variable 12 (author id), and Variable 13 (location id) specified the order of the message in its thread, the date

\(^{62}\) Variable 1, Variable 2, and Variable 3, etc. are shown as V1, V2, V3, etc. respectively in Figure 4.5.
and time of posting of the message, the author ID of the message poster, and the location ID of the message poster respectively.

The remaining variables, Variables 14 to 21, were all linguistic in nature. Variables 14 to 17 provided detail about the linguistic features of the message itself i.e. its script, arithmographemic content, linguistic code, and Arabic content respectively, while Variables 18 to 21 provided the same detail about any quoted material within the message if the message did in fact contain quoted material from a previous message.

With this conversion complete, it was now possible to carry out a series of simple cross-tabulations in SPSS in order to determine various distribution patterns of the linguistic codes across the corpus.

4.5.2 Determining code distribution across text type

The second main research question of the study focuses on discovering code distribution patterns in the corpus across four different variables: text type, overarching forum topics, thread length, and poster type (i.e. prolific vs. non-prolific poster). In light of this, the first part of the question sought to establish code distribution across the four main text types found to exist in each message: 1) thread title texts, 2) seed message texts, 3) following message texts, and 4) quoted message texts. In order to determine code distribution across these types, initially, WordSmith 5.0 concordance searches were carried out using the linguistic tags specific to each of these text types. For instance, the tag <title_code id="1"> was searched across the entire corpus to yield the total number of thread titles composed in Code 1 (Arabic-scripted Arabic) in the corpus. The same concordance search procedure was carried out for the thread title tags of each of the remaining codes (e.g. <title_code id="2"> (BNC English), <title_code id="3">...
(Arithmographemic Latin-scripted Arabic with English), etc.). Results from these searches were inputted into an Excel spreadsheet. In this way, it was possible to obtain a complete breakdown of the number thread titles composed in each of the distinct linguistic codes in the corpus. The same method was used to determine the number of seed messages composed in each of the 16 codes (i.e. Codes 1-14, the 14 linguistic codes and the two non-linguistic codes, Codes 15 and 16). In this case the search tags used were <seedMessage_0>, <seedMessage_1>, <seedMessage_3>, and so on. Again, these searches yielded total numbers of seed messages composed in each code. This method was again repeated for following messages using the tag series <code_0>, <code_1>, <code_2>, etc. However, in order to avoid the double counting of seed messages, it was necessary to exclude all seed messages in the corpus. This was achieved by selecting the “sections to cut out” function in WordSmith 5.0, indicating that all text including tags which were found between the tags <message n=“1”> to </message n=“1”>. Thus, it was possible to systematically exclude all seed messages across the corpus from the concordance. The resulting code tag figures for each code represented the following messages. Next, the tag series <codeQuote_0> through <codeQuote_14> was searched to determine precise figures for the codes used in each quoted message. Once figures from all four text types had been obtained, these were inputted into a separate Excel file and their relative percentages calculated based on total numbers for each category of text type. In this way, it was possible to determine which codes were more or less prevalent across thread titles, seed messages, following messages, and quoted messages. As a final step, using the ‘Descriptives’ function under the ‘Descriptive Statistics’ tab within the ‘analyse’ menu in SPSS (see Figure 4.6 below), frequencies and percentages of all codes as they occurred across each of the four texts types were tabulated within their own bar charts (see Figure 5.2).
The methodology used to examine code distribution across text-type has just been discussed. The information provided by this procedure helped to establish the broad distribution patterns of the codes across the corpus, highlighting which codes were more prevalent and which were more obscure. However, in order to investigate possible reasons for code use in the corpus, it was deemed valuable to explore the potential relationship between code, on the one hand, and topic, thread length and poster, on the other, recurrent themes in the literature on code-switching.\footnote{Blom and Gumperz in their seminal work on code-switching in Norway coined the term ‘metaphorical’ code-switching to refer to topic-related code choice as opposed to ‘situational’ code-switching which they understood to mean context or situation-related code choice (Blom & Gumperz, 1972). Subsequently, several theorists have explored the relationship between topic and code choice further (see Bentahila, 1983; Myers-Scotton, 1993b; Warschauer et al., 2002). Meanwhile, in CMC code-switching studies, Androutsopoulos (2007) has looked at forum topic and code choice while Wodak and Wright (2007) have examined thread length and code use.}
methodology utilized to investigate such relationships in this study is outlined in the section 4.5.4 below.

4.5.3 Conflating the 14 linguistic codes in the corpus

In order to test the statistical significance of the 14 codes' frequencies, a Chi-squared test was carried out with the value p-0.05 using the SPSS 'cross-tab' function. Initially, when this chi-squared test was run for all 14 codes, there were too few occurrences of the biscriptal codes i.e. Codes 4-9 and 11-13 to establish statistical differences between them (see Chapter 5). Consequently, these codes were conflated into a single 'Code 4' while all remaining monoscriptal codes were left as before. This conflation resulted in the following six codes:

- Code 1 (Arabic-scripted Arabic - the original Code 1, no conflation)
- Code 2 (BNC English - the original Code 3, no conflation)
- Code 3 (Arithmographemic Latin-scripted Arabic with English - the original Code 3, no conflation)
- Code 4 (Biscriptal content-only - Codes 4-9 and 11-13, conflated)
- Code 10 (Non-Arithmographemic Latin-scripted Arabic with English - the original Code 10, no conflation)
- Code 14 (Non-BNC English - the original Code 14, no conflation)

Now that the conflation of the 14 codes has been described, the method for examining code distribution across variables such as forum topic, thread length, and poster can be discussed.

4.5.4 Determining code distribution across overarching forum topic

Recalling the research question referred to in Section 4.5.2 above, once overall patterns of code occurrence across the four text types in the corpus had been identified, it was deemed necessary to investigate the occurrence of codes across the variable of forum topic. In order to address this question, following message code was selected as the most reliable indicator of this potential relationship. This was because code use was found to differ considerably between seed messages
and following messages, and following messages made up the vast majority of all messages in the corpus (see Section 5.2).

A second step involved recoding all 41 forums into eight overarching topics adapted from Bentahila (1983) in his study of topic-related code choice involving Arabic and French. The adapted topics in the present study were 1) Humour and jokes, 2) Poetry, 3) Work and study, 4) Friends and family, 5) Local/Regional culture, nationality, and politics, 6) Hobbies and pastimes, 7) Gender and age-related forums, and 8) General discussion/opinion (see Bentahila, 1983). Initially, forums were coded into these topics based on their forum titles. However, in order to verify that these forums were in fact connected to the topics into which they had been placed, threads from all 41 forums were randomly sampled using a WordSmith 5.0 concordance of each forum. The sampled threads were then read in their entirety and classified as pertaining to one of the eight topics. This process did in fact confirm the original topical coding of most forums.64

Next, the SPSS corpus data file containing all message cases was copied as a second SPSS data file and a new variable was added to the existing 21 variables in order to indicate the topic of the forum in which the message appeared. In this manner, all message cases received a numeric value between 1 and 8 to reflect the topic of their respective forum. For instance, all Religion forum messages received the numeric value 5 indicating that they belonged to the topic of Local/Regional culture.

Once this labelling had been carried out, using the Cross-Tab function in SPSS, a table was produced that displayed the percentage of following messages composed in Codes 1, 2, 3, 4, 10, 64 A notable exception was the Soul Retreat forum which was initially coded as a General Discussion/opinion forum. Upon random sampling of its threads it was found to belong in fact to the Poetry forum.
and 14 across each of the eight overarching topics. The first, second and third most used code for each overarching topic were highlighted in yellow, green, and blue respectively. As a result, a picture of topic-related code choice could emerge from the data.

While topic-related code choice patterns are certainly valuable in helping to delineate the contrastive functions of the linguistic codes and their metaphorical uses (see Blom & Gumperz, 1972), in order to gain a better understanding of how such topic-related code use patterns might relate to identity-formation, it was deemed desirable to explore other possible patterns of code use in the corpus such as thread length-related code use and author-related code use. The methodologies employed to explore these other patterns in the corpus will be discussed in the next two sections.

4.5.5 Determining code distribution across thread length

The third part of the second research question was to examine code distribution across longer and shorter threads. This question was motivated by research done by Wodak and Wright (2007), who had discovered that multilingual discussion threads of different lengths tended to reflect different code usage patterns i.e. the longer the discussion thread the more mixed it was linguistically. In a bid to determine whether any salient differences in code use might exist between longer and shorter discussion threads in the mahjoob.com data, a decision was taken to restructure the data set in such a way that longer and shorter threads could be compared and contrasted. First, using SPSS descriptive statistics to calculate averages, it was determined that the average number of messages in a thread was 25 messages. However, it was also discovered that several hundred threads were in fact very long, containing 100 or more messages. Since such long threads provided ample message data to observe code use patterns, it was decided to use the
100th message mark as a cut off point. This meant dividing the SPSS corpus messages into two groups: those containing 100 or more messages and those containing less than 100 messages.

To effect this restructuring of the data, using the SPSS data set, all threads with 100 or more messages were grouped together in order to be contrasted with threads that contained less than 100 messages. This was achieved by first obtaining descriptive statistics for each thread. A total of 476 threads were found to contain 100 or more messages each (see Appendix E file “thread_n_descending_count”). Next, using the SPSS RECODE variable function, the thread_n variable was recoded into a new SPSS variable entitled, “Largest_to_smallest_threads_recoded” (see Appendix E file “Recoded_threads_by_length”). In this process, each message from these longer threads was reassigned the value 1 under the new variable. All of the remaining messages in the dataset, which were all from threads containing less than 100 messages, were reassigned the value 2. At the end of this process, this restructured data set was saved as “restructured_data_set_by_thread_2” (see file in Appendix E). Next, using this data set, a cross-tab analysis was executed in order to contrast code use frequencies and percentages between “Shorter threads” (i.e. those containing 1-99 messages) and “Longer threads” (i.e. those containing 100 or more messages). In this way, a clearer picture of code use across thread length was obtained.

Such thread length data provided valuable information for identifying and examining atypical threads as loci for identity-creation in Chapter 7. In this connection, detail concerning author-related code choice was also desirable. The methodology used to explore author-related code choice is the focus of the next section.
4.5.6 Determining code distribution across author posting frequency

At the outset of the research, as the various forums and threads were browsed online, it appeared that certain forum contributors were quite prolific, posting messages to several different threads. Thus, it was decided to investigate whether such posters were consistent in their code use and whether such consistent code use might reveal identity-construction processes. This question corresponded to the fourth part of the research question outlined in Section 4.5.2: How are script-code pairings distributed in terms of author posting frequency? To address this, as was done for thread length, the data set was restructured again in terms of most prolific authors. Initially, information on the most prolific authors was extracted from the corpus by performing concordance searches in WordSmith 5.0 with the search tag <author id=*>. These concordances revealed that there were indeed numerous prolific authors who had posted at least 1000 messages in the corpus. To explore further the possible impact of such prolific authors on code distribution in the corpus, the top 10 authors were grouped together in order to compare their code use patterns to the remaining 1,251 authors. The regrouping of these authors entailed defining a new SPSS variable (“top10_authors”) as was done for longer vs. shorter threads. Recoding all messages posted by a top 10 author with the value “1” and assigning the remaining messages the value “2”. It was now possible to examine author posting behaviour across both code choice and topic (see Section 5.5). This yielded insights into how top 10 authors differed from other authors in terms of the specific topic areas they tended to post to and the codes which they used for such postings.

65 The impact of such prolific posters could not be overlooked since it was determined that the top 10 authors alone accounted for roughly 20% of all forum messages in the corpus.
The examination of code distribution patterns across text type, topic, thread length, and author offered a range of insights into possible uses and values attached to each of the linguistic codes in the corpus. However, at best, such insights are valuable at a bird’s eye level since they focus merely on overall distribution trends and frequencies. In order to ensure that results obtained were not due to chance, the SPSS cross-tab function was used to measure observed code frequencies against expected frequencies. The p-value was set to 0.05, signifying that any differences in code distribution across the chosen variables of text type, overarching topic, thread length, and poster frequency had a 5% or less likelihood of having occurred by chance. And indeed, all differences were found to be significant (see Chapter 5).

In order to explore further whether any substantive differences existed between the codes in the corpus in terms of topic, stylistics, and function, it was deemed necessary to investigate specific instances of code use in a systematic way using corpus analysis methods. Section 4.6 below describes how this was achieved.

**4.6 Topical content and stylistic features of the main linguistic codes**

One key goal of the research was to establish the differences among the script-code pairings i.e. linguistic codes in corpus in terms of topics and stylistic features. In this regard, examining the distribution of linguistic codes across several variables including topic, as described in the preceding section, provided several valuable details about these codes such as which of these could be considered the main linguistic codes in the corpus; i.e. those which were found most frequently across the entire corpus as well as among specific forums, threads, and authors.\(^6\)

---

\(^6\) In Chapter 5, it is revealed that Code 1 (Arabic-scripted Arabic), Code 2 (BNC English), and Code 3 (Arithmographemic Latin-scripted Arabic with English) are in fact the main (i.e. most prevalent) codes in the
With specific reference to topic, patterns of code use for different topics could also be explored to an extent by focusing on code occurrence and frequency across the various topical forums. By clustering these forums into even broader, overarching topics, a picture of topic-conditioned code use could also emerge from the data. However, this level of macro-analysis was still quite abstract and superficial. In response to this limitation, the third research question of this study was conceived:

3) What do frequent lexical patterns in the main script-code pairings in the corpus reveal about the topical content and stylistic features of those pairings?

In order to address this question, several steps were taken. Specifically, these involved the use of two specific WordSmith 5.0 applications: frequency wordlists and concordances. These procedures will be described in the next few sections.

4.6.1 Compiling frequency wordlists for the main codes in the corpus

In order to examine topic more closely, Baker (2006) establishes the usefulness of frequency wordlists in revealing potential insights into the latent discourses within texts. By compiling wordlists, he shows that the most frequent lexical and grammatical items in a corpus can be discovered, which can then help generate hypotheses about the discursive content of such texts. For the purposes of this research, this method was adopted to uncover what kinds of topics might be foregrounded in a text or set of texts. Clearly, however, as Baker (2006, p. 68) alludes, further corpus analytical procedures are required in order to test such early hypotheses.
To create these wordlists, numerous procedures had to be followed. First, since the corpus is composed of several thousand text files based on discussion threads which can and do contain messages composed in several different linguistic codes, it was necessary to compile discrete frequency wordlists of each of the three main codes in the corpus i.e. Code 1 (Arabic-scripted Arabic), Code 2 (BNC English), and Code 3 (Arithmographemic Latin-scripted Arabic with English) so that their most salient lexical content could be investigated and contrasted (see Chapter 5). Using the wordlist function in WordSmith 5.0 and the linguistic tags built into the corpus, it was possible to scan all the message data in the corpus in order to produce a distinct wordlist for each of these three codes (see Appendix F). In order to achieve this, message texts in the corpus had to be chosen from which to derive the wordlists. Since message texts written in any code might occur in any thread in the corpus, it was necessary to include all 21,626 discussion thread text files in the wordlist compilation process. However, given the possibility that a certain code might not feature in a specific thread, it was decided to resort to the “tags as selectors” function. This ensured that only files which contained message texts composed in a specific code were scanned to compile each code’s wordlist (see Figure 4.7 below).
Another parameter had to be set. To avoid including parts of files such as tags, thread titles, and quoted message content which would skew the wordlist, it was necessary to specify which parts of the corpus files were to be included in the wordlist. Accordingly, the “Select Only sections of text files” function was employed (see Figure 4.8 below):
As shown in Figure 4.8 above, by using the tags `<quote>` and `</quote>`, `<quote id=*>` and `</quote id=*>` as well as `<new>` and `</new>` with the “sections to cut out” operation, it was possible to ignore all file content extraneous to the actual new message content during wordlist compilation. Next, in order to specify that the wordlist compilation process captured message text lexis from a specific code and not from others, the “sections to keep” function was employed with a tag pair `<code_1>` to `</content>`. This particular combination of tags resulted in a wordlist containing lexis only from Code 1 messages. By maintaining the same “text to cut out” tags while changing the “sections to keep” tag pair to `<code_2>` to `</content>`, `<code_3>` to `</content>`, and so on, it was possible to compile wordlists respectively for Code 2 messages, Code 3 messages, etc.

### 4.6.2 Examining the Top 10 open class lexical words in the main codes

Once wordlists for each of these codes had been compiled, it was decided to identify the top 10 open-class lexical words in each wordlist as a means of detecting topical content, following the
lead of Baker (2006), who profitably explored the ten most frequent open class lexical words in his discourse analysis of a corpus of holiday brochures. The decision to use Baker’s approach here was motivated by several factors. First, by focusing on only the most frequent lexical items in a given code, it would be possible to generate initial hypotheses about the topical focus of that code. For instance, if the word *Allah* occurred frequently in Code 3 (Arithmographemic Latin-scripted Arabic with English), it would be worthwhile to explore whether Code 3 texts might be used to talk about God or religion.\(^6\) Thus, by sampling the most frequent ten open class lexical words from each code, a broad sense of a variety of possible topical foci and stylistic uses was determined for each code and ideas were generated about whether there might be any overlap in these topics for a given code.

Pragmatically, this relatively small number of items also made it easier to compare the surface topical similarities between the main codes and provide a deeper level of analysis for each of these. Clearly, a greater number than ten lexical words could have been selected, but given the vast number of items in each wordlist, a cut-off point had to be selected especially since a certain amount of repetition was observable among frequent lexical items in each wordlist such as Code 1 (Arabic-scripted Arabic)’s top 10 frequency items *yawm* 

\[
\text{yawm ‘day’ and } \text{al-yawm ‘the day’}
\]

and Code 2 (BNC English)’s THANKS (wordlist item no. 93) and THANK (wordlist item no. 116 respectively). However, an important point regarding representativeness and distribution of the chosen top 10 words from each code should be made. Despite the likelihood of the co-occurrence of these highly frequent words in several of the same threads, adding together the percentage of threads in which each top 10 word occurred appeared to cover a wide range of the

---

\(^6\) This hypothesis was not, in fact, borne out as will be seen in Chapter 6.
corpus. To illustrate, consider that Code 1 (Arabic-scripted Arabic) top 10 words occurred in 25,033 threads i.e. 115% of all threads in the corpus (indicating overlap of terms in the threads), while Code 2 (BNC English) top 10 words occurred in 19,559 texts i.e. 90.5% of all threads and Code 3 (Arithmographemic Latin-scripted Arabic with English) top 10 words were found in 25,179 i.e. 116% of all threads in the corpus (again allowing for overlap of terms in the threads). These figures suggest that even as low a number as ten frequent open class lexical items can cover a relatively large amount of the texts of the corpus, providing an overall sense of the usage patterns of a code across the whole corpus.

The decision to focus on 100 random concordance lines drawn from across the corpus for each of the ten items in Codes 1, 2, and 3 respectively should also be mentioned here. The choice of 100 lines reflected a desire to strike a balance between depth and breadth of analysis. Initially, I considered using only 50 random concordance lines per word, but I felt that these might offer too few lines to delineate salient usage patterns associated with a given word. This led me to double the number of lines to 100. Certainly, more than 100 lines could have been analyzed, but such an exhaustive analysis would surely have exceeded the space limits of the thesis.

In light of the above, the claims made about the the topical and stylistic features for each of these codes cannot be taken as absolute or exhaustive for each code. At best, they are indication of salient themes and styles associated with each code in the context of its most frequent open class lexis. Nevertheless, the in-depth analyses of the top 10 lexical items from each of the three main codes did in fact reveal certain distinctive characteristics of each of these three codes (see Chapter 6), findings that inform the micro-analyses of identity featured in Chapter 7.
In order to select the top 10 open-class lexical items for each code, the UCREL CLAWS7 Tagset\textsuperscript{68} was used as a measure of determining whether a given lexical item was an open-class one. In the case of Code 3 (Arithmographemic Latin-scripted Arabic with English), several homograph cases were encountered in which a word could have been either English or Latin-scripted Arabic. These ambiguous items were hand-checked to determine whether they functioned as open-class items or closed-class items such as prepositions. If a Code 3 item functioned as an open-class noun, adjective, or verb in 50% or more of all cases, it was kept in the Code 3 top 10 list. Otherwise, it was discarded. In this regard, the item MEN which had two possible meanings i.e. the plural form of the English word ‘man’ and the Arabic preposition meaning ‘from/of” was discarded since it was found to occur more than 50% of the time as the closed-class Arabic preposition. Once such items had been discarded, the remaining top 10 lexical item lists for all three codes were compiled into a table for comparison (see Section 6.3).

A next step in the analysis of specific topical focus of the linguistic codes of the corpus was to examine further the top 10 lexical items in detail. Since a full corpus analysis of all ten items in

\textsuperscript{68} UCREL stands for ‘University Centre for Computer Corpus Research on Language’, a corpus linguistics research centre at Lancaster University, UK. The original CLAWS (Constituent Likelihood Automatic Word-tagging System) is a part-of-speech tagging system developed at Lancaster, Oslo, and Bergen Universities in the early 1980s (Garside, 1987). Since then, UCREL has updated the tagset which is currently at version 8. The CLAWS7 Tagset used in this study consists of a list of short abbreviations used to annotate lexical items in stretches of text according to which part of speech they belong to (see http://ucrel.lancs.ac.uk/claws7tags.html). The list also includes several examples of items which belong to the different parts of speech. For the purposes of this study, the CLAWS7 Tagset provided examples of closed-class and open-class lexical items which informed the selection of the top 10 open-class lexical words analysed in Chapter 6.
each code’s list (i.e. a total of 140 items) would have exceeded the limits of space for this study, a decision was taken to concentrate on the top 10 lexical items from only the most prevalent codes in the corpus. Specifically, these were found to be the three mono-scriptal codes: Code 1 (Arabic-scripted Arabic), Code 2 (BNC English), and Code 3 (Arithmographemic Latin-scripted Arabic with English) as established by the code distribution analysis procedures described in Section 4.5 above (see also Chapter 5). Thus, for the top 10 lexical items in each of these three codes, a series of further WordSmith 5.0-based investigations involving concordances were carried out in order to help gain an understanding of their most frequent topical focuses and their stylistic differences. These follow-up procedures will be described in the next section.

4.6.2 100-line concordance analysis

In order to render the analysis feasible, it was decided to investigate a random sampling of 100 concordance lines for each of the top 10 lexical items in Codes 1 (Arabic-scripted Arabic), 2 (BNC English), and 3 (Arithmographemic Latin-scripted Arabic with English) respectively. This was achieved by first compiling full concordances for each top 10 word based on its respective sub-corpus. Once the main concordance was compiled, the WordSmith function “reduce number to” was employed to randomly select 100 lines from each concordance.

69 In fact, Code 1, Code 2, and Code 3 as the three main codes in the corpus could also be considered the most representative of Arabic-scripted Arabic, Latin-scripted English, and Latin-scripted Arabic, the key linguistic ‘ingredients’ in the remaining biscriptal codes (Section 3.3).

70 Notably, as will be seen in Chapter 5, the two other mono-scriptal codes, Code 10 (Non-arithmographemic Arabic with English), and Code 14 (Non-BNC English), were more prevalent than the biscriptal codes in the corpus. Still, the relative occurrence of both Codes 10 and 14 was dwarfed by Codes 1, 2, and 3.
The 100-line concordance tables were then copied and pasted into Microsoft Excel spreadsheets (see Appendices H, I, and J). Each line was then read and annotated to indicate its type of language (vernacular/standard, formal/informal), topic (sports, religion, relationships, etc.), discursive function (rhetorical question, assertion, exclamation, etc.) (see Callahan, 2004, pp.70-75), level of involvement of text composer and/or addressee with the text (involved for first and second person references, non-involved for third person references), and stylistically, whether it contained smileys. In this way, all words were analysed in context.

Such insights into frequent lexical tendencies in Code 1 (Arabic-scripted Arabic), Code 2 (BNC English), and Code 3 (Arithmographemic Latin-scripted Arabic with English) helped to establish how these three codes were used by forum participants to discuss certain topics or to carry out certain stylistic functions. This in turn suggested the general patterns regarding topical circumstances under which forum posters might opt to use one code instead of another.

However, when a specific code is used by a forum poster outside of an expected topical context, the question arises as to what motivates such a change from established practice. The more recent literature on code-switching seems to suggest that perhaps identity-construction motives are at hand in such cases of unexpected code use (see Hinrichs, 2005; Rampton, 2007; Sebba &

---

71Although smileys were systematically excluded from the top 10 lexical item lists for each code, one methodological challenge which had to be remedied was the fact that both Code 2 (BNC English) and Code 3 (Arithmographemic Latin-scripted Arabic with English) featured the frequent item LOVE which was also the name of a smiley in the corpus. As a result, all 100-lines of LOVE in Code 2 and Code 3 respectively had to be hand-checked in order to eliminate lines containing the LOVE smiley. These lines were then replaced with other random lines from the LOVE concordance which were in fact bona fide samples of LOVE as a lexical item as opposed to a smiley (see Chapter 6).
Wooton, 1999). It is this question and a suitable methodology to address it that the remainder of the chapter describes.

4.7 Identity construction through strategic use of linguistic codes

By examining code distribution across the corpus and exploring the topical content of the main codes in the corpus, a broad picture of normative code use and function began to emerge. Positing that forum posters are in fact sophisticated linguistic beings with the capacity to create their own distinct identities by strategically using the linguistic resources available to them (see Smedley, 2006; Su, 2003), it was deemed desirable to explore this hypothesis by investigating specific cases of variable code use among forum posters in depth. This led to the formulation of the final research question of the study:

4) How do forum posters use the available script-code pairings to construct identities?

The methodology used to address this question was heuristic and relied on both quantitative and qualitative approaches grounded in the findings of the earlier research questions. Also, given the size of the data set, it was decided to limit the analysis to a relatively small pool of data in order to provide modest but tangible results. Consequently, no attempt is made to generalize the findings here to other forum posters or to the corpus as a whole.

To begin the analysis, it was decided to focus on longer threads (i.e. containing 25 or more messages) as these would provide an opportunity to investigate instances of identity creation over a number of messages. It was then necessary to select threads for analysis. This involved scanning the entire corpus to find longer threads which were somehow salient in terms of code
use\textsuperscript{72}. As a result, Thread 206940 was detected which, of all lengthy threads in the corpus, was unique in that it was the only thread to contain more messages composed in Code 10 (non-Arithmographemic Latin-scripted Arabic with English) than in Code 3 (Arithmographemic Latin-scripted Arabic with English)\textsuperscript{73}. This thread was also salient in another way: it occurred in the Religion Forum, a forum containing a higher proportion of Code 1 (Arabic-scripted Arabic) messages than other codes. Thus, it was reasoned that these two distinguishing features of Thread 206940, i.e. an unusually high Code 10 content and its location in a forum dominated by Code 1 messages, might provide the opportunity to observe clear instances of strategic code use in support of identity construction.

Subsequently, the original Thread 206940 was accessed from the website and printed. Once printed, the thread was read and divided into its constituent discussions based on whether or not a shift in topic occurred at any point. In this regard, the first 75 messages of the thread were found to focus on a single topic and to contain the majority of Code 10 (non-Arithmographemic Latin-scripted Arabic with English messages), so these were selected for more detailed quantitative and qualitative analyses.

This analysis involved several procedures. First, each message in the discussion was translated into English (see Appendix G). Next, a table was calculated to show the number of authors who

\textsuperscript{72} It had already been established that Code 3 (Arithmographemic Latin-scripted Arabic with English) was the code most often used for such long threads (see Chapter 5).

\textsuperscript{73} A relatively high occurrence Code 10 in Thread 206940 was salient in and of itself due to the fact that of all the mono-scriptal codes in the corpus i.e. Codes 1, 2, 3, 10, and 14, Code 10 is the least common (see Chapter 5).
contributed to the first 75 messages and the codes in which they posted. Within this thread, certain posters were found to be especially prolific notably, Snipe_aac, Kharoof Tayeh, Muslim4, and Argumentive_36.

At this point, discourse analysis was carried out using Fairclough’s notion of ‘linguistically realized styles’ (2003, p. 160). This required that each message be read several times to determine how its author appeared to portray himself both in terms of stylized code use (Su, 2003) and discursive content. Messages were not read in isolation: the analysis also involved reading the texts as components of an interaction in which differential code use might reveal distinct and contrasting styles among interactants (Fairclough, 2003). Once this had been done, it was then possible to identify symmetries between different rhetorical positions among the thread participants and their code use. These symmetries were interpreted as examples of identity construction where authorial stance and stylized code use combined to create distinct and salient styles in the context of the threaded discussion.

To illustrate this process, consider the first four messages of Thread 2069040 (see Figure 4.9 below), the discussion thread which forms the basis of the discursive analysis of identity-related code use in Chapter 7. These four messages were initially categorized in terms of code use and

74 In this particular thread, it appeared that all contributors were males based on their grammar use in Arabic, their choice of author IDs, and their discussion of their attitudes towards and relationships with women. Nevertheless, there was no way to confirm this supposition.

75 For the purposes of illustration and in the interests of space here, all tags irrelevant to the discourse analysis of these messages have been stripped from the messages shown in the example. Specifically, redundant linguistic tags such as <AG id="yes"> and end tags such as </code_3> are not displayed here.
put into a table, showing author, message number, and code. As a result, Messages 1 and 3 were identified as being composed in Code 3 (Arithmographemic Latin-scripted Arabic with English), while Messages 2 and 4 were listed as Code 1 (Arabic-scripted Arabic). Next, the messages were read several times and categorized based on the code used in each and translated into English where necessary.

**Figure 4.9: First Four Messages of Thread 206940**

<thread n="206940">
<forum n="26" name="Religion_Forum">
<thread_title name="masha2a allah masha2a allah">

<content>
<quote>
كشف تقرير أصدره معهد التمويل الدولي أن الولايات المتحدة الأمريكية اجتذبت غالبية إيرادات النفط الخليجية، وما زيد على نصف روؤش الأموال المصدرة من هذه الدول في السنوات الخمس الماضية. وذكر التقرير بحسب ما نشره جريدة "الرأي العام" الكويتية اليوم الجمعة 1-6-2007 أن مشتريات الأصول المقومة بالدولار ربما تكون قد ارتفعت بعد هجمات الحادي عشر من سبتمبر/أيلول 2001. الايرادات وفرت المتعاقدين وهو راحة عالمية لمساهمات عتبة تزيد أكثر من 360 عضوا على مستوى العالم إجمالا إيرادات صادرات دول مجلس التعاون الخليجي وهم البحرين والكويت، عمان وأفغانستان والإمارات بما يزيد على 1.5 تريليون دولار في فترة من 2002 إلى 2006. وأضاف "نحو 542 مليار دولار منها ذهبت إلى الولايات المتحدة، وبالتالي هو نحو 542 مليار دولار يمثل أموالاً فائضة ذهبت أصول روؤس الأموال الحالية وأسهمت في زيادة ما بعدها دول المجتمع من أصول أخرى، 300 مليون دولار وقد ثبت التقرير أن نحو 300 مليون دولار ذهبت إلى الولايات المتحدة بينما ذهب 100 مليون إلى أوروبا و نحو 60 مليون إلى أسا مثيراً إلى أن مشترات المستثمرين الخليجيين لأصول المحومة بالدولار من خلال حرف ذاك بلغت متوسطات مرتفعة. ونادر ما يبدع ذلك ارتفاع بأسعار أصول درهم رخيص، رفع Empresa "السماح". التقرير: "بشكل عام فقد أن حصة الولايات المتحدة من الأصول المقومة بالدولار في حافلة دول مجلس التعاون الخليجي استمرت خلال الأعوام الخمسة وربما تكون قد زادت".</quote>

<new> source wbeejow begolooy enno elhai2a btenteqed elmashayekh. Gayeb...does anyone know the phone number for the hai2a? I’d like to inform them that 300 billions of gulf money is helping the economy of “their enemy” as they claim. <smiley>whistlingface</smiley></new>
</content>

<content>
<new> الله يرحم الملك فيصل</new></content>

<content>
<new> Amen ya rab el 3alameen</new></content>

<content>
<new> الله يرحم الملك فيصل</new></content>
</content>
</thread>

<thread n="2">
<date id="1st June 2007, 10:37:19 PM">
<author id="Guillotine">
<location id=":rant:">
<code_1>
<content>
<new> الله يرحم الملك فيصل</new></content>
</thread>

<thread n="3">
<date id="1st June 2007, 10:43:16 PM">
<author id="3affash">
<location id="Where ever the road may lead">
<code_3>
<content>
<quote id="Guillotine">
الله يرحم الملك فيصل</quote></content>
</thread>

<thread n="4">
<date id="1st June 2007, 11:27:47 PM">
<author id="Muslim4">
<location id="n/a">
<code_1>
<content>
<new> Amen ya rab el 3alameen</new></content>
</thread>
After several readings of these messages (as well as subsequent ones), it was possible to
determine that Messages 1, 2, and 3 were all Anti-establishment in tone. For instance, Message 1
composed in Code 3 (Arithmographemic Latin-scripted Arabic with English) and written by a
poster named Kharoof Tayeh (lit. ‘Sheep who has gone astray’) essentially criticizes the religious
authorities for turning a blind eye to the anti-Islamic behaviour of the Saudi authorities. Message
2 composed in Code 1 (Arabic-scripted Arabic) and written by poster, Guillotine, and Message 3
composed in Code 3 and written by poster, 3affash affirm this view by invoking blessings upon
the deceased patriarch of the Saudi family, King Faisal, implying that if he were alive today he
would be horrified at what his country’s government was doing. In contrast, Message 4
composed in Code 1 by poster, Muslim4 counters the attack on the religious authorities,
defending the important role they play in safeguarding the morals of society. Thus, from this
small example, it was seen that two Anti-establishment posters used Code 3 while one Anti-
establishment poster used Code 1 and the lone Pro-establishment poster used Code 1. These
trends were examined further across the remainder of the debate, yielding a clear picture of
which codes were used by Pro-establishment vs. Anti-establishment authors. Re-reading of the
entire debate elucidated further linguistic differences between the various posters such as stylistic
formality and orthographic patterns. These contrasting patterns provided the evidence for
identity-creation on the part of posters.
In order to corroborate the findings from Thread 206940, it was decided to follow the postings of its main contributors across the rest of the corpus in order to gain a deeper and richer understanding of the practices of this small size sample (see Smedley, 2006). This meant carrying out concordances using the <author id=* tag to locate all messages composed by these contributors across the entire corpus. Once this had been done, another code usage table was compiled to display which author used which codes most often across the entire corpus. Such data facilitated a cross-thread analysis of the authors’ strategic code use in support of their own identity creation work (see Lee, 2007; Smedley, 2006). An example will clarify this process. Using the tag <author id=“Muslim4”> as a search word, all messages of Muslim4 were collected in a single concordance file. Next, these messages were sorted by code. Then, following the same discourse analysis methodology employed for analysing Thread 206940, Muslim4’s messages were read and re-read in the context of the thread in which they had been posted along with the messages of other posters found in close proximity to Muslim4’s messages. In this way, it was possible to gain a clearer understanding of how and why Muslim4 varied his code use across threads which contributed to his creating different, contextually-relevant identities as a result. This procedure was followed not only for Muslim4 but for the three other main contributors to Thread 206940 as well i.e. Kharooftayeh, Snipe_aac, and Guillotine (see Chapter 7).

4.8 Summary

This chapter has outlined the different methodologies employed to address the second, third, and fourth research questions of this study. SPSS software was a key analysis tool in addressing questions pertaining to the distribution of the various codes in the corpus across text type, topic, thread length, and author. WordSmith 5.0 software served to address the question of topical
focus and stylistic functions across the main linguistic codes in the corpus. Finally, a combination of SPSS and WordSmith 5.0 applications were employed to help discover threads and authors for in-depth discourse analysis of identity creation. Addressing each question in order moved the analysis from macro-level perspectives to ever deeper and finer levels of understanding, starting with a view of broad patterns and ending with unique instances of identity construction work. In the next three chapters, the results of each of these methodologies will be discussed.
Chapter 5: The distribution of codes in the corpus

5.1 Overview

In Chapter 3, the data were described and the classification system used to analyse the data was presented. In Chapter 4, the methodologies used in this study involving corpus and discourse analysis were outlined. The focus of the present chapter is to present findings in connection with the second main research question of the study:

2) How are script-code pairings distributed across certain types of postings in terms of
   a) text type (thread title, first message, following messages, quoted content messages)
   b) overarching forum topics
   c) thread length
   d) author posting frequency

Regarding sequence, code distribution patterns across the entire corpus are presented first followed by a forum-by-forum breakdown. Next, the relationship between code distribution and overarching forum topic (see Androutsopoulos, 2007) will be examined more closely to see whether any significant patterns can be determined. Finally, the relationships between code distribution and thread length and author posting frequency will be explored. Recurrent patterns, where found, will be considered again in Chapter 7 as they relate to the strategic use of codes for identity construction.

In order to provide a comprehensive overview of script and code occurrence in the corpus, the distributional patterns of Arabic-scripted messages, Latin-scripted messages, and Mixed Arabic-Latin script messages will now be discussed. As mentioned in Chapter 4, the thread titles, seed messages, following messages, and quoted content messages of the corpus were selected as the
main text types in the corpus for analysis. Consequently, the first part of this chapter will reveal patterns of code use across these four main text types. It is worth mentioning that SPSS-based Chi-squared tests were employed to ensure that the distributional patterns discovered were not due to chance alone using a 5% significance level, i.e. rejecting the null hypothesis of relationship being due to chance when the p-value of the chi-squared test is less than 0.05 (see Appendix E file "Chapter_5_Chi_squared_tests.spv").

An important point should be mentioned here regarding Code 16. As described in Section 3.3.3.6, Code 16 (Graphic content only) denoted texts that were composed of a hyperlink or imported graphic material and did not exhibit any readily parsable textual content\(^7\). As such, while potentially interesting in terms of establishing the hyper-textual nature of much of the website, as with Code 15 (Numeric content-only), no certainty could be reasonably established regarding which language the author was using when composing these texts unless these messages were each examined on a case by case basis. Consequently, apart from a mention of their frequency, Code 16 texts will not be considered further here. The discussion begins now with the overall code distribution patterns of the remaining codes in the main corpus.

\(^7\) This is not to say that all graphic content was text-free. Indeed, several of these graphic-based messages were found to exhibit textual content such as Qur'anic verses, cartoons, maps, etc. However, from a methodological point of view, the texts in these messages could not be automatically converted into text file format and assigned a code value. Consequently, in terms of the influence of prior text on subsequent code choice, their presence is an obscuring factor and was excluded from the analysis on these grounds.
5.2. Overall code distribution of the 14 linguistic codes in the corpus

5.2.1 Overall code trends in the corpus across thread titles, seed messages, following messages and quoted messages

As discussed in Section 3.3, the 14 distinct linguistic code tags were applied to four text types in the corpus: the thread titles, the seed messages, the following messages, and, where present, the quoted messages that were embedded within the following and seed messages. These four tags respectively were: \textit{title\textunderscore code\_1} to \textit{title\textunderscore code\_14}, \textit{seedMessage\_1} to \textit{seedMessage\_14}, \textit{code\_1} to \textit{code\_14}, and \textit{codeQuote\_1} to \textit{codeQuote\_14}. The distribution of the 14 codes across these text types was by no means even. There were in fact several dominant codes with relation to all four text types.

To put all of the trends into perspective, Figure 5.1 (below) was compiled (using Microsoft Excel), reflecting the percentages for all four text type variables: thread title code, seed message code, following message code, and quoted content message code:

\textbf{Figure 5.1: Code distribution across the four text types}\footnote{In this figure, I have omitted data for Codes 15 and 16 in order to put its overall frequency in context. Again, Code 16 is roughly stable at 6.5\% for both seed message and following message.}
Based on this figure, a number of points are observable. First, the mono-scriptal codes: Code 1 (Arabic-scripted Arabic), Code 2 (BNC English), and Code 3 (Latin-scripted Arithmographemic Arabic with English), and to a far lesser extent, Code 10 (Latin-scripted Non-arithmographemic Arabic with English), and Code 14 (Non-BNC English) dominated the entire corpus across all four text types with a combined percentage of 97% of all thread titles, 88.3% of all seed messages, 96.8% of all following messages, and 96.4% of all quoted content messages. Averaged together, 94.6% of all textual content across all four text types is accounted for by the five mono-scriptal codes alone. In terms of biscriptal codes, only Code 6 (Arabic-scripted Arabic with BNC English) has any sort of salient frequency as it accounts for 6.4% of all seed messages.

Regarding occurrence, it is clear that Code 1 (Arabic-scripted Arabic) is the most consistently used code across all four text types. It accounts for between 30-40% of all thread titles, seed messages, following messages, and quoted content messages. Code 2 (BNC English), on the other hand, reveals intriguing patterns. For instance, it is the second most widely encountered code for thread titles at well over 35%, yet it is far less common for seed messages, following messages, and quoted content messages where it ranges between 13-17%. Code 3 (Latin-scripted Arithmographemic Arabic with English) shows an almost converse pattern to Code 2 (BNC English) in this regard. It is relatively rare in thread titles (8.4%), becoming much more common in seed messages (22.9%), and finally rising to become the most frequently used code for both following messages and quoted content messages (approximately 35%). Code 10 (Non-Arithmographemic Latin-scripted Arabic with English), like Code 1 (Arabic-scripted Arabic) is relatively stable across all four variables fluctuating slightly between 2 and 4%. Finally, Code 14
(Non-BNC English) is also relatively stable though it is quite noticeably more common in seed messages.

5.2.2 Thread title code occurrence across the corpus

Thread titles occupy an important place in the corpus. They appear at the top of each discussion thread and present forum visitors with a brief indication of the topic of a given thread. Consequently, besides the generic topical forum name itself, thread titles are one of the first lines of text would-be posters read before deciding whether to contribute to a discussion thread. As such, they play a crucial role in conveying topical information and enticing posters to post. In theory, thread titles can be composed in any of the 14 linguistic codes. However, certain codes were found to be highly favoured for composing thread titles, while other codes, especially biscriptal codes, were virtually excluded from the thread titles. Also worth noting, of all the input fields analysed here, thread titles are unique in that they cannot contain graphics or links. As a result, while a small number of Code 15 (Numeric content-only) thread title samples were found e.g. a thread title consisting of a number or dollar amount or a typographic smiley such as “:)
there were no cases of Code 16 (Graphic content-only) thread titles observed in the corpus. These exceptions aside, Figure 5.2 displays the distribution of Codes 1-14 across thread titles:
Clearly, at the level of the thread title, mono-scriptal codes dominate the corpus. Code 1 (Arabic-scripted Arabic) is the most frequent choice accounting for 41.6% of all thread titles. In close second, Code 2 (BNC English) accounts for 38.4% of all thread titles. The distant third choice is Code 3 (Arithmographemic Latin-scripted Arabic with English) with only 8.4%, followed by Code 14 (Non-BNC English) with a mere 7.3%. Code 10 (Non-Arithmographemic Latin-scripted Arabic with English), is the least commonly used mono-scriptal code in thread titles with only 2.3%. As for biscriptal codes, Code 6 (Arabic-scripted Arabic with BNC English) and Code 8 (Arabic-scripted Arabic with Arithmographemic Latin-scripted Arabic) have the largest percentages with 1% and 0.7% respectively. This indicates that biscriptal codes in general have very little presence in thread titles.

Taken together, these findings highlight the strong preference for both Code 1 (Arabic-scripted Arabic) and Code 2 (BNC English) among posters when composing thread titles. This, in turn,

---

78 Code 15 thread titles which accounted for 2,236 thread titles (i.e. 0.5%) have been omitted from the chart.

Percentages have been recalculated accordingly.
suggests that perhaps Code 1 and Code 2 are favoured for more informational and specifying communication as opposed to phatic and involved communication when compared to other available codes such as Code 3 (see Chapter 6). Or perhaps because they are relatively short, thread titles do not provide text-producers as much opportunity to code or script-switch. Regardless of the reason, the prominence of Code 2 in thread titles becomes even more pronounced when Code 2’s frequency is considered as seen below.

5.2.3 Seed message code frequencies across the corpus

As with thread titles, seed messages as the initial messages in a thread are important in establishing the topic of the thread and inviting responses and replies (Wodak & Wright, 2007). Unlike thread titles, however, the seed messages can be composed of a simple graphic or link to a separate audiovisual file such as a Youtube vignette. In fact, a total of 1,398 Code 16 (Graphic content-only) seed messages were found in the entire corpus which represented 6.5% of all seed messages. This indicates that seed messages occasionally introduce graphic content and/or links in order to start off a discussion topic. Nevertheless, as mentioned above, due to limitations in analysing this highly variable, non-textual data as well as its relative paucity, Code 16 data will not be explored further in this study.

Returning to the remaining 14 codes, it is interesting to compare their respective rates of occurrence within seed messages and thread titles (see Figure 5.3 below):
Compared to thread title code distribution, several differences are notable regarding seed messages. Code 1 (Arabic-scripted Arabic) is the most common seed message code across the whole corpus at 37.1%, representing over one third of all seed messages. In contrast, Code 2 (BNC English) has a mere 13.7% compared with the 38.4% that it accounted for within thread titles. This signifies a 66% drop in frequency. Meanwhile, Code 3 (Arithmographemic Latin-scripted Arabic with English) has increased by over 14%, signifying that it has almost tripled in percentage compared to its presence in thread titles. Code 14 (Non-BNC English) has also risen slightly in occurrence from 7.2% for thread titles up to almost 10% of all seed messages. The remaining mono-scriptal code, Code 10 (Non-arithmographemic Latin-scripted Arabic with English) is relatively constant at just over 3%. Concerning biscriptal codes, Code 6 (Arabic-scripted Arabic with BNC English) has increased dramatically from 1% to 7.3%. Similarly, Code 7 (Arabic-scripted Arabic with Latin-scripted Arabic) has increased to almost 3% from 0.1% for

---

79 As noted above, Code 16 with 19,364 seed messages (i.e. 4.2% of all seed messages) and Code 15 with 108 (i.e 0.02%) have been omitted from Figure 5.2 and percentages for the rest of the codes have been adjusted accordingly.
thread titles. Still, on the whole, the biscriptal codes are relatively marginal when considered individually and when combined account for only 13.3% of all seed messages. The most striking finding here, perhaps, is that Code 2 (BNC English) is saliently less frequent in seed messages than it was in thread titles. This observation alone calls into question the status of English as the main language on the English-language website of mahjoob.com. Also, that Code 3 (Arithmographemic Latin-scripted Arabic with English) should feature so much more prominently in seed messages when compared to thread titles suggests that there may be different stylistic values attached to these various codes (see Chapter 6).

5.2.4 Following message code frequencies across the entire corpus

Following messages are the largest grouping of texts in the corpus, representing all of the messages which make up each thread minus the seed messages (see Figure 5.4 below). Following messages were treated separately from seed messages for two reasons. First, unlike seed messages, they do not explicitly introduce new topics within threads, so if a forum’s overall topic has any bearing on code choice, then these messages should readily reflect this. Second, since following messages represent the overwhelming majority of messages in the corpus measuring their relationship to topic is considered a stronger indication of whether code choice is influenced by the topical forum in which these messages are posted. Indeed, Wodak and Wright (2007) discovered that there were salient differences in code use between seed messages and following messages in their study of multilingual websites in the EU.

Figure 5.4 below provides overall frequencies of codes in following messages:
Some interesting findings regarding code distribution across thread titles, seed messages, and following messages are apparent here. It should be noted that quoted message content is not included here as it occurs originally in the corpus as other seed messages or following messages.

First, Code 1 (Arabic-scripted English) while relatively stable across thread title and seed message has decreased in occurrence within following messages: Code 1 accounts for only 32.3% of all following messages, a clear decrease from the 41.6% which Code 1 represented among thread titles and the 37.1% which it accounted for among seed messages. At 32.3% of all following messages, Code 1 accounts for roughly one third of all following messages. Code 2 (BNC English), on the other hand, accounts for only 17.5% of all following messages, i.e. Code 2 accounts for roughly half the number of following messages that Code 1 comprises. In
comparison with Code 2’s seed message percentage, at 17% of all following messages Code 2’s percentage in following messages represents a 4% increase from its percentage among seed messages where it accounted for only 13.7%. Nevertheless, with 13.7% and 17.5% for seed messages and following messages respectively, Code 2’s frequency has clearly dropped considerably when compared to its presence in thread titles which was well over one third at 38.4%.

One intriguing finding, however, is that the trend toward ever more use of Code 3 (Arithmographemic Latin-scripted Arabic with English) first noted in seed messages is highly conspicuous here. Again, this may reveal that when textual data is longer or more abundant, it is more likely to encounter code-switching or that Code 3 is relatively more phatic and less informational in its function and uses (see Chapter 6). Consider that between thread title and seed message there was an increase in its occurrence from 8.5% to 22.4%. This trend continues with following messages: Code 3 accounts for 35.5% of all following messages, ranking it as the most commonly encountered code among these. Regarding the other mono-scriptal codes, Code 10 (Non-arithmographemic Latin-scripted Arabic with English) increased slightly to 4.1% of all following messages while Code 14 dropped slightly from 9.9% of all seed messages to 7.5% of all following messages.

As for the nine biscriptal codes, again, these codes accounted for an almost negligible number of following messages. Together, they accounted for slightly over 3% of all following messages. Evidently, biscriptal following messages are a relative rarity in the corpus, a fact that will ultimately necessitate their conflation (see 5.2.6 below).
5.2.5 Quoted message code frequencies across the corpus

A total of 295,372 following messages (i.e. 64% of all following messages) contained quoted material from previously posted messages. In other words, these messages were responses to earlier messages in a given thread and included the quoted material which was being responded to. It is important to mention that in the following analysis quoted content messages are treated as entire messages i.e. they include both the originally quoted parts and any new parts composed in response to the quoted parts.

Figure 5.5 was adjusted to exclude Code 15 (Number content-only) and Code 16 (Graphic content-only) messages, which accounted for 0.2% and 4.1% of all quoted content messages respectively.

![Figure 5.5: Quoted content message code overall percentages](image)

Although a relatively smaller number, quoted content messages show similar distribution patterns to following messages. In fact, all mono-scriptal messages i.e. Code 1 (Arabic-scripted Arabic), Code 2 (BNC English), Code 3 (Arithmographemic Latin-scripted Arabic with English),
Code 10 (Non-arithmographemic Latin-scripted Arabic with English), and Code 14 (Non-BNC English) are all relatively stable with little fluctuation in percentages between their following message frequencies and their quoted content message frequencies. As with following messages discussed above in Section 5.2.4, quoted content messages are dominated by quoted content in Code 1, Code 3, and Code 2 in that order. To illustrate, comparing quoted content message percentages to following message percentages the only noticeable changes are that Code 1 which comprised 32.3% of non-quoting following messages accounts for 35.7% of quoted content following messages, and Code 2 which has equivalent figures of 17.5% and 14.3% signalling that the latter’s messages are quoted slightly less than their occurrence in following messages, while Code 1 messages appear to be quoted slightly more often than they occur in following messages.

In order to scrutinize these frequencies for statistical significance, a null hypothesis was formulated that there was no difference in distribution between all 14 codes across the four text-types presented here i.e. thread titles, seed messages, following messages, and quoted content messages. Non-parametric, one-sample chi-square tests carried out using SPSS led to the rejection of the null hypothesis for all four text types given that the p-values in each case were less than 0.001, which is less than the critical value of 0.05, confirming the significantly different distribution of these codes (see Appendix E file "Chapter_5_Chi_squared_tests.spv").

It was also desirable to test significance for the data collected about the remaining aspects of the research question i.e. code use across variables such as overarching forum topic, thread length, and poster frequency. Yet with 14 codes, some of which, had very low counts, cross-tabulation chi-squared tests to test the independence of two variables (e.g. code use and overarching forum
topic) were unadvisable. This is so because chi-square tests cannot establish statistical difference for expected cell counts of less than 5. To address this issue, it was decided to conflate such obscure codes, a process described in the next section.

5.2.6 Conflating biscriptal codes into one code, Code 4 (Biscriptal content-only)

As shown above, the biscriptal codes account for a very small percentage of text types in the corpus. Consequently, as mentioned in Chapter 4, due to their low frequencies, all of the biscriptal codes (i.e. Codes 4-9 and 11-13) were conflated into a single code i.e. Code 4 (Biscriptal content-only) so that chi-squared tests of significance could be carried out more reliably because all expected cell counts in a cross-tabulation were now greater than 5. It is worth mentioning again that Codes 2 (BNC English) and 14 (Non-BNC English) were not conflated nor were Codes 3 (Arithmographemic Latin-scripted Arabic with English) and 10 (Non-arithmographemic Latin-scripted Arabic) despite their similarities. This was because hand-checking of their messages as well as scrutiny of their wordlists suggested certain stylistic differences between them. For instance, in addition to Code 14's non-English content, Code 14 texts also revealed more political content than Code 2 messages, featuring key words such as ISRAEL, ISRAELI, and GOVERNMENT when compared with Code 2 messages. Similarly, Code 10 exhibited relatively more Islam-related lexical content than Code 3. The remainder of this chapter will demonstrate that these four codes were also significantly different in terms of their overall code distributions.

Having reviewed the overall trends toward code occurrence across the whole corpus, it is time to explore patterns of code distribution according to specific variables in the corpus such as topical forum, thread length, and poster frequency.
5.3 Code frequency by overarching forum topic

In Section 2.4.1, several previous studies were cited which indicated that in bilingual web-based discussion forums language use often varies by topical forum (see Androutsopoulos, 2007; McLellan, 2005; Paolillo, 1999). Indeed, this earlier research has provided the motivation for investigating the relationship of topic to code choice in the present study. Consequently, it will be helpful to present all 41 topical forums in terms of the main topics they refer to. Table 5.1 below displays all 41 topical forums along with a brief reference to their primary topical content:

<table>
<thead>
<tr>
<th>No.</th>
<th>Name</th>
<th>Topical Content</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>Suggestions and Complaints</td>
<td>General Discussion</td>
</tr>
<tr>
<td>5</td>
<td>Announcements</td>
<td>General: classifieds</td>
</tr>
<tr>
<td>6</td>
<td>Girls Talk</td>
<td>Male/Female: Female</td>
</tr>
<tr>
<td>7</td>
<td>On Focus</td>
<td>General Discussion</td>
</tr>
<tr>
<td>8</td>
<td>e7ke w fadfed (trans. “Get it off your chest”)</td>
<td>General Discussion</td>
</tr>
<tr>
<td>9</td>
<td>Politics, Business and Economy</td>
<td>Work/Study</td>
</tr>
<tr>
<td>11</td>
<td>Sentiments</td>
<td>General Discussion</td>
</tr>
<tr>
<td>12</td>
<td>General Forum</td>
<td>General Discussion</td>
</tr>
<tr>
<td>14</td>
<td>Sports</td>
<td>Hobby: Sports</td>
</tr>
<tr>
<td>15</td>
<td>Humanities</td>
<td>Hobby: interesting books</td>
</tr>
<tr>
<td>18</td>
<td>A7la Alkalam (trans. “Sweetest Pen”)</td>
<td>Poetry: Arabic language</td>
</tr>
<tr>
<td>19</td>
<td>Kuluna Al Ordun (trans. “All of us are Jordan”)</td>
<td>Local Culture: Jordan</td>
</tr>
<tr>
<td>21</td>
<td>Tech Talk</td>
<td>Hobby: Computers</td>
</tr>
<tr>
<td>22</td>
<td>Teens Thoughts</td>
<td>Male/Female</td>
</tr>
<tr>
<td>24</td>
<td>Health and Science</td>
<td>Work/Study: Health and Science</td>
</tr>
<tr>
<td>25</td>
<td>Men’s corner</td>
<td>Male/Female: Male</td>
</tr>
<tr>
<td>26</td>
<td>Religion Forum</td>
<td>Local Culture: Religion</td>
</tr>
<tr>
<td>27</td>
<td>Entertainment</td>
<td>Hobby: Music</td>
</tr>
<tr>
<td>28</td>
<td>Palestine</td>
<td>Local Culture: Palestine</td>
</tr>
<tr>
<td>29</td>
<td>Iraq Under Occupation</td>
<td>Local Culture: Iraq</td>
</tr>
<tr>
<td>31</td>
<td>Joke Zone</td>
<td>Humorous</td>
</tr>
<tr>
<td>35</td>
<td>Family Matters</td>
<td>Familiar: family-related content</td>
</tr>
<tr>
<td>38</td>
<td>Mobiles (Sub-forum of “Tech Talk”)</td>
<td>Hobby: Mobile phones</td>
</tr>
<tr>
<td>39</td>
<td>Art Gallery</td>
<td>Hobby: Art</td>
</tr>
<tr>
<td>40</td>
<td>Automotives (Sub-forum of “Sports”)</td>
<td>Hobby: Cars</td>
</tr>
</tbody>
</table>

Although there are 41 forums, numerically these range from 3 to 73, reflecting the fact that forums were started at different times and that forums from the Arabic website (not dealt with here) have also started in between the various forums on the English website. Thus, forum no. 3 would have been started before forum no. 5, but forum no. 4, which occurs on the Arabic website, was started after no. 3 but before no. 5.
Structurally, the 41 forums were organized into main forums and sub-forums seen below:

Figure 5.6: All 41 forums displayed as main forums and sub-forums

- Suggestions and Complaints
- Announcements
- Girls Talk
- On Focus
  - SIG (Special Interest Groups)
- e7ke w fadfed (trans. “Get it off your chest”)
- Politics, Business and Economy
- Sentiments
- General Forum
  - Wishes and Greetings
- Sports
  - Automotives
- Humanities
- A7la Alkalam (trans. “Sweetest Pen”)
  - Copied Material
  - Soul Retreat
- Kuluna Al Ordun (trans. “All of us are Jordan”)
- Tech Talk
  - Graphic Design and Architecture
  - Mobiles
- Teens Thoughts
- Health and Science
  - Engineering
- Men’s corner
- Religion Forum
- Entertainment
  - Tv and Movies

Main forums are shown leftmost in List 5.1 while sub-forums are found immediately beneath the main forum in which they occur and are indented to the right. As can be seen here, only eight main forums contain sub-forums.
Considering Figure 5.6 above, it should be noted that in terms of the corpus content, messages posted to main forums are distinct from messages posted to their constituent sub-forum(s). Thus, text and message totals presented for any main forum below represent messages from that forum only and do not include the messages from its sub-forum(s). Instead, these sub-forum messages were totalled separately and treated as distinct forum messages which they in fact are. To illustrate, the *A7la Alkalam* forum is a main forum that contains two sub-forums: *Copied Material* and *Soul Retreat*. Posters can post messages to threads within *A7la Alkalam* itself or they can choose to post messages to one of its two sub-forums. In the forum data below, when *A7la Alkalam* message totals are presented, these signify only messages that were posted directly to threads under the *A7la Alkalam* forum heading and not messages that were posted to either of *A7la Alkalam*’s two sub-forums. In contrast, messages that were in fact posted directly to threads within the *Copied Material* and *Soul Retreat* sub-forums were totalled separately from *A7la Alkalam* main forum messages and are shown as separate forum totals in the data tables below.
As mentioned in 5.2.4, given that the bulk of the corpus is composed of primarily of following messages\(^\text{82}\) and that seed messages were shown to have code distribution patterns different to following messages, following messages are used as the most reliable indicators of the potential relationship between topic and code choice.

**5.3.1 Forums grouped into 8 overarching forum topics**
Following Bentahila’s (1983) study of code-switching in Morocco, suggesting that a number of distinct topics occasioned differentiated code use, all 41 forums were grouped into eight broader topical themes (see Table 5.2 below):

---

\(^{82}\) Out of a total of 460,220 messages in the corpus, there are 438,594 following messages vs. 21,626 seed messages.

In other words, seed messages make up only 4.7% of the corpus.
To explore topic-related code choice further, SPSS descriptive analyses revealed overall percentages of the conflated codes across overarching topics (see Table 5.3):

### Table 5.3: The six conflated codes across overarching forum topics

<table>
<thead>
<tr>
<th>Topic</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>10</th>
<th>14</th>
</tr>
</thead>
<tbody>
<tr>
<td>General Topics</td>
<td>22.3%</td>
<td>22.1%</td>
<td>39.6%</td>
<td>2.2%</td>
<td>4.3%</td>
<td>9.3%</td>
</tr>
<tr>
<td>Humour</td>
<td>66.7%</td>
<td>4.3%</td>
<td>19.5%</td>
<td>5.3%</td>
<td>2.1%</td>
<td>2.0%</td>
</tr>
<tr>
<td>Gender/Age-related</td>
<td>24.4%</td>
<td>13.5%</td>
<td>48.8%</td>
<td>1.4%</td>
<td>5.3%</td>
<td>8.4%</td>
</tr>
<tr>
<td>Hobbies</td>
<td>8.7%</td>
<td>8.7%</td>
<td>44.7%</td>
<td>1.5%</td>
<td>5.9%</td>
<td>13.3%</td>
</tr>
<tr>
<td>Local Culture</td>
<td>39.5%</td>
<td>16.5%</td>
<td>32.0%</td>
<td>5.8%</td>
<td>4.8%</td>
<td>7.9%</td>
</tr>
<tr>
<td>Poetry</td>
<td>70.6%</td>
<td>16.3%</td>
<td>10.2%</td>
<td>7.1%</td>
<td>1.7%</td>
<td>3.0%</td>
</tr>
<tr>
<td>Family/Friends</td>
<td>20.4%</td>
<td>15.9%</td>
<td>49.4%</td>
<td>1.9%</td>
<td>5.5%</td>
<td>6.8%</td>
</tr>
<tr>
<td>Work/Study-related</td>
<td>8.9%</td>
<td>36.4%</td>
<td>28.4%</td>
<td>2.3%</td>
<td>4.9%</td>
<td>19.1%</td>
</tr>
<tr>
<td>Total</td>
<td>32.3%</td>
<td>17.5%</td>
<td>35.5%</td>
<td>3.2%</td>
<td>4.1%</td>
<td>7.5%</td>
</tr>
</tbody>
</table>
Code 1 (Arabic-scripted Arabic) was found to dominate the topics of Poetry, Humour and to a much lesser extent, Local Culture. It was anticipated that humour-related messages might actually be composed in Code 3 (Arithmographemic Latin-scripted Arabic with English) as it is often linked stylistically with Vernacular humour (see Chapter 6). However, it was discovered that Code 1’s presence in the humour topic was accounted for primarily by the Joke Zone forum where it was the dominant code. Indeed, Joke Zone messages account for well over half of all Code 1 messages in the entire corpus. A frequency wordlist of Joke Zone forum’s Code 1 lexis, revealed that its most frequent items were actually Vernacular Arabic items written in Arabic script as opposed Modern Standard Arabic (see Appendix F “Joke_Zone” files). The Local Culture topic which included nationalistic and religious forums was also largely written in Code 1, however Code 3 was relatively common here as well which seemed understandable given the cultural link of such topics to vernacular Arabic. In contrast, Hobbies and Work/Study-related topics were least commonly expressed in Code 1. Instead, these same two topics were most commonly discussed using Code 2 (BNC English) and Code 3, followed by Code 14 (Non-BNC English). Regarding Work/Study-related forums, these had an especially high percentage of Code 14 messages, suggesting that technical neologisms may have been prevalent in these forums. In contrast to Code 1, Code 2 was least common in both the humour and poetry topics. These observations reveal a pattern of Code 1 and Code 2 being diametrically opposed in terms of the topics which they most frequently pair with. Code 3 was found to dominate Family and Friends, Gender/Age-related forums, Hobbies, and General Discussion forums.

Interestingly, Code 4 (Biscriptal content-only) was prevalent for humour as well. Again, this suggests that material may have been imported from other Arabic websites where a certain
amount of Latin script was present alongside Arabic script due to hyperlink-related strings. Meanwhile, Code 10 (Non-arithmographemic Latin-scripted Arabic with English) reveals a relatively low frequency across virtually all topics with 7.5% overall. Still, this is more than double the amount of Code 4 following messages (3.2%), further highlighting the paucity of biscriptal following messages in the corpus.

To sum up, there are several clear patterns of code choice related to topic. Topical forums that are both local and formal in content feature relatively high amounts of Code 1 (Arabic-scripted Arabic). In this regard, Arabic language poetry, Arabic politics and nationalism, and Islamic religion all favour the use of Code 1. The major exception to this formal Arabic usage trend is the preponderance of Vernacular Arabic language jokes in the Joke Zone forum. Regardless, thematically, all of these forums could be described as connected to local Arabic culture. In contrast, topical forums dealing with more specialized technical content especially related to fields of work and study are dominated by Code 2 (BNC English) messages, perhaps because sources for such content are to be found primarily in English language elsewhere on the web (e.g. the Graphic Design and Architecture forum, see folder in Appendix B). Code 3 (Arithmographemic Latin-scripted Arabic with English), the most frequently encountered code in the corpus was found to dominate topical forums which were less formal and more intimate in content (see Chapter 6) as well as forums which encouraged the sharing of general discussion and opinions. In this connection, Hobby-related forums were also dominated by Code 3 messages, though both Code 2 (BNC English) and Code 14 (Non-BNC English) were also frequently featured in such forums. Again, given the non-culturally-localized nature of many of the hobbies, this was not a surprising finding.
These results were again tested using the Chi-squared function in SPSS and were found to be significant given that the p-value was less than 0.001, which is less than the critical value of 0.05 (see Appendix E file "Chapter_5_Chi_squared_tests.spv"). Now that these trends have been reviewed, it will be useful to consider code use across thread length as this may shed further light on how the codes are used differently by authors.

5.4 Code use across threads
In order to uncover salient code usage patterns, it was desirable to establish which codes were used in a typical thread in the corpus. However, given the large number of threads (21,626) and the relative linguistic heterogeneity of these, rather than determine a simple average of code use across all threads, it was decided to group threads into two distinct categories: those that were relatively long, i.e. contained 100 or more following messages, and those with less than 100 following messages. The longer and shorter threads were found to account for approximately 20% and 80% of all messages respectively. As mentioned in Chapter 4, the decision to create these two categories of threads was motivated by previous research which indicated that discussion thread lengths reflected different code choice patterns (see Wodak & Wright, 2007). Consequently, Figures 5.7 and 5.8 below were produced to contrast the following message code distribution patterns among longer threads and shorter threads respectively:
These two figures reveal certain similarities and differences between the longer threads and shorter threads. For instance, Code 1 (Arabic-scripted Arabic) is comparatively more common in
longer threads, comprising 37.5% of all following messages in these threads. Yet in shorter
threads, Code 1 accounts for 34.1% of all messages. These figures highlight the fact that Code 1
is one of the most common codes in the corpus. Code 2 (BNC English) represents 14.3% of all
following messages in longer threads, but rises to 18% of in shorter threads. The occurrence of
Code 3 (Arithmographemic Latin-scripted Arabic with English) also increases slightly in shorter
threads from 34 to 34.9%. The remaining codes show very little fluctuation with Code 4
(Biscriptal content-only) messages decreasing in shorter threads while Codes 10 (Non-
arithmographemic Latin-scripted Arabic) and 14 (Non-BNC English) messages increase slightly.
As for Code 4 (Biscriptal content-only), it is marginally present in both longer and shorter
threads where it accounts for a mere 3.8-3.5% of following messages. Clearly, then, Code 1 is
much more of a mainstay in longer threads, while Code 2 increases and Code 3 retains the same
relative frequency in shorter threads.

To add another dimension to these findings, it was decided to examine whether there were any
differences in topic and code choice between longer and shorter threads (see Figure 5.9 below):
Figure 5.9: Following message code choice by topic and thread length

Overarching topics in longer threads

Overarching topics in shorter threads
Before commencing a comparison of this panelled chart, it is important to bear in mind that the percentages shown here are not overall percentages. As mentioned earlier, messages in longer threads make up only 20% of all messages. Consequently, the percentages in each chart were calculated based on the total number of long thread messages and short thread messages respectively. In this way, the messages from the two types of threads could be compared in terms of their proportional occurrences relative to the rest of messages within their own groupings. Consequently, in the top chart in 5.9 where it shows 32% for the topic of Humour, this percentage refers to the entire number of long thread following messages, but not the entire number of following messages in the corpus. Similarly, the percentages in the bottom chart in Figure 5.9 were calculated solely based on the total number of short thread following messages.

Having clarified these points, a number of observations can be made from these two charts. First of all, among long thread messages, over 30% of these were connected to humour i.e. the Joke Zone forum. In contrast, just over 25% of short thread messages were connected to Humour. The topic of Poetry also contains relatively more messages (i.e. slightly under 10%). However, the topics of Work/Study and Family/Friends are virtually absent among long threads, indicating that short threads are favoured for content involving these topics. Local Culture content occurs proportionally more in short threads, while Hobby-related messages are more frequent in long threads. Gender/Age-related topic messages are proportionally more common in long threads where they comprise well over 25% compared to 20% for short threads. General Discussion-related messages, however, were far more common in short threads comprising 35% compared to only 16% of long thread messages.
Clearly, the Humour, Poetry, Hobbies, and Gender/Age-related topics are proportionally more common in long threads while Work/Study-related topics and Family/Friends are virtually absent. Short thread messages occur proportionally more often within General Discussion and Local Culture forums.

In terms of code use, short thread messages contain relatively more Code 3 (Arithmographemic Latin-scripted Arabic with English) except in Hobby-related and Gender/Age-related forums. Long thread messages favour Code 1 (Arabic-scripted Arabic) for Humour and Poetry. Otherwise, no marked differences can be found between code use among long and short thread messages concerning topic. This suggests that code use is relatively stable over topic regardless of thread length (see Figure 5.10 below To test these findings, two separate chi-squared test were run, one for thread length vs. code use, the other, for thread length vs. topic. In both cases, thread length was found to vary significantly across code and topic respectively where the p-values were less than 0.001 which is less than the critical value of 0.05.(see Appendix E file "Chapter_5_Chi_squared_tests.spv").
5.5 Author-related code use

Thus far, it has been seen that certain codes predominate in certain topics and that thread length has relatively little impact on code choice patterns. Now it will be interesting to explore whether code use varies by type of poster. In order to investigate this further, a similar method was applied to the one used for thread length: the top 10 most prolific posters were grouped into one category and the remaining posters into another. Again, this allowed for a comparison of the groups to see whether there were any marked in code use patterns. In this way, it was possible to glean a better understanding of what prolific and non-prolific posters do with language in the corpus. It needs to be borne in mind again that the data presented in Figure 5.11 below represents percentages within the respective category total of each grouping of posters and does not represent the overall percentages. In terms of overall percentages, however, it is important to note that messages composed by the top 10 most prolific posters account for a full 20% of all following messages in the entire corpus.
In order to compare these two broad categories of posters, consider the panelled display shown in Figure 5.11 below:

**Figure 5.11: Top 10 posters vs. non-Top 10 posters across topic and code**
As was noticed with long threads and short threads, several topical and linguistic trends are common to both groups. For instance, both prolific and non-prolific posters write following messages in all topics. Generally, if a topic is infrequent among the top posters, it is also infrequent among non-top posters. However, there are some key observable differences proportionally between the two groups. The most salient difference is that over 60% of all top poster following messages are found in the Joke Zone forum compared to less than 20% for non-top posters. This indicates that top posters are atypical of the majority of posters in terms of their strong preference for posting to a Humour-related forum. In contrast, non-top posters are relatively more balanced topically: their preference is to post messages in general topic forums roughly 40% of the time. Non-top posters also contribute to Gender/Age-related forums relatively more often with 25% of their messages falling into this category compared to less than 10% for top posters. Hobby-related forum messages account for 10% of non-top poster messages whereas they comprise roughly 5% among top posters. Local Culture-related forums are another area where non-top posters post relatively more messages. On the other hand, top posters compose messages in poetry-related forums relatively more often than non-top posters do.

Linguistically, top posters are notable for greater use of Code 1 (Arabic-scripted Arabic). This is not surprising given their tendency to post to Humour and Poetry-related forums which have been shown to be connected to Code 1 in the corpus. However, top posters also appear to use Code 1 relatively more often for General Topic messages at about the same rate that they use

---

83 This appears to be connected to the fact that so many of these texts are copied from other sources on the Internet. And copying and pasting is typically easier than original composition.
Code 3 (Arithmographemic Latin-scripted Arabic with English) for these. In contrast, non-top posters tend to use Code 2 (BNC English) and Code 3 (Arithmographemic Latin-scripted Arabic with English) far more often. Indeed, for Hobby forums, Gender/Age-related forums, and General Discussion forums, the non-top posters prefer Code 3 and, to a lesser extent, Code 2.

To sum up, top posters contribute to Humour and Poetry-related forums more often and make use of Code 1 (Arabic-scripted Arabic) primarily for these. In this sense, they are atypical of the average non-prolific poster who frequents General Discussion, Gender/Age-related, Hobby, Local Culture, and Family/Friends-related forums relatively more often. Indeed, non-prolific posters appear to be more diverse in their use of codes and their preference for forums. This creates a skewed image of the corpus where much Code 1 use is accounted for by a small group of posters, posting in a limited range of forums. Indeed, as was shown earlier in this chapter, Code 1-based Joke Zone following messages alone account for 66.7% of all Code 1 messages in the entire corpus. In this connection, it is also interesting to note that the top posters, who are only ten in total, account for a full 37.5% of all Code 1 following messages whereas the remaining 1,251 posters account for the remaining 62.5% of Code 1 following messages. In fact, these top 10 poster Code 1 messages represent a full 12% of all following messages in the entire corpus. Clearly, the impact of the top posters’ linguistic preference on code distribution in the corpus cannot be ignored (see Figure 5.12 below). Again, these findings were confirmed to be significant using a Chi-squared test where the p-value was found to be less than 0.001, less than the critical value of 0.05 (see Appendix E file "Chapter_5_Chi_squared_tests.spv").
5.6 Conclusions and summary

Several salient findings have been presented in this chapter. These can be summarized under the headings of 1) Code and text type, 2) Overarching topics 3) Thread and 4) Poster as follows:

1) Code and text type:

- Arabic-related codes i.e. Code 1 (Arabic-scripted Arabic), Code 3 (Arithmographemic Latin-scripted Arabic with English) and Code 10 (Non-arithmographemic Latin-scripted Arabic with English) are more prevalent than English Codes 2 (BNC English) and 14 (Non-BNC English) within a corpus drawn from English forums
- Codes 1, 2, and 3 alone account for over 80% of all text types
• Biscriptal codes make up less than 4% of all following messages combined, necessitating their conflation as 'Code 4' (Biscriptal content-only) for Chi-squared testing
• Code 2 is the most prevalent code in Thread titles, less common in seed messages, and the least common code in following messages while Code 3 exhibits a converse pattern

2) Overarching topics:
• Code 1 is most used for Humour, Poetry, and Local Culture
• Code 2 is most used for Work/Study
• Code 3 is most used for General Discussion, Family
• Codes 10, 14, and conflated Code 4 are relatively rare

3) Thread:
• 80% of threads are under 100 messages long
• Shorter threads dominate in General Discussions, longer threads in Poetry and Humour
• Shorter threads contain more Code 3 messages, Longer threads, more Code 1

4) Poster:
• Top 10 posters compose almost 20% of all messages in the corpus, 12% of which in Code 1
• Top 10 posters mainly post Code 1 messages i.e. 64% of all their posts
• Top 10 posters contribute mainly to longer threads
• Top 10 posters post mainly in Humour
• Non-top 10 posters contribute across several topics, especially General Discussion
• Non-top 10 posters post mainly in Code 3 and then Code 2
First, regarding code choice across text type, it is intriguing that both Arabic-scripted Arabic (i.e. Code 1) and Latin-scripted Arabic with English (i.e. Codes 3 and 10) should be so prevalent on a purportedly 'English' section of a website. This seems to underscore the ability and desire of posters to use the languages of their choice despite externally-imposed constraints. The prevalence of Arabic in the English section perhaps highlights the self-expression of posters who are, themselves, linguistically part Arab and part English. Another notable finding is that the five mono-scriptal codes, Code 1 (Arabic-scripted Arabic), Code 2 (BNC English), and Code 3 (Arithmographemic Latin-scripted Arabic with English), Code 10 (Non-arithmographemic Latin-scripted Arabic with English) and Code 14 (Non-BNC English) were the most prevalent codes in the whole corpus across all text types, accounting for well over 90% of all textual content. This reveals that script-switching, while observable in the corpus, is much more the exception than the norm. On reflection, there are a couple of plausible explanations for this. Firstly, perhaps most posters are either unable or unwilling to script-switch (see Al Share, 2007). This may be due to one or more of the following factors: 1) technical/typographical constraints limiting a poster’s easy access to a bilingual keyboard, 2) linguistic constraints reflecting a poster’s inability to use either Latin script or Arabic script, and 3) a poster’s habitual preference for, or comfort in, using one script over another (see Bianchi, 2005). However, problems related to accessing an Arabic keyboard set are less likely because of the fact that the mahjoob.com website actually allows its posters to select Arabic characters when composing messages from its own web-based Arabic keyboard set.

Another surprising finding is that Code 3 (Arithmographemic Latin-scripted Arabic with English) should increase in frequency, moving from thread titles, to seed messages, to following
message while Code 2 (BNC English) exhibits a converse pattern. Perhaps, these codes carry different communicative functions, where Code 2 is more informational in content and Code 3 is more phatic hence the former's prevalence in thread titles and seed messages which are both necessarily informational as they present the topic of each thread. Interestingly, Code 1 (Arabic-scripted Arabic) also features prominently in both thread titles and seed messages when compared to Code 3, suggesting that Code 1 as well often plays an informational function. The informational role of Code 2 is supported by the observation that more technical and professional forums tend to contain relatively more Code 2 while familial and opinion-type forums favour the use of Code 3. Additionally, if the content of several seed messages is imported from other websites as cursory observation seems to indicate (e.g. the quoted material in the seed message of Thread 206940 in Chapter 7), it would only be natural for these imported text sources to be written in either standard Arabic (Code 1) or English (Code 2). However, once a topic has been established and discussion has commenced, the use of Code 3 to convey phatic and/or emotionally-charged content would be expected. In fact, in Chapter 6, it will be seen that Code 3 is most frequently associated with smileys, strongly suggesting that as a linguistic code, it serves a highly phatic function.

The next group of findings reflected code choice patterns in following messages across overarching topic, where Code 1 (Arabic-scripted Arabic), Code 2 (BNC English), and Code 3 (Arithmographemic Latin-scripted Arabic with English) predominated across all eight

---

84 As mentioned in Chapter 4, it was decided to use following messages as the most reliable indicator of the relationship between code choice and topic due to the fact that following messages accounted for over 95% of all messages in the corpus and yet showed different percentage patterns from seed messages.
overarching topics. Several differences among the codes were also noted here. Considering Code 1, its distribution, while stable across the four text types as noted above, was found to be highly variable across topics. For instance, Code 1 was the code of choice for forums from the overarching topics of Poetry, Humour, and, to a lesser extent, Local Culture. While Poetry written in Arabic is not surprising, it was expected that the Humour topic would have tended to Code 3 use because of its connection to vernacular culture. However, one reason for Code 1’s dominance in both of these topics is that material in these forums appears to be frequently copied from other Arabic language websites. Also, as suggested by its overwhelming occurrence in the Joke Zone forum, Code 1 should not be equated with Standard Arabic (see Chapter 6). Regarding Poetry, which has a long and venerable tradition within Arabic literature predating even the Holy Qur’an (see Bentahila, 1983), even if message texts were not copied from other Arabic language websites, typically, original poetic compositions would likely be written in Modern Standard Arabic by forum participants. In contrast, Code 2 was found to be most common in Work/Study-related forums. This finding parallels the fact that across the Middle East excluding Francophone North Africa, professional and academic disciplines besides religious studies and history are traditionally studied in English (see Section 2.3.1). Again, in a similar fashion to Code 1, which often features copied material from Arabic-language websites, the potential impact of copied material from English-language websites cannot be ruled out in Code 2-dominated forums. Finally, Code 3 was most common among the largest number of topics: General Discussion, Gender/Age-related, Hobbies, Family/Friends-related forums. Interestingly, Code 3-dominated forums appear to encourage the most original composition as opposed to the importation of copied material as seen in the Code 1 and Code 2-dominated forums above. However, the claim that Code 1 and Code 2 following messages feature more copied material than Code 3 following
messages is difficult to substantiate fully. This is so because of the relative ease with which message composers can copy and paste texts from across the Internet, rendering it difficult to ascertain the origin of such texts unless each suspected imported message is subjected to the time-consuming and laborious process of investigation using a search engine such as Google (Hoffmann, 2007). But even then, sometimes the original source of the text has been removed from the Internet, obscuring the ultimate source of authorship of the text.

Concerning code occurrence patterns across thread length, longer thread following messages tended to be written in Code 1 (Arabic-scripted Arabic) messages while shorter thread following messages were composed mostly in Code 3 (Arithmographemic Latin-scripted Arabic with English) or else in Code 2 (BNC English). In terms of overarching topics, longer thread following messages occurred most often in Humour and Poetry-related forums while shorter thread following messages were found most often in General Discussion, Work/Study, and Family/Friends-related forums. Again, this finding is not altogether surprising given the earlier finding that Humour and Poetry-related forums also contained relatively more Code 1 following messages than messages composed in other codes, whereas General Discussion and Family/Friends-related forums were dominated by Code 3 following messages, and Work/Study-related forums were dominated by Code 2 following messages. Notable exceptions to these trends were the Hobbies, Gender/Age-related, and Local Culture-related forums, which were dominated by Code 3, Code 3, and Code 1 respectively. In the cases of the Hobbies and Gender/Age-related forums, longer threads were more frequent despite the prevalent use of Code 3 in these forums. Conversely, the Local Culture-related forums, which tended toward Code 1 use, featured relatively more shorter thread following messages.
In terms of poster behaviour, the most remarkable finding was the impact of Top 10 posters on the corpus: these ten authors accounted for almost 20% of all following messages. Alone, they composed 12% of all Code 1 (Arabic-scripted Arabic) following messages in the whole corpus. This seemed to reflect their strong tendency to post their messages in the Joke Zone forum or in Poetry-related forums. These findings help to explain why Code 1 is the most common code in both Humour-related and Poetry-related forums. In contrast, non-prolific posters tended to use Code 3 (Arithmographemic Latin-scripted Arabic with English) and to post across a wider range of topics especially General Discussion, Gender/Age-related, Hobby-related, and Local Culture-related forums. Thus, the average mahjoobian poster favours Code 3, a point underscored in Chapter 7.

In order to investigate further the relationship between poster and code choice, Chapter 7 will explore salient linguistic behaviour on the part of a select number of posters as a means of identifying potential examples of identity-construction work achieved through strategic code choice. But first, it will be useful to explore in greater depth the potential relationship between topic and code as suggested by the quantitative data in this chapter. To this end, the next chapter will present and contrast the results of a discourse analysis of the three dominant codes in the corpus i.e. Codes 1 (Arabic-scripted Arabic), 2 (BNC English), and 3 (Arithmographemic Latin-scripted Arabic with English) in terms of their respective frequent topical content.
Chapter 6: Discursive tendencies of the main codes on Mahjoob.com

6.1 Overview

This chapter addresses the third major research question of this study:

3) **What do frequent lexical patterns in the main script-code pairings in the corpus reveal about the topical content and stylistic features of those pairings?**

In practical terms, this question entails an examination of the contrastive lexical tendencies of the major script-code pairings i.e. codes found on mahjoob.com as indicators of topical focus among them. Having identified in the previous chapter that the mono-scriptal codes of Code 1 (Arabic-scripted Arabic), Code 2 (BNC English), and Code 3 (Arithmographemic Latin-scripted Arabic with English) were the most prevalent linguistic codes across the whole corpus (see Figure 5.1 above)\(^85\), this chapter will explore the topical and stylistic characteristics of these three codes as revealed by the most frequently encountered lexical words in their frequency wordlists derived from their respective sub-corpora, highlighting unexpected findings of this analysis.

In this connection, it is worth highlighting the key methodological differences between this and the preceding chapter. Chapter 5 helped to establish that certain forum topics occurred more frequently with certain linguistic codes by using an indirect method that relied on relating forum topics to code use based on sampling messages from the forums, reading these, and concluding

\(^85\) As mentioned in Chapter 5, together, Codes 1, 2, and 3 accounted for roughly 90% of all thread titles, 70% of all seed messages, 85% of all following messages, and 95% of all quoted content messages among the linguistic codes in the entire corpus. It should be noted that these percentages exclude Code 15 and 16 which were found to contain no inputted verbal text data and consisted entirely of imported graphic content and numerics respectively.
that a certain topical forum such as “Copied material” was in fact dedicated to sharing of poetry copied from other websites. These forums were then coded into broader topics such as Work/Study, Hobbies, etc. that could then be analysed for code use patterns. The approach in this chapter differs from the previous chapter in a couple of important aspects. First, it takes the actual frequent lexis of a given linguistic code as the starting point for analysis without making any a priori assumptions about what topic might be articulated with a given code, offering a method of triangulating the findings of Chapter 5. Second, by employing a close textual reading of specific messages in the corpus, it is possible to determine what, if any, stylistic differences exist between the main codes in the corpus, especially regarding textual features such as involvedness vs. informationality, formality vs. informality, and seriousness vs. humour (see Biber, Conrad, & Reppen, 1998). And as revealed in Chapter 7, understanding such stylistic differences among codes will be crucial for examining instances of identity-related code choice.

6.2 Some methodological issues

Before beginning the analysis, it is worth discussing a few points of methodology. First, the rationale for concentrating on the three mono-scriptal codes will be given. Next, some relevant issues surrounding frequency will be considered. Finally, the limitations of the analysis presented here will be discussed.

Apart from their sheer volume in the corpus, one compelling reason to concentrate on Code 1 (Arabic-scripted Arabic), Code 2 (BNC English), and Code 3 (Arithmographemic Latin-scripted Arabic with English) is the fact that, lexically, the biscriptal codes are actually derived from combinations of these mono-scriptal codes. For instance, as mentioned in Chapter 4, the non-conflated Code 4 (not to be confused with conflated Code 4, the amalgam of all biscriptal codes
mentioned in Chapter 5) was defined as BNC English with a lesser amount of Arabic-scripted
Arabic, while Code 5 was based on Latin-scripted Arabic with a lesser amount of Arabic-
scripted Arabic and Code 6 was composed of Arabic-scripted Arabic with some BNC English.
Consequently, by examining the lexical features of the mono-scriptal codes, it is expected that
important clues about the biscriptal codes may be discovered, which might help to account for
why forum posters would have reason to code and script-switch within their texts (see Chapter
7).

Regarding the use of frequency lists, within corpus linguistics, there is no commonly agreed
upon definition of ‘frequent’ vs. ‘infrequent’. Indeed, these labels can only be applied in relative
terms. To illustrate, consider that WordSmith 5.0 software generates a frequency wordlist based
on the number of occurrences of a word (i.e. a token) within a corpus composed of many tokens.
The tokens which occur statistically more often than others are then ranked in descending order.
In the case of many languages such as English and Arabic, grammatical words i.e. articles,
quantifiers, prepositions, etc. occur more frequently than lexical words i.e. proper nouns, non-
modal verbs, descriptive adjectives and adverbs, etc. (Baker, 2006, pp. 121-150). Still, when
considered from the point of view of their occurrences as a percentage of the overall occurrence
of all words in a corpus, the most frequent words are often found to occur with what appear to be
very low percentages. For instance, in the present research, the grammatical word ‘the’ as
encountered within the sub-corpus of Code 2 (BNC English) messages is actually the most
frequent word (see Appendix F, Code 2 wordlist). Nevertheless, its actual occurrence is only
3.5% i.e. out of an average of every 100 words in the Code 2 sub-corpus, the word ‘the’ occurs
only 3.5 times per 100 words. In terms of distribution, ‘the’ is found in 38% of all threads which
contain Code 2. Clearly, these small numbers might lead one to conclude that ‘the’ is relatively obscure in Code 2. However, in relative terms, when compared with all other lexical items, ‘the’ is still the most frequent word. The lesson here is that even the most frequent word within a given corpus will typically yield a deceptively low percentage.

Consequently, the method adopted in this chapter is to first identify the ten most frequent open class lexical items in each of the three main monoscriptal codes in the entire corpus i.e. Codes 1 (Arabic-scripted Arabic), 2 (BNC English), and 3 (Arithmographemic Latin-scripted Arabic with English). This “first brush” gives an overall sense of what the topical foci of each of these codes might be. Next, the top 10 frequent words of each of these codes words are hand-checked using a 100-line concordance in order to establish their respective usage patterns in the corpus, suggesting, in turn, broad stylistic differences between the codes themselves.

6.2.1 Limitations and other considerations
This chapter describes only the broadest salient topical patterns associated with each code as suggested by the top 10 frequent lexical items along with examination of random samples of 100 concordance lines of each of these frequent items. These highly frequent words are used as the measure in determining to what extent each code resembles or differs from others in terms of topical content (see Baker, 2006, pp. 121-150). Where specific topics, references, and functions are cited for the concordance line of a specific lexical item, it is important to bear in mind that these were determined solely by inferring them from the immediate context of the item within the boundaries of its concordance line of between 10-15 words. This was done because the time-consuming process of referring back to the original message for each of the 3,000 concordance lines in order to specify beyond doubt the topic of each concordance line would have proven
highly unfeasible. Rather, the ensemble of concordance lines was used in a holistic way to help further inform the inferences about what the topic and reference of specific instances of a lexical word within its concordance might have been. Furthermore, additional clues as to topic, reference, and function of an item were provided by the presence of smileys and other stylistic features such as standard grammar and formal vocabulary. As a result, there were several lines within each concordance whose topic could not be identified with absolutely certainty. However, such lines often still exhibited stylistic features such as smileys or discursive functions such as criticisms. In several cases, an utterance could be construed as belonging to more than one topic e.g. wearing hijab as a form of female dress or as a form of Islamic practice. Again, the whole concordance was relied upon when making such judgements regarding topics. Inevitably, though, overlapping and blurring of boundaries between topics is often apparent (cf. Callahan, 2004: 70), although I have attempted to be consistent in my categorizations and judgements.

A further point needs to be noted here: several related topics are likely articulated across all three major codes. For instance, religious topics can easily be discussed in either of Code 1 (Arabic-scripted Arabic), Code 2 (BNC English), or Code 3 (Arithmographemic Latin-scripted Arabic with English). Thus, the analyses presented in this chapter should not be considered in any way exhaustive of all possible topics or even of all major topics associated within a given code. Nevertheless, by examining the most frequent lexical patterns associated with each code, certain salient linguistic tendencies become apparent, indicating that a given code is generally favoured for discussing a certain kind of topic or for conveying a certain kind of style or level of formality and familiarity. I argue that this broad brush heuristic method has made it possible to determine that certain codes are indeed associated with specific topics and styles by their users (Su, 2003).
The same may be said of the relationship between topic and single lexical items. For instance, each occurrence of a term such as allāh ﷺ ‘Allah/God’ within Code 1 (Arabic-scripted Arabic) messages should not be interpreted automatically as a religion-related reference (see 6.4.1 below). Indeed, in popular speech allāh ﷺ ‘Allah/God’ is often used as an exclamation to express a wide variety of emotions (similar to the English “O God!” and “O my God!”). Furthermore, allāh ﷺ ‘Allah/God’ often occurs alongside other words in various other expressions and to form certain compound names via the Arabic construct case. It is precisely because of such multiple uses that concordance lines provide valuable insights. It should be stressed, though, that concordance analyses cannot establish absolute lexical usage patterns. At best, concordances merely reveal broader topical and stylistic tendencies. Bearing this limitation in mind, one important rationale for using a concordance approach is that, in order for a linguistic code to provide any sort of indexicality, it must first be interpreted by its users to be

86 The construct syntactic formation or structure refers to a grammatical feature of both Modern Standard and Colloquial Arabic whereby grammatical possession is expressed primarily through word order. Essentially, in a construct phrase, the ‘possessor’ noun is placed immediately after the ‘possessed’ noun which loses any articles and may undergo slight morphological changes. To illustrate, English possessive phrases such as ‘Ahmed’s friend’ or ‘The friend of Ahmed’ are both rendered into Arabic syntactically as سدئق أحمد (lit. ‘friend Ahmed’). In Arabic, these words are not joined in any way so they appear as two distinct nouns. Context provides the cues to interpret the phrase as a possessive construction. Typically, construct phrases when featured as proper names in Arabic are transliterated into English as single terms with the possessor noun affixed to the possessed noun. Consider the following examples: Hezbollah (‘party God’) meaning ‘(political) Party of God’, (Brunei) Darussulam meaning (‘House Peace’), Darfur (‘House Fur’) i.e. Fur Region).
habitually associated with specific topics, domains, and styles (Su, 2003). By using concordances, these associations can be profitably measured, looking at frequent broader lexical patterns in context across a large number of texts. These patterns then shed light on recurrent and therefore dominant topical and stylistic patterns associated with a given linguistic code.

A couple of final notes on the citation of example lines from the frequent lexis concordances are in order. Regarding the presentation of the concordance lines cited in this chapter: First, rather than writing out in full “Line 1” I employ the shorthand “L1” here. Second, in each boxed concordance line, the top 10 concordance word has been bolded to set it apart from the other words in the concordance line. Third, where present, smileys are indicated by italics. Fourth, Arabic-scripted English text is featured in bolded and italicized text. These same conventions are used for the translations of concordance lines provided below the original boxed concordance lines where necessary.

Having clarified these points, the analysis begins with a contrastive overview of the top 10 most frequent lexis across Codes 1 (Arabic-scripted Arabic), 2 (BNC English), and 3 (Arithmographemic Latin-scripted Arabic with English).

6.3 The Top 10 lexical words in the wordlists of Codes 1, 2, and 3

In order to provide a general sense of the kinds of words which feature in Code 1 (Arabic-scripted Arabic), Code 2 (BNC English), and Code 3 (Arithmographemic Latin-scripted Arabic with English), Figure 6.1 below was compiled to display the top 10 lexical words for all three codes (note that grammatical or ‘closed-class’ words such as pronouns, articles, determiners, modal verbs, auxiliary verbs, conjunctions and prepositions are not included in the table). Instead, the focus here is on open class ‘content words’ (lexical nouns, adjectives, verbs and adverbs), which help to reveal more about topics:
Table 6.1: Top 10 lexical words across Codes 1, 2, and 3

<table>
<thead>
<tr>
<th></th>
<th>Code 1</th>
<th>Code 2</th>
<th>Code 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>الله</td>
<td>KNOW</td>
<td>ALLAH</td>
</tr>
<tr>
<td></td>
<td>God</td>
<td></td>
<td>God</td>
</tr>
<tr>
<td>2</td>
<td>قال</td>
<td>THINK</td>
<td>KNOW</td>
</tr>
<tr>
<td></td>
<td>He said</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>والله</td>
<td>GOOD</td>
<td>THINK</td>
</tr>
<tr>
<td></td>
<td>By God</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>الناس</td>
<td>PEOPLE</td>
<td>LOVE</td>
</tr>
<tr>
<td></td>
<td>People</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>يوم</td>
<td>LOVE</td>
<td>TIME</td>
</tr>
<tr>
<td></td>
<td>day</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>صلى</td>
<td>TIME</td>
<td>GOOD</td>
</tr>
<tr>
<td></td>
<td>He blessed</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>يبني</td>
<td>SEE</td>
<td>WALLAH</td>
</tr>
<tr>
<td></td>
<td>I want</td>
<td></td>
<td>By God</td>
</tr>
<tr>
<td>8</td>
<td>اليوم</td>
<td>GO</td>
<td>MAN</td>
</tr>
<tr>
<td></td>
<td>The day</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>وسلم</td>
<td>THANKS</td>
<td>PEOPLE</td>
</tr>
<tr>
<td></td>
<td>And he saved</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>طيب</td>
<td>WANT</td>
<td>WAY</td>
</tr>
<tr>
<td></td>
<td>Good</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 6.1 reveals a number of interesting lexico-semantical similarities across the codes. For instance, Code 1 (Arabic-scripted Arabic), Code 2 (BNC English), and Code 3 (Arithmographemic Latin-scripted Arabic with English) share one semantically-related highly frequent concept in common: ‘people’. This indicates that in all three codes references to people are common, suggesting that perhaps the topic of ‘people’ or generalizing statements employing

---

87 Words in the Code 2 and Code 3 lists are given in capitals reflecting the WordSmith 5.0 convention of displaying frequency wordlist items in capitals.

88 This word can also be translated as ‘and God’ according to context.

89 This word can also be translated as ‘and God’ according to context.
the word ‘people’ may be prevalent across all three codes. Another concept that these codes have in common is ‘good’ (see Code 1: item 10, Code 2: item 3, and Code 3: item 6) Again, on the surface, these words imply that something (or someone) is frequently described in a positive manner.

Other sets of similarities are discernible between these three codes. For instance, in addition to the concept of ‘people’, Codes 1 (Arabic-scripted Arabic) and 3 (Arithmographemic Latin-scripted Arabic with English) also show the concept of ‘Allah/God’ to be highly frequent as both codes feature the words ALLAH (‘God’) and WALLAH (‘By God’ or ‘and God’). Such surface lexical similarities in word list items suggest that perhaps the topic of God or religion may be commonly discussed in both of these codes. When the wordlists of Code 1 and Code 2 are examined in conjunction, again, considerable overlap is apparent. Codes 2 and 3 also share a number of lexical items in common. In fact, these codes feature a total of six identical top 10 words in their respective sub-corpora. In addition to the words PEOPLE and GOOD (which also had semantic counterparts in Code 1), Codes 2 and 3 have four other top 10 words in common: KNOW, THINK, TIME, and LOVE. The words KNOW, THINK, and LOVE suggest that personal viewpoints, opinions, and feelings may often be expressed frequently in Codes 2 and 3. As an aside, the fact that Code 3 shares semantically-related concepts in common with both Code 1 and Code 2 serves to underscore Code 3’s code-mixed nature as a ‘fused lect’ between Arabic and English (see Auer, 1998; McLellan, 2005).

90 The only difference between these items is the script in which they are composed in each code i.e. in Code 1 the words are written in Arabic script as الله and الله while in Code 3 these same words are written in Latin script as ALLAH and WALLAH respectively.
An important cautionary note needs to be borne in mind: surface similarities should not be taken too uncritically, and without further evidence from samples drawn from specific concordance lines, it would be premature to conclude that these three codes employ the common concepts cited here in the same manner. Indeed, when concordance line data is presented below, ample evidence will be offered to highlight that such seemingly similar lexis is in fact often employed in different ways by users of these three codes.

Having provided a brief overview of the similarities between the three codes, the top 10 list of each code will now be explored in depth beginning with Code 1 (Arabic-scripted Arabic) in the next section.

### 6.3.1 Code 1 (Arabic-scripted Arabic)

Considering Tables 6.2 and 6.3 below, Code 1’s wordlist features four items that appear to be directly related to the topic of God and/or religion: \textit{allāh} الله ‘Allah/God’\textsuperscript{91}, \textit{wallāh} وَلله ‘by Allah/God’, \textit{Salla} صلى ‘may (God) bless (him)’ and \textit{wa-sallam} وسلم ‘and may (God) save (him)’\textsuperscript{92}. The remaining items appear more general: \textit{qāla} قال ‘he said’, \textit{al-nās} الناس ‘people’, \textit{yawm} يوم ‘day’, \textit{baddī} بدي ‘I want’, \textit{al-yawm} اليوم ‘the day’, and \textit{Tayyib} طيب ‘good’.

\textsuperscript{91} Following the works of Brustad (2000) and Holes (2004), Arabic-scripted items are typically cited here by first giving their transliteration (in italicized bold), next the actual item in Arabic script, and third a quoted translation. So, the Arabic scripted item الله is cited as \textit{allāh} الله ‘Allah/God’. This convention is used throughout the remainder of this thesis.

\textsuperscript{92} The expression \textit{wa-sallam} وسلم ‘and may (God) save (him)’ might be translated more accurately as ‘and may (God) grant him salvation’ i.e. admit him into paradise. It is a formulaic honorific expression used exclusively for the Prophet Mohammed.
Table 6.2: Top 10 lexical words in the Code 1 sub-corpus

<table>
<thead>
<tr>
<th>Code 1 Top 10 Lexical Words</th>
<th>Transliteration</th>
<th>Meaning</th>
<th>Occurrence</th>
<th>%93</th>
<th>Threads</th>
<th>%95</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>الله</td>
<td>allāh (‘Allah/God’)</td>
<td>46839</td>
<td>0.8</td>
<td>6013</td>
<td>27.8</td>
</tr>
<tr>
<td>2</td>
<td>قال</td>
<td>qāla (‘he said’)</td>
<td>11516</td>
<td>0.2</td>
<td>2539</td>
<td>11.7</td>
</tr>
<tr>
<td>3</td>
<td>والله</td>
<td>wallāh (‘by God’)</td>
<td>6455</td>
<td>0.1</td>
<td>3117</td>
<td>14.4</td>
</tr>
<tr>
<td>4</td>
<td>الناس</td>
<td>al-nās (‘the people’)</td>
<td>6129</td>
<td>0.1</td>
<td>2247</td>
<td>10.4</td>
</tr>
<tr>
<td>5</td>
<td>يوم</td>
<td>yawm (‘day’)</td>
<td>5772</td>
<td>0.1</td>
<td>2547</td>
<td>11.8</td>
</tr>
<tr>
<td>6</td>
<td>صلى</td>
<td>Salla (‘may (God) bless (him)’)</td>
<td>5646</td>
<td>0.1</td>
<td>721</td>
<td>3.3</td>
</tr>
<tr>
<td>7</td>
<td>بدي</td>
<td>baddī (‘I want’)</td>
<td>5548</td>
<td>0.1</td>
<td>2391</td>
<td>11.1</td>
</tr>
<tr>
<td>8</td>
<td>اليوم</td>
<td>al-yawm (‘the day’)</td>
<td>5436</td>
<td>0.1</td>
<td>2497</td>
<td>11.5</td>
</tr>
<tr>
<td>9</td>
<td>وسلم</td>
<td>wa-sallam (‘and may (God) save (him)’)</td>
<td>5344</td>
<td>0.1</td>
<td>705</td>
<td>3.3</td>
</tr>
<tr>
<td>10</td>
<td>طيب</td>
<td>Tayyib (‘good’)</td>
<td>4459</td>
<td>0.1</td>
<td>2256</td>
<td>10.4</td>
</tr>
</tbody>
</table>

93 These figures refer to the total number of occurrences of the frequent word within the Code 1 (Arabic-scripted Arabic) Sub-corpus. Tables 6.4 and 6.6 have similar columns for their respective sub-corpora.

94 This percentage is based on the ‘Occurrence’ column (see footnote 93 above), reflecting the average number of times the frequent word occurs in every 100 words. So, the first frequent word allāh ḥi ̄ ‘God/Allah’ occurs 0.8 times per 100 words in the sub-corpus. The same column is found in Tables 6.4 and 6.6.

95 The figures in this column indicate the number of threads in which the frequent word is found within Code 1 messages. So, allāh ḥi ̄ ‘God/Allah’ as a word in Code 1 messages is found in 6013 threads in all. The same type of column is found in Tables 6.3 and 6.4.

96 The percentage here is based on the ‘Threads’ column. Taking the example of the frequent word allāh ḥi ̄ ‘Allah/God’ again, the column indicates that 27.8% of all threads feature this frequent word within Code 1 messages.

97 This word can also be translated as ‘and God’ according to context.
Table 6.3: Code 1 Top 10 lexical words showing topical and linguistic features

<table>
<thead>
<tr>
<th>Rank</th>
<th>Code 1 Top 10 Lexical Words</th>
<th>Transliteration</th>
<th>Meaning</th>
<th>Language: MSA vs. Vern.</th>
<th>Recurrent Topics</th>
<th>Involved vs. Informational</th>
<th>Smileys</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>الله</td>
<td>allāh</td>
<td>‘Allah/God’</td>
<td>MSA 66%</td>
<td>Religion (Islam), Christianity, Palestine</td>
<td>Inf. 60%</td>
<td>30%</td>
</tr>
<tr>
<td>2</td>
<td>قال</td>
<td>qāla</td>
<td>‘he said’</td>
<td>MSA 64%</td>
<td>Religion (Islam, mostly Hadith) 64%, Humour (36%)</td>
<td>Inf. 100%</td>
<td>8%</td>
</tr>
<tr>
<td>3</td>
<td>والله</td>
<td>wallāh</td>
<td>‘by God’</td>
<td>Vern. 69%</td>
<td>General, religious (15%)</td>
<td>Inv. 67%</td>
<td>32%</td>
</tr>
<tr>
<td>4</td>
<td>الناس</td>
<td>al-nās</td>
<td>‘the people’</td>
<td>MSA 55%</td>
<td>General, Religion (Islam, 35%)</td>
<td>Inf. 60%</td>
<td>10%</td>
</tr>
<tr>
<td>5</td>
<td>يوم</td>
<td>yawm</td>
<td>‘day’</td>
<td>Vern. 60%</td>
<td>Islam (25%), Romance, Narratives, politics, Jokes, food</td>
<td>Inv. 61%</td>
<td>19%</td>
</tr>
<tr>
<td>6</td>
<td>صلى</td>
<td>Salla</td>
<td>‘may (God) bless (him)’</td>
<td>MSA 100%</td>
<td>Prophet Mohammed (100%)</td>
<td>Inf. 100%</td>
<td>0%</td>
</tr>
<tr>
<td>7</td>
<td>بدي</td>
<td>baddī</td>
<td>‘I want’</td>
<td>Vern. 100%</td>
<td>General (no religion), songs, food, clothing, relationships</td>
<td>Inv. 100%</td>
<td>47%</td>
</tr>
<tr>
<td>8</td>
<td>اليوم</td>
<td>al-yawm</td>
<td>‘the day i.e. today’</td>
<td>Vern. 50%</td>
<td>General, news, religion</td>
<td>Inv. 66%</td>
<td>32%</td>
</tr>
<tr>
<td>9</td>
<td>وسلم</td>
<td>wa-sallam</td>
<td>‘and may (God) save (him)’</td>
<td>MSA 100%</td>
<td>Prophet Mohammed (100%)</td>
<td>Inf. 100%</td>
<td>0%</td>
</tr>
<tr>
<td>10</td>
<td>طيب</td>
<td>Tayyib</td>
<td>‘good’</td>
<td>Vern. 97%</td>
<td>Jokes, love, food, songs, well-wishing</td>
<td>Inv. 80%</td>
<td>74%</td>
</tr>
</tbody>
</table>

98 This word can also be translated as ‘and God’ according to context.
To investigate these topical assumptions, a random sample of 100 concordance lines was tabulated for each of these top 10 lexical words (the 100-line concordances for each top 10 lexical word in Code 1 are found in Appendix H).

Interestingly, the most frequent open-class lexical word was the noun *allāh* الله ‘Allah/God’.

Since *allāh* الله ‘Allah/God’ appears to be potentially connected to the topic of religion, it was important to examine the 100-line concordance for evidence of this and other uses. As a result, it was found that over two thirds of all 100 concordance lines involving this word were in fact connected to the topic of religion, usually to Islam. And of these 66 specific religion-related uses of *allāh* الله ‘Allah/God’, a large number of these lines made reference to the Prophet Mohammed directly using one of his most common epithets ‘Messenger of Allah/God’ as in L9 (i.e. ‘concordance line 9’):

```
عَلَىٰ...حُرْرَىَّةَبِنَ...اللهِ...ضَيْعُ...فَمَا...وَسَّمَىَهُ...اللهِ...صَلَّىَاللهِ...وَسَلَّمَ...فَقَالَ...;
```

...may (God) be glorified. According to Abu Huraira, verily, the Messenger of *Allah/God*, may *Allah/God* bless and save him, hath said:

Typically, such religion-related uses were written in Modern Standard Arabic as opposed to Vernacular Arabic, signalling a higher level of formality. Further, these religion-related lines tended to be non-involved as opposed to involved i.e. they did not contain first or second person references (see Biber et al., 1998, pp. 145-149). Surprisingly, though, not all of these religious uses of *allāh* الله ‘Allah/God’ were related to Islam. Indeed, a small number of lines were connected to Christianity, e.g. L28:

```
كتِابٍ...مَفْصَلُهَا...اللهِ...وَلَكِنَّهَا...مَا...اللهِ...؛...وَاللهِ...مَنْ...يَأْتِيَ...اللهِ...هُوَ...مَلَكُ...الْجَهَنَّمَ...;
```

...the Holy Bible, “…in that Jesus Christ is *Allah/God* so that all of us know that the divinity...
Another interesting point to note was that stylistically, not all references to *allāh* ‘Allah/God’ occurred in Modern Standard Arabic. Nor were all references to *allāh* ‘Allah/God’ non-involved. Examples of vernacular and involved uses of this word, which accounted for roughly 30% of all 100 concordance lines, usually consisted of well-wishing in the form of blessings or prayers invoked on behalf of either second person addressees, oneself, an inclusive ‘we’ or others such as Palestinian martyrs. Very often such well-wishing concordance lines featured smileys alongside the invocation of blessing as in L68:

![huggingfriend](https://example.com/huggingfriend)

68 *huggingfriend* Allah/God help you in this world and in the next *huggingfriend*...

Besides clear references to the topic of religion, a couple of other instances of the word *allāh* ‘Allah/God’ were connected to Palestine such as the L56 reference to Ramallah (*rām allāh* رام الله), a city in Palestine, and as part of an exclamation within a personal narrative describing the hardships of living in Palestine in L57 (i.e. *allāh ‘ā’lim* الله أعلم ‘God (only) knows’). Other narrative contexts were found in the concordance in which *allāh* ‘Allah/God’ occurred as part of idiomatic expressions of exclamation or disbelief. In this regard, it is interesting to observe that in one case *allāh* ‘Allah/God’ was actually used to curse someone.

This brief description of the contexts of use and functions of *allāh* ‘Allah/God’ in Code 1 (Arabic-scripted Arabic) messages, has highlighted the range of uses which this word serves as well as the kinds of topics that can be articulated in Code 1. It also underscores the stylistic

---

99 In *allāh* ‘Allah/God’ concordance L68 featured above, note the presence of the smileys *huggingfriend* written in Latin script. This use of Latin script in what are otherwise entirely Arabic-scripted texts is a by-product of the conversion process whereby the graphical smileys were assigned alphabetic names in order to facilitate their processing and analysis in WordSmith 5.0 as discussed in Chapter 3.
versatility of Arabic-scripted Arabic within the corpus, ranging from serious informational religion-related utterances to light-hearted, involved phatic purposes such as greetings and blessings.

The second most frequent lexical word in the Code 1 (Arabic-scripted Arabic) list was the verb *qāla* ‘he said’. It also revealed an interesting dichotomy between its use in Classical Arabic texts on religious topics (62 lines alone quoted Prophet Mohammed) and its appearance in jokes in Vernacular Arabic as seen here

> وش رأيك في الزواج المبكر... قال يعني الساعه كم 2 عسكري غلب

“What do you think about getting married early...?” He said, “Like, what time?” A soldier was absent...

The next frequent word in Code 1 (Arabic-scripted Arabic) was *wallāh* ‘by Allah/God’. On the surface, this one-word expression is semantically close to *allāh* ‘Allah/God’ discussed above. However, in its most common usage, this word functions as a discourse marker in Vernacular Arabic to convey sincerity and emphasis and might be best translated into English as ‘really’, ‘truly’, or more literally as ‘I swear to God’. Another common use of this expression is as a question in order to convey surprise or disbelief upon hearing unexpected or dubious information. By hand-checking the 100 random concordance lines for this expression, the word *wallāh* ‘by Allah/God’ was found to occur regularly in involved utterances. In fact, out of 100 lines, 67 were found to be involved as opposed to non-involved. Stylistically, a similar number of lines (69 in total) were found to be written exclusively in Vernacular Arabic, with near total overlap between involved and vernacular concordance lines as L33 illustrates:

---

100 A second possible meaning is ‘and God’ based on the fact that the word *wallāh* وَالله actually consists of *allāh* اللّه ‘Allah/God’ prefixed by the conjunctive particle *wa* و ‘and’. 
As exemplified here, almost one third of all its concordance lines contained smileys, highlighting its informal, emotive style. However, despite such involved and expressive uses, 15 lines featuring this word were found to mean “and Allah/God” within Islam-related texts. Otherwise, topically, wallāh ‘by Allah/God’ occurred mainly with a range of general topics. Again, a clear distinction between informal Vernacular uses and formal Classical Arabic uses is apparent here. Topically, The next frequent word in Code 1 (Arabic-scripted Arabic) was al-nās ‘the people’. As mentioned in Section 6.1 above, this word suggests that perhaps generalizations involving ‘people’ are frequently made in Code 1. Its concordance lines seemed to confirm this as well over half of all the concordance lines revealed the use of al-nās ‘the people’ in a generic way to make statements or questions referring to people in general across a wide variety of topics such as succeeding in life:

...between us Buggedouteyes I don’t understand how (the) people get ahead (in life) eyeswatering This example illustrates its use in general references to people written in Vernacular Arabic with smileys. Yet, 55 concordance lines written in Modern Standard Arabic were also found. Such formalized utterances were evident in L33 (driven mad by people), L34 (people under pressure), L37 (relationships between people), L59 (judging people), and L78 (how to treat people).

Interestingly, though, the majority of these stylistically formal references were actually found to be related thematically to Islam. Islam-related references accounted for well over one third of the
entire concordance. A recurrent pattern among these lines was the quotation of the Qur’an or Hadith, where Mohammed addresses people collectively:

صلى الله عليه وسلم : “يا أئها الناس ألا إن ربك واحد وأن آب
May God bless and save him (Prophet Mohammed): “O ye (the) People! Is your Lord not one and father (of)...

The remaining top 10 lexical words in the Code 1 (Arabic-scripted Arabic) sub-corpus, could be described as reflecting formal, Vernacular, and mixed stylistic purposes. Regarding formal uses, the two highly frequent verbs *Salla صلى* ‘may (God) bless (him)’ and *wa-sallam وسلم* ‘and may (God) save (him)’ are Classical Arabic lexical items that co-occur routinely in the formulaic expression *Salla allāh ‘alayhi wa sallam صلى الله عليه وسلم* ‘may God bless and save him’, reserved solely for the Prophet Mohammed. The presence of these two verbs within the top 10 frequent lexical item list for Code 1 highlights a clear connection between Code 1 and Islam especially for narrative and informational purposes.

In stark contrast, the inherently Vernacular Arabic item *baddī بدي* ‘I want’, reflects the frequently involved and personalized use of Code 1 (Arabic-scripted Arabic). All of its concordance lines reflect *de facto* involved as opposed to an informational discursive style since *baddī بدي* ‘I want’ is grammatically first person singular. Further, almost half of its lines, 47 in total, exhibited one or more smileys. Widespread use of smileys betrays a humorous, playful style as in L27 where the poster appears to feign outrage:

---

101 In English texts written by Muslims this same phrase is generally translated as “may the peace and blessings of Allah be upon him” or simply “peace be upon him” (‘PBUH’ is its common abbreviated form) (see Chapter 7).

102 Although grammatically not a true verb, *baddī بدي* ‘I want’ (lit. ‘my wish’) functions verbally and takes predicate noun and verb phrases (Brustad, 2000, p. 155).
While covering a wide range of topics, conspicuously, religion-related concordance lines were absent here, suggesting that Vernacular Arabic style with smileys is inappropriate for such a weighty topic. Further highlighting the linguistic flexibility of Code 1 (Arabic-scripted Arabic) is the fact that one line featured the relatively rare phenomenon of script-switching with Arabic-script used to transcribe (here indicated by bold italicized font and transliterated with English below):

\[\text{Abrīshīyт yūūr 'ayfūrts mān, I wanted to ask you for some za'atar, but the sesame in it is too expensive.}
\]
\[\text{Appreciate your efforts, man}\]

Again, such atypical linguistic behaviour indicates that not only is Code 1 (Arabic-scripted Arabic) used for relatively informal and playful vernacular communication, but that it can also contain even more playful samples of language such as Arabic-scripted English.

Another, predominantly Vernacular Arabic word was \textit{Tayyib} ‘good’. Though found in both Modern Standard Arabic and Vernacular Arabic, as with the Vernacular form \textit{baddī} ‘I want’ mentioned above, \textit{Tayyib} ‘good’ occurred in Vernacular contexts in virtually all of its concordance lines (97%). The most common use of \textit{Tayyib} ‘good’ was as a discourse marker meaning “okay”, “alright”, “fine then” in order to issue a challenge to the addressee(s) in the form of a rhetorical question, an imperative, a counter-assertion, or to express indignation or disapproval e.g.:

\[\text{Za’atar is a popular Middle Eastern mix of dry herbs (thyme) and spices (salt, sumaq, etc.) used to flavor breads and pastries. A key ingredient in a typical za’atar mix is whole sesame seeds, which the poster here complains are too expensive and therefore inhibit him from asking his addressee to give him some.}\]
Most of these seemingly provocative utterances are in fact tempered by the use of smileys, rendering the challenges playful and light-hearted in tone. Indeed, stylistically, Tayyib ‘good’ co-occurred in 74% of lines with smileys, underscoring its connection with informality and humour. Again, a vast range of topics was apparent here, but religion was not one of them.

The remaining highly frequent words were yawm ‘day’ and the semantically-related al-yawm ‘the day’. The main differences between these two words consist in their relative degrees of definiteness as nouns, in their levels of formality (see Footnote 115 above), and in their adverbial functions. In this last regard, al-yawm ‘the day’ often functions as a temporal adverb meaning ‘today’ whereas yawm ‘day’ does not. Indeed, within its concordance lines al-yawm ‘the day’ was found to mean ‘today’ in no less than 58 instances. Another difference is that yawm ‘day’ frequently occurs in the Arabic construct structure in order to generate the days of the week such as Monday, Tuesday, etc. (see Footnote 86 above).

Now, regarding their respective contributions to an understanding of the uses and functions of Code 1 (Arabic-scripted Arabic), the word yawm ‘day’ co-occurred with adverbs of time e.g. ams ‘yesterday’ and ghada ‘tomorrow’ to reinforce the temporal proximity of an event. Topically, yawm ‘day’ whether an adverb or noun, was found to occur with several themes: especially Islam with 25% of all lines (e.g. the 'Day of Resurrection', the 'Day of Reckoning', the

---

104 Grammatically as mentioned above, yawm ‘day’ occurs frequently as the first element in the Arabic construct formation meaning ‘day of…’ such as ‘day of judgement’, ‘day of wedding’ i.e. wedding day, etc. In contrast, al-yawm ‘the day’ can only occur as the second element in such a construct phrase e.g. ‘joke of the day’.
'Day of Ashura', etc.), but also romantic relationships, narratives, politics and political news, jokes, and food. Stylistically, although 60 concordance lines of *yawn* يوم ‘day’ were Vernacular Arabic, only 19 lines were found to contain smileys. Again, a relatively low count of smileys might be expected when Islamic topics were so prevalent here. Non-religious uses of *yawn* يوم ‘day’ were generally references to days of the week or in expressions such as 'every day', 'wedding day', 'the next day', etc. In such cases, the utterances are typically part of Vernacular-style personal narratives. References to specific days involved news pieces usually to Palestine and/or Israel.

Similar to *yawn* يوم ‘day’ discussed above, *al-yawn* اليوم ‘the day’ was found to be connected to a variety of different topics: politics involving Islam, Jordan, Palestine, or Iraq, Islam in a religious sense as opposed to a political sense, food and eating, football, jokes, work, and male-female relationships and love. This mixture of topics revealed a division in style between more serious topics such as religion and politics written in informational formal Modern Standard Arabic and lighter topics such as food, sports, and romance written in involved Vernacular Arabic, exemplified here:

> بعد سنة واحدة بس بالغيره اننا اليوم كتب من عمان ياعا,
> After (having spent) only a year and a half in the West, today I am writing from Amman...hey...

---

105 In fact in Arabic, the use of *yawn* يوم ‘day’ as part of a noun phrase to indicate a day of the week such as Thursday i.e. *yawn al-khamis* يوم الخميس (lit. ‘day the fifth’) is stylistically-informal when compared to the more standard form *al-yawn al-khamis* اليوم الخميس (lit. ‘the day the fifth’) in which *yawn* carries the definite article *al* in order to agree grammatically with its post-positioned modifier adjective *al-khamis* (see *al-yawn* اليوم ‘the day’ below).
In contrast to *yawm* يوم ‘day’, *al-yawm* اليوم ‘the day’ appeared in the numerous personalized narratives. A smaller subset of utterances mentioning ‘today’ were part of news reports and were more non-involved than involved, written in Modern Standard Arabic instead of Vernacular Arabic. In this connection, several other lines containing *al-yawm* اليوم ‘the day’ were linked to news reports and more formal discourse especially when referring to specific dates or days of the week.

In summary, Code 1 (Arabic-scripted Arabic) is not entirely uniform in terms of style, topical focus, or even form. Stylistically, it featured both involved and non-involved utterances written typically in Vernacular and Classical/Modern Standard Arabic respectively. This highly salient finding alone shatters the dictum that Written Arabic is, or should always be, Standard Arabic. Highlighting the dichotomous nature of Code 1 in the corpus is that fact that most of Code 1’s religion-related frequent lexical words tend to be used in formal utterances or in a mixture of formal utterances and informal utterances, as illustrated by items such as *qāla* قال ‘he said’, *al-nās* الناس ‘the people’, *al-yawm* اليوم ‘the day’, and *yawm* يوم ‘day’ while the items *wallāh* وَالله ‘by Allah/God’, *baddī* بدي ‘I want’, and *Tayyib* طيب ‘good’ provided clear evidence of vernacular use, tempering the notion that Arabic-scripted Arabic is used exclusively for serious topics such as Islam. Having said this, the theme of religion, especially Islam, was encountered very frequently throughout most Code 1 concordances. In some cases, whole concordances were discovered to be connected to the Prophet Mohammed (e.g. *Salla* صلى ‘may (God) bless (him)’ and *wa-sallam* وسلم ‘and may (God) save (him)’). In others such as *allāh* اللّه ‘Allah/God, a majority of lines (i.e. 66%) were religion-related. In these cases, virtually all utterances were written in a highly formalized style of Modern Standard Arabic appropriate to the weighty topic
of religious discourse, consisting of a mixture of Quranic (and sometimes Biblical) quotations, narrations of Hadith, spiritual advice, or non-involved content about Islam. Even seemingly lay words such as qāla قال ‘he said’, al-nās الناس ‘the people’ and yawm يوم ‘day’ were frequently connected to Islam, their religion-related concordance lines totalling 64%, 35%, and 25% respectively. Although these observations seem to concur with findings presented in Chapter 5, linking Code 1 to the Religion Forum, the vernacular and smiley-containing utterances connected to a variety of secular topics ranging from food to songs and from romance to work and career. This topical diversity underscores the relative versatility of Code 1 as a means of discussing the very topics that are reflected in the titles of the various forums that make up the corpus.

At this point, concordance findings for Code 2 (BNC English) and Code 3 (Arithmographemic Latin-scripted Arabic with English) will be presented, exploring whether these codes are in fact discursively different or similar to Code 1 (Arabic-scripted Arabic) and to one another. This comparison begins with Code 2 in the next section.

6.3.2 Code 2 (BNC English)

This sub-corpus consists of items which match lexis from the BNC World Corpus and in this sense, can be considered lexically closest to samples of both written and spoken British English of the 1990s (Oxford, 2005). Tables 6.4 and 6.5 reveal that Code 2’s top 10 lexical words exhibit certain topical differences when compared to those of Code 1 (Arabic-scripted Arabic):
Table 6.4: Top 10 lexical words in the Code 2 sub-corpus

<table>
<thead>
<tr>
<th>Rank</th>
<th>Code 2 top 10 lexical words</th>
<th>Occurrence</th>
<th>%</th>
<th>Threads</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>KNOW</td>
<td>5292</td>
<td>0.3</td>
<td>2785</td>
<td>12.9</td>
</tr>
<tr>
<td>2</td>
<td>THINK</td>
<td>4297</td>
<td>0.3</td>
<td>2458</td>
<td>11.4</td>
</tr>
<tr>
<td>3</td>
<td>GOOD</td>
<td>3999</td>
<td>0.3</td>
<td>2368</td>
<td>11.0</td>
</tr>
<tr>
<td>4</td>
<td>PEOPLE</td>
<td>3596</td>
<td>0.2</td>
<td>1643</td>
<td>7.6</td>
</tr>
<tr>
<td>5</td>
<td>LOVE</td>
<td>3402</td>
<td>0.2</td>
<td>1485</td>
<td>6.9</td>
</tr>
<tr>
<td>6</td>
<td>TIME</td>
<td>3284</td>
<td>0.2</td>
<td>2019</td>
<td>9.3</td>
</tr>
<tr>
<td>7</td>
<td>SEE</td>
<td>2801</td>
<td>0.2</td>
<td>1706</td>
<td>7.9</td>
</tr>
<tr>
<td>8</td>
<td>GO</td>
<td>2746</td>
<td>0.2</td>
<td>1709</td>
<td>7.9</td>
</tr>
<tr>
<td>9</td>
<td>THANKS</td>
<td>2575</td>
<td>0.2</td>
<td>1936</td>
<td>8.9</td>
</tr>
<tr>
<td>10</td>
<td>WANT</td>
<td>2442</td>
<td>0.2</td>
<td>1450</td>
<td>6.7</td>
</tr>
</tbody>
</table>
Table 6.5: Code 2 Top 10 lexical words showing topical and stylistic features

<table>
<thead>
<tr>
<th>Rank</th>
<th>Code 2 Top 10 lexical words</th>
<th>Topics</th>
<th>Involved vs. Informational</th>
<th>Smileys</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>KNOW</td>
<td>General, relationships, songs/lyrics, politics, religion, food, clothing</td>
<td>Inv. 90%</td>
<td>25%</td>
</tr>
<tr>
<td>2</td>
<td>THINK</td>
<td>General, posting, songs/lyrics, local and social issues, relationships homosexuality</td>
<td>Inv. 98%</td>
<td>7%</td>
</tr>
<tr>
<td>3</td>
<td>GOOD</td>
<td>Varied: food, health, politics, fashion, religion, songs/lyrics, hobbies</td>
<td>Inv. 58%</td>
<td>32%</td>
</tr>
<tr>
<td>4</td>
<td>PEOPLE</td>
<td>Qualified groups e.g. Jordanians, Jews, Palestinians, etc.(69%), Generic references (31%), lack of jokes</td>
<td>Inf. 59%</td>
<td>8%</td>
</tr>
<tr>
<td>5</td>
<td>LOVE (noun or verb)</td>
<td>General: affection, romance, songs/lyrics, hobbies</td>
<td>Inv. 84%</td>
<td>19%</td>
</tr>
<tr>
<td>6</td>
<td>TIME</td>
<td>General: relationships, food, work, songs/lyrics</td>
<td>Inv. 62%</td>
<td>19%</td>
</tr>
<tr>
<td>7</td>
<td>SEE</td>
<td>General: songs/lyrics, local culture, posting, relationships, religion</td>
<td>Inv. 93%</td>
<td>17%</td>
</tr>
<tr>
<td>8</td>
<td>GO</td>
<td>General: songs/lyrics, health and fitness, internet, places</td>
<td>Inv. 91%</td>
<td>18%</td>
</tr>
<tr>
<td>9</td>
<td>THANKS</td>
<td>General: well-wishing, health/fitness</td>
<td>Inv. 100%</td>
<td>58%</td>
</tr>
<tr>
<td>10</td>
<td>WANT</td>
<td>General: songs/lyrics, posting, relationships, family/children, local culture, sex</td>
<td>Inv. 80%</td>
<td>13%</td>
</tr>
</tbody>
</table>
Tellingly, in contrast to Code 1 (Arabic-scripted Arabic), Code 2 (BNC English) does not feature any ostensibly religion-related lexis in this list. Rather, more general and personalized lexical items predominate such as verbs and/or nouns of feeling, sensing, and thinking e.g. KNOW, THINK, LOVE, SEE, and WANT. However, this initial impression merits closer scrutiny by examining the concordance lines of each item.

The most frequent Code 2 (BNC English) lexical verb was KNOW (the 100-line concordances for each top 10 lexical word in Code 2 are found in Appendix I). Its 100 concordance lines revealed a variety of topics especially relationships and English-language songs, including quoted lyrics. Political discussions were frequent as were references to the Middle East, Palestine, and Arab language and culture. In this last regard, it was also interesting to observe a question about the Christian Arabic version of the name ‘Jesus’:

16 s the arabic word for Jesus? i know muslim says Isa, but wha

Other observable topics were hacking and gaming, clothing and fashion, food, living in the UAE, and discussion of specific groups of people such as women including an apparent slur against ‘blonds’ meant to be taken humorously:

56 Ask one of the blonds you know to keep an eye on it stickingtongueout

A relatively large number of lines were difficult to decipher in terms of precise topic but were nonetheless illustrative of various communicative functions carried out by Code 2 (BNC English). In this group were utterances expressing personal opinions and assertions, admissions of not knowing certain things, and different kinds of questions including rhetorical ones.

106 In Arabic, there are two names for ‘Jesus’, يِسُوعَ which is used by Christians and عيسى used by Muslims. Thus, in L16 the poster is apparently inquiring about the Christian Arabic name for Jesus.
Stylistically, concordance lines for KNOW varied from formal English devoid of smileys and contracted words to highly informal English with Netspeak-type orthography (see Crystal, 2001) i.e. non-standard punctuation and spelling such as omission of the apostrophe in contractions, small-case where capitals are normally necessary, e.g. ‘i’ for the first person pronoun ‘I’, as well as other orthographic shortcuts such as ‘u’ for ‘you’, ‘ur’ for ‘your’, ‘ppl’ for ‘people’, and the acronym ‘lol’ for ‘laugh out loud’. Incidentally, 25 lines featured smileys. Also, 90% of its concordance lines were involved. L32 illustrates a number of these informal and personalized features:

```
32  lol..tough one! i dont know what i would do..contented
```

At this point, it will be interesting to consider evidence from the concordances of other Code 2 (BNC English) top 10 frequent words in order to determine whether the trend toward involved generalized discourse across a variety of topics is also observable among them.

The next Code 2 (BNC English) frequent item was THINK, which had much in common with KNOW. For instance, THINK revealed a variety of topics: comments pertaining to reading and writing posts and moderating the forums, computers, songs and lyrics, Arabs including Palestinians and Lebanese, clothing and attire, social and moral opinions, questions surrounding cases or situations of murder, violence, and drug dealing, discussion surrounding women and girls, children, and friendship, love, and relationships. One unexpected finding in this regard, was a single line stating the desirability of gay males as friends:

```
39  i wanna gay friend 2.. i think gay guys make great frie
```
This finding is particularly noteworthy in light of the officially hostile attitudes toward homosexuality in the Middle East region in general.

Stylistically, THINK concordance lines were overwhelmingly involved, exhibiting first person personal opinions and observations, statements of self-disclosure, and statements of advice and questions posed to second person addressees. Though Netspeak-type orthography was evident throughout, smileys were quite infrequent (only 7 lines total), implying that much of the discourse containing THINK is rather serious and sincere in tone as suggested by this candid self-reflection:

```
58 with such issues. That made me think of how pathetic I am, t
```

The third most frequent lexical word in Code 2 (BNC) was the adjective GOOD. As with the previously mentioned Code 2 frequent words, GOOD revealed a range of topics. In addition to those already mentioned for KNOW and THINK were football, health and fitness, hair care and hairstyling, tourism, the economy, work and careers, and pastimes such as photography, TV, films, and film-making. Interestingly, religion, Islam, the Qur’an, and Christianity were also relatively common here. But in some cases, as with the other Code 2 items, the precise topical references of certain lines was difficult to determine. Despite this occasional topical ambiguity, a large number of these general utterances were clearly stylistically involved, consisting of words of appreciation, praise, encouragement, or compliments directed toward addressees. Many of these praising utterances were also accompanied by smileys, augmenting their supportive and appreciative effect:

```
95 offersflower good words thanx
```

Other uses were in expressions of well-wishing and farewell e.g. ‘good luck’, ‘good night’, and ‘have a good one’ often accompanied by smileys (32% total), though some were critical
statements and questions where a negative was present. In brief, GOOD tended toward personalized and involved utterances in line with other frequent Code 2 (BNC English) items. The remaining frequent lexical words of Code 2 (BNC English) reflected similar characteristics to those mentioned thus far and with a similar range of topics: local culture, living overseas, songs/lyrics, food, health and fitness, sports, relationships, marriage, family and children, posting to the forums, hobbies, etc. More serious topics such as regional politics, gender and human rights (e.g. wearing hijab), and social issues (e.g. abortion, anti-gun laws, corruption, etc.) were also encountered periodically. Religion-related discussions were also apparent, notably Christian ones in addition to Islam-related ones, though not with any of the frequency witnessed for Code 1 (Arabic-scripted Arabic).

The frequent word PEOPLE was somewhat different from the others in the Code 2 (BNC English) list since it was informational as opposed to involved 59% of the time, featuring only 8% smileys. It was also mostly used to refer to specific ethnic groups such as Jordanians, Jews, Palestinians, etc. Otherwise, PEOPLE occurred in generic statements.

One of the top 10 Code 2 words, THANKS was inherently involved and phatic in usage, which explains why it was accompanied 58% of the time with smileys. Others, such as the verbs SEE, WANT, and LOVE were also overwhelmingly involved in style, but were used to express personal opinions and desires as well as questions. Unsurprisingly, LOVE when used as a noun, engendered discussions of marriage and romance, or else functioned phatically to express affection toward an addressee. GO, on the other hand, was mostly used to encourage an addressee to do something or else in personalized narratives. TIME was also used primarily in personalized narratives and in well-wishing.
Compared to Code 1 (Arabic-scripted Arabic), Code 2 (BNC English) featured a broader range of topics including taboo ones such as homosexuality or sex as seen in these two lines from the WANT concordance:

| 47 | sex and food ...do they ever really **want** anything else | 47 | sex and food ...do they ever really **want** anything else |
| 75 | in the west the guys who are into this stuff **want** just to have sex for pleasure | 75 | in the west the guys who are into this stuff **want** just to have sex for pleasure |

Briefly, Code 2 (BNC English) highly frequent lexis revealed several patterns. On the whole, involved utterances were more common than non-involved utterances throughout most of the concordances. Levels of formality varied among the utterances: some conformed closely to standard written English as found in offline contexts, while others exhibited Netspeak-type orthography and frequent use of smileys which functioned as personalizing markers of levity and humour. Utterance types ranged from opinions, assertions, simple and rhetorical questions, requests, self-disclosing comments to greetings. Several recurrent topics seemed to reflect the themes of the various forums discussed in Chapter 5 such as jokes, food and cooking, politics, local culture, work and study, relationships between males and females, family, religion, health and fitness, and well-wishing. While these also paralleled certain aspects of Code 1 (Arabic-scripted Arabic) topics, the diversity of topics in Code 2 seemed greater given far less concentration on religion. Also, while Code 1 exhibited numerous stylistically-informal utterances (see **wallāh** ‘by Allah/God’ and **baddī** ‘I want’), overall its use tended toward a relatively formal and non-involved style when compared to Code 2.

To complete the picture of contrastive topical focus and function among the three main codes of the corpus, the concordance lines of the frequent lexis of Code 3 (Arithmographemic Latin-scripted Arabic with English) will now be examined.
6.3.3 Code 3 (Arithmographemic Latin-scripted Arabic with English)

Code 3 is the most linguistically unconventional of the three codes by virtue of its mixed nature, featuring both English and Arabic lexis, the latter written in Latin script often with numerals. Its linguistic hybridity is observable in its top 10 frequent lexical items seen in Tables 6.6 and 6.7:

<table>
<thead>
<tr>
<th>Code 3 Top 10 Lexical Words</th>
<th>Occurrence</th>
<th>%</th>
<th>Threads</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALLAH</td>
<td>11094</td>
<td>0.27</td>
<td>4748</td>
<td>22.0</td>
</tr>
<tr>
<td>WALLAH³⁰⁷</td>
<td>3837</td>
<td>0.09</td>
<td>2399</td>
<td>11.1</td>
</tr>
<tr>
<td>WALLAH</td>
<td>107</td>
<td>0.09</td>
<td>2399</td>
<td>11.1</td>
</tr>
<tr>
<td>KNOW</td>
<td>6765</td>
<td>0.16</td>
<td>3207</td>
<td>14.8</td>
</tr>
<tr>
<td>THINK</td>
<td>4895</td>
<td>0.12</td>
<td>2439</td>
<td>11.3</td>
</tr>
<tr>
<td>LOVE</td>
<td>4615</td>
<td>0.11</td>
<td>1856</td>
<td>8.6</td>
</tr>
<tr>
<td>TIME</td>
<td>4442</td>
<td>0.11</td>
<td>2384</td>
<td>11.0</td>
</tr>
<tr>
<td>GOOD</td>
<td>4417</td>
<td>0.11</td>
<td>2408</td>
<td>11.1</td>
</tr>
<tr>
<td>WAY</td>
<td>3467</td>
<td>0.08</td>
<td>1999</td>
<td>9.2</td>
</tr>
</tbody>
</table>

³⁰⁷ WALLAH can also mean ‘and God’ according to context (see Footnote 90 above).
Table 6.7: Code 3 Top 10 lexical words showing topical and stylistic features

<table>
<thead>
<tr>
<th>Rank</th>
<th>Code 3Top 10 Lexical Words</th>
<th>Language; Vern. vs. English</th>
<th>Topics</th>
<th>Involved vs. Informational</th>
<th>Smileys</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>ALLAH</td>
<td>Vern. 100%</td>
<td>General: relationships, child-bearing, health/fitness, condolences, (no religion)</td>
<td>Inv. 100%</td>
<td>42%</td>
</tr>
<tr>
<td>2</td>
<td>KNOW</td>
<td>Mostly Eng.</td>
<td>General: marriage/marital status/relationships, fashion, family</td>
<td>Inv. 100%</td>
<td>21%</td>
</tr>
<tr>
<td>3</td>
<td>THINK</td>
<td>Mostly Eng.</td>
<td>General: Palestine/Israel, gender roles, Islamic practice, posting</td>
<td>Inv. 94%</td>
<td>17%</td>
</tr>
<tr>
<td>4</td>
<td>LOVE</td>
<td>Mostly Eng.</td>
<td>General: male/female romantic behaviour, romance/marriage, songs/lyrics, food, hobbies</td>
<td>Inv. 80%</td>
<td>26%</td>
</tr>
<tr>
<td>5</td>
<td>TIME</td>
<td>Mostly Eng.</td>
<td>General: posting, gender issues, Islam, Middle East, business/work,</td>
<td>Inv. 80%</td>
<td>20%</td>
</tr>
<tr>
<td>6</td>
<td>GOOD</td>
<td>Mostly Eng.</td>
<td>General: sports, food/cooking, posting, relationships</td>
<td>Inv. 87%</td>
<td>36%</td>
</tr>
<tr>
<td>7</td>
<td>WALLAH&lt;sup&gt;108&lt;/sup&gt;</td>
<td>Vern. 100%</td>
<td>General: posting, marriage, local culture (no religion)</td>
<td>Inv. 100%</td>
<td>62%</td>
</tr>
<tr>
<td>8</td>
<td>MAN</td>
<td>Vern. 50%</td>
<td>General: gender roles, relationships, jokes</td>
<td>Inv. 74%</td>
<td>41%</td>
</tr>
<tr>
<td>9</td>
<td>PEOPLE</td>
<td>Vern. 25%</td>
<td>General: posting to mahjoob, local/social issues, Islam</td>
<td>Inv. 72%</td>
<td>9%</td>
</tr>
<tr>
<td>10</td>
<td>WAY</td>
<td>Mostly Eng.</td>
<td>General: family/relationships, local content, Arabic songs, food/cooking</td>
<td>Inv. 74%</td>
<td>18%</td>
</tr>
</tbody>
</table>

<sup>108</sup> WALLAH can also mean ‘and God’ according to context (see Footnote 90).
Despite surface lexical commonalities between both Code 1 (Arabic-scripted Arabic) and Code 2 (BNC English), the frequent lexis of Code 3 (Arithmographemic Latin-scripted Arabic with English) had to be scrutinized for topical focus and stylistic functions to reveal to what extent Code 3 resembled (or differed from) the other two codes.

As in Code 1 (Arabic-scripted Arabic), ALLAH ‘Allah/God’ was the most frequent word in Code 3 (Arithmographemic Latin-scripted Arabic with English) (the 100-line concordances for each top 10 lexical word in Code 3 are found in Appendix J). However, in contrast to its Code 1 counterpart, ALLAH occurred in only four religion-related utterances out of the one hundred concordance lines examined, i.e. belief in God, becoming Muslim, Prophet Mohammed’s wife Aisha, and Islamic songs. Most of the remaining lines revealed functions such as well-wishing, congratulating, and offering blessings invoked on behalf of a first person singular or plural, a second person addressee, or a third party. Several of these also mentioned the addressee by name or contained terms of endearment such as 7abebii ‘love’. A much smaller number of lines reflected intentions via the Arabic expression of hope, IN SHA’ ALLAH ‘Allah/God willing’. Interestingly, three lines featured curses directed at others as here:

```
80       allah yokheth.hom wa7ad wa7ad
       May Allah/God take them away one by one
```

42 lines contained smileys, highlighting the personalized function of ALLAH in several cases. Linguistically, Vernacular Arabic, which was linked to personalized content in Code 1 (Arabic-scripted Arabic), characterized the majority of lines though a few lines exhibited Latin-scripted Modern Standard Arabic as in this stylistically formal utterance despite the smileys:

```
12    in happyfacesmallsmile jazzaki allah khayran huggingfriend
...in happyfacesmallsmile May Allah/God grant you a portion of goodness huggingfriend...
```
Lexically, throughout the concordance, content ranged from utterances featuring mainly arithmographemic Latin-scripted Arabic to those mainly composed in English. In terms of topics, frequent references were connected to relationships, marriage, families, and having babies as in this line about wishing for a baby boy:

19 | a pink or a blue? blue bi ezn Allah tab3an. How r ur prepar
... a pink (girl) or a blue (boy)? Blue Allah/God willing of course. How r ur prepar...

Others concerned health and illness, condolences, food, cars, and specific countries such as Canada, Jordan, and Kuwait.

Although occurring in seventh place, it is opportune to discuss the lexically and semantically similar term WALLAH ‘by Allah/God’ at this point. As with its Code 1 (Arabic-scripted Arabic) counterpart, WALLAH functioned mostly as an intensifier in Code 3. However, while a few lines of its Code 1 counterpart were found to mean ‘and Allah/God’, no such usage was detectable for WALLAH in Code 3. Stylistically, WALLAH was used almost exclusively in involved utterances while its Code 1 counterpart occurred in non-involved utterances roughly 33% of the time. Regarding smileys, compared to its Code 1 equivalent, WALLAH exhibited almost twice the number i.e. 62 of 100 lines, suggesting a comparatively more personalized and light-hearted use of WALLAH in Code 3. Further, virtually all lines contained Vernacular as opposed to Modern Standard Arabic, underscoring the informal connotation of WALLAH:

86 | wallah saba2teeni stickingtongueout
Hey, you beat (Fem. Sing.) me to it stickingtongueout

56 | o5te fa 7adret janabha will stay in amman till aug!!! wak wallah gaharatne
...my sister, so Her Royal Highness will stay in Amman till August!!! Anyway, she really used to boss me around

109 *Pink and Blue* is the name of a forum devoted entirely to expectant and new mothers (see Chapter 5).
In terms of topics and functions, a range was apparent: school subjects, food, mobile phones, money, sports, smoking, posting to mahjoob.com, jokes, shopping malls, summer vacation, cars, downloading CDs, references to the Middle East such as places, and people such as Jordanian girls and Saddam Hussein. Others lines concerned wearing hijab, friends, family, marriage including choosing a wife, romantic relationships, and relationship advice. Discursively, self-disclosure statements and personal narratives were very common as were well-wishing statements, exclamations, questions, opinions, and assertions. Briefly, WALLAH was similar to its Code 1 counterpart in terms of topics but had apparent no connection to the theme of religion.

The next set of Code 3 (Arithmographemic Latin-scripted Arabic with English) items discussed here are the stative verbs KNOW and THINK, both also found in the Code 2 (BNC English) top 10 list. Interestingly, in clear contrast to their Code 2 counterparts, both KNOW and THINK were frequently accompanied by Latin-scripted Arabic items such as discourse markers e.g. ‘ba3den’ (‘and then’), ‘5ala9’ (‘that’s enough’) or the Arabic subordinate conjunctions ‘inno’, ‘eno’, and ‘enno’ (‘that he/she/it is’) or ‘eny’ (‘that I am’). As ostensibly English-language items, perhaps it is not surprising that their respective concordance lines contained relatively little Latin-scripted Arabic compared to both ALLAH and WALLAH, which featured such items in almost each line of their concordances. Nevertheless, sporadic use of Vernacular Latin-scripted Arabic appeared to underscore text-producers’ attempts to forge a direct link to local, popular Arab culture. Some noteworthy Latin-scripted items present in the concordance lines were Arabic proper names e.g. ‘7attar’, cultural terms such as ‘a7maq’ (‘fool’), ‘ashkaljeyeh’ (‘trouble-maker’), ‘fay3a’ (‘hip’, ‘cool’), ‘9atyat’ (‘rude girls’) or very short phrases and exclamations like ‘(ma) 2dert 2adal sakta anymore’ (‘I couldn’t keep quiet anymore’).
Topically, like their Code 2 (BNC English) counterparts, KNOW and THINK exhibited a variety of themes: video clips, food, gender issues, relationships, single life, and marriage, health, (female) dress and clothing, children and family, friends, music and Arabic-language songs, Islam and Muslims, morality including terrorism, career/work, studying, politics including references to Arab-related places and politics especially Palestine and Israel, and forum posting. Stylistically, the vast majority of lines revealed involved style with ‘I’/‘i’ or ‘you’/‘u’ as the most frequent subjects. Both concordances exhibited a mix of formal and informal English, especially Netspeak features. Discursively, both KNOW and THINK were similar, featuring assertions, opinions, and self-disclosure statements, as well as various types of questions, though THINK also revealed several statements of intent. Regarding smileys, KNOW had 21 lines with smileys while THINK had only 17 lines, suggesting that more serious discussion often took place with these words as seen here:

| 93 | It is completely illogical to think that blowing yourself u |

In brief, KNOW and THINK behaved similar to their Code 2 (BNC English) counterparts with the exception that Vernacular Arabic elements occurred, typically highlighting Arabic cultural content such as names, expressions and exclamations.

LOVE was the fourth most frequent Code 3 (Arithmographemic Latin-scripted Arabic with English) item. It should be noted that as with Code 2, in 33 lines LOVE was found to function as a smiley (see Footnote 84 above). And in three more lines, LOVE was part of an author ID i.e. Happy Love. Consequently, as was done for its Code 2 (BNC English) counterpart, the 33 concordance lines containing the smiley LOVE were eliminated from the concordance and a randomized sample of 33 new concordance lines containing valid cases of LOVE was collected.
and appended to the original concordance in order to carry out a fuller analysis. And as with Code 2, the Code 3 top item LOVE featured several topics in common with its Code 2 counterpart as well as topics observed across other Code 2 and Code 3 frequent item topics: social commentary and critique, posting, personality types, well-wishing, a Qur’anic verse translated into English, references to music; both Arabic and English-language as seen here:

(Do you) have a song named ‘Everlasting Love’? blushingface

Unsurprisingly, as with its Code 2 (BNC English) counterpart, a large number of LOVE’s lines dealt with topics related to love: male and female romantic behaviour and across cultures, falling in love, relationships and advice, and marriage and proposals. Regarding discursive function, LOVE was often used phatically toward an addressee i.e. ‘love u’, ‘love ya’ with or without a term of endearment such as ‘sis’. Further, stylistically, 80 concordance lines were involved while 26 contained smileys, indicating an overall personalized style in the use of the word LOVE. Other utterances featured narratives, personalized questions, assertions, and especially positive evaluations of places such as Jordan, specific people e.g. ‘i love her outgoing personality’ or ‘I just love this guy huggingfriend’ and even local food such as the popular Middle Eastern vegetable stew, Molokhia:

I love molokhia (Jew’s Mallow) lovefilled

Notice this use of Latin-scripted Arabic content for local cultural references as seen with the other frequent Code 3 lexis. Again, Arabic discourse markers, expressions, and exclamations were also observed.
The next Code 3 (Arithmographemic Latin-scripted Arabic with English) item was TIME, also found in Code 2 (BNC English). In terms of topics, the same kinds of themes were discovered as with its Code 2 counterpart and elsewhere in the other concordances: forum members and their posts, sports like football and games, cooking and food, photography, gender issues and differences, and female rights, e.g. not wearing hijab, Islam, its teachings, religious leaders and followers, Middle East politics, rulers, and wars involving Palestine, Israel, Lebanon, and Afghanistan, playing songs such as English songs as well as Latin-scripted Arabic references to Arabic songs and singers, business, work, study, time management, vacations, friendship, relationships, marriage, motherhood and child rearing, and health and skin care.

Beyond specific references to Arabic proper nouns such as ‘3olama2’ (i.e. ulama علماء ‘Ulema’, Islamic religious scholars), in the TIME concordance lines, Latin-scripted Arabic items, while relatively infrequent, served similar functions as seen before: exclamations, untranslatable expressions, and discourse markers. Stylistically, 80% of TIME's lines were involved. However, only 20 lines contained smileys, suggesting that most utterances were more serious than frivolous. This was seen in several utterances featuring self-disclosure statements, serious questions, criticisms, personal narratives, warnings, assertions, and advice, using expressions containing TIME: ‘at the same time’, ‘any time’, ‘at this/that time’, ‘some time’, from ‘time to time’, ‘the first/last time’, ‘a long time ago’, and ‘it’s time to’.

The 6th most frequent Code 3 (Arithmographemic Latin-scripted Arabic with English) item GOOD, which also occurred in Code 2 (BNC English), resembled its Code 2 counterpart in several ways. First, GOOD in Code 3 had a similar number of lines containing smileys to its Code 2 counterpart (36 and 32 respectively). Next, both concordances featured a majority of
involved utterances. However, GOOD in Code 3 exhibited substantially more involved lines than Code 2, i.e. 87% vs. 58%. Regardless, both concordances featured evaluations of people (e.g. mahjoob posters) and things either positively or negatively, i.e. ‘not good’. Moreover, personalized greetings (e.g. ‘good morning/evening/night’, well-wishing (e.g. ‘good luck’), compliments (e.g. ‘good job/one), questions about quality (e.g. ‘is it good?’), and advice (e.g. ‘a good way’) were common in both Code 3 and Code 2 uses of GOOD. Code 3 topical similarities to the Code 2 concordance of GOOD were evidenced by references to sports like football, food and cooking, health and fitness, work, study, and careers, pastimes such as songs, art and photography, posting to, reading, and moderating the forums as well as discussing or addressing specific posters, and Middle East politics including anti-corporatism. Curiously, there were no obvious references to religion. Other common references involving GOOD in Code 3 were to love, marriage, relationships, parents, children and childrearing, clothing, cars, and shopping.

The next frequent Code 3 (Arithmographemic Latin-scripted Arabic with English) item was MAN, which had no counterpart item in the top 10 lists of the other two codes. Hand-checking of the concordance revealed that in 85 lines it was used to refer to males. The reminder of instances were either references to author IDs e.g. ‘K_man’, football clubs e.g. ‘man city’ for Manchester City, or Latin-scripted Classical Arabic where man means ‘who’ or ‘whoever’\textsuperscript{110}. Also, three more lines were examples of the Modern Standard Arabic relative pronoun man من ‘who’ that had been transcribed using Latin script. In each of these cases, quotations of Classical

\textsuperscript{110} Two more lines were excluded because they appeared to have been wrongly identified as Code 3 due to verse numbers being attached to the first word in each verse, creating pseudo-arithmographemic Latin-scripted Arabic items e.g. “8For man did not come from woman…. ….”.
In terms of discursive function, references to MAN were found in 46 lines to consist of vocatives and/or exclamations rather than as subjects or objects of verbs:

42 specially for zalmate offersflower Welcome Back, man huggingfriend Weenak
...specially for my man offersflower Welcome Back, man huggingfriend where’ve you been?

Here, notice the semantic redundancy of the Jordanian Vernacular Arabic ‘zalamate’ (i.e. ‘my man’) and the vocative use of the English ‘man’ later on. Such utterances underscore the use of MAN to express peer relationships between males. In this regard, there were no less than 11 occurrences of the awkward-sounding hybrid English cum Latin-scripted Arabic expression ‘ya man’ (‘hey, man’), which combines the Arabic vocative marker ‘ya’ meaning ‘hey’ or ‘yo’ with the English word ‘man’ exemplified here:

40 eyeswatering wallah ya man kolo tamam bs zae ma 2olt enta elsho'3ol fo2 rasi
eyeswatering really, man, everything is fine but as you said, work is over my head

In fact, involved utterances using MAN were evident in 74 lines out of the 85 lines where MAN occurs meaning ‘male’. Stylistically, Vernacular Arabic and Netspeak were very frequently encountered and mixed throughout the concordance in over 50 lines, further suggesting informal communication. Among these utterances, complimenting, greeting, inviting, well-wishing, and thanking were very common. Moreover, 35 of the 85 lines contained smileys as the above examples illustrated, indicating informality, playfulness, and affection. In this last connection, the public expression of affection and emotion between males, which is very acceptable in Arab culture, was frequently observed here as suggested by the huggingfriend and eyeswatering smileys. This is so despite pressure on males to project a virile heterosexual image toward others as seen here:
This last example appears to have been written by a female. Regardless, it underscores expectations for men to be macho on mahjoob.com (see references to ‘gay friends’ in Code 2 (BNC English) above).

Recurrent topics were marriage, divorce, and relationships including desirable qualities in a male partner, women’s rights vis-à-vis men, harassment, and male-dominated society. Topics common to other top 10 concordances were computers, food, TV and movies, money (e.g. “money can’t buy u…a decent man”), American politics, Middle East government and politics involving Jordan, Palestine, and Israel, Islamophobia and Anti-Shi’ism, and childrearing. Discursively, several lines were parts of narratives or jokes. The remaining utterances consisted of assertions, self-disclosure, and questions often expressing incredulity e.g. “Man get a grip, what the hell are you talking about?”.

PEOPLE was the next item in the Code 3 (Arithmographemic Latin-scripted Arabic with English) list, also found in Code 2 (BNC English). In contrast to MAN, PEOPLE featured fewer lines with Latin-scripted Arabic i.e. 29 out of 100. Apart from discourse markers, Latin-scripted Arabic here tended to consist of hard-to-translate expressions and proper names such as ‘majlis nowab’ (‘assembly of deputies’). Stylistically, while involved in 72% lines, smileys were found in only nine lines. Further, Netspeak was found in less than one third of the concordance. 40% of
lines showed PEOPLE in a general sense, indicating that generalizations were relatively common. Combined, these features suggest generally involved but serious discussion as was the case with Code 2 PEOPLE. This observation was confirmed by the relatively weighty topics frequently encountered: relationships and marriage, gender issues, warning and criticisms about posting to mahjoob.com, study and careers, appearance and dress, non-humorous narratives, politics, and economics, especially of Palestine and Jordan, social issues such as war, injustice, corruption, poverty, and unskilled social classes, and, related to these previous themes, Islam at the centre theological and social debate including references to jihad and de facto religious police (see Chapter 7) as seen here:

<table>
<thead>
<tr>
<th>55</th>
<th>does sharee3a allows people to become ameer by force too?</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Does Sharia (Islamic law) allow people to become rulers by force too?</td>
</tr>
</tbody>
</table>

| 93 | religious groups to run wild in the country and apply islam on poor people... |

In terms of discursive function, references to specific kinds of people were usually part of generalizing assertions about ‘other people’, ‘few people’, ‘some people’, ‘many people’, ‘lots of people’, ‘most people’, and ‘people you know’. More descriptive references were to ‘old people’, ‘Muslims’, ‘people in Jordan’, ‘Maan and Zarqa people’, and ‘our people’. Briefly, assertions and opinions followed by questions were the most typical types of utterances involving PEOPLE as was the case in Code 2 (BNC English).

The final Code 3 (Arithmographemic Latin-scripted Arabic with English) item was WAY. Typically, this word occurred in expressions describing a manner or method of doing or being e.g. ‘a timely and prompt way’, ‘the same way’, ‘is no way to treat…’, and ‘a sane way’. Other examples were as parts of discourse marker expressions such as ‘by the way’, ‘any way’, or as an
amplifier, e.g. ‘way better’ and ‘no way u can compare’. Occasionally, WAY was involved in prepositional phrases such as ‘Islam is the way of life’ and ‘your twisted way of thinking’.

Stylistically, 74% of utterances were involved, though smileys were found in only 18 lines. As with PEOPLE above, WAY appeared to be featured most often in serious topics: health and fitness, gender and equality, family issues, marriage, and relationships, and heated discussions about moderating and freedom of speech in posting:

<table>
<thead>
<tr>
<th></th>
<th>u are distroying this site by your <strong>way</strong>, and treat us as ur child and u are the fathers here,</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>follow up on your word and keep this thread. Freedom of speech is a <strong>two way</strong> road, after</td>
</tr>
</tbody>
</table>

Other serious topics were politics and social criticism, especially involving Arabs in general as well as Palestine, Israel, Lebanon, and political parties like Hamas, Fatah, Hezbollah, and the Muslim Brotherhood, and religion, especially Islam including this attack on Islam presumably by a Christian poster:

| 94 | **christians to adopt islam, but not the other way around? shu el islam msh le3beh?** |

... (ok for) Christians to adopt Islam, but not the other way around? What? Islam isn’t a game (and Christianity is)?

Less serious topics were also observable such as food and cooking, songs, especially Arabic ones featuring Latin-scripted Arabic singers and song titles, jokes, cars, and computers.

As for Latin-scripted Arabic, as seen in the rhetorical question above, Arabic language expressions were often employed in order to add emphasis to an assertion, a question, or a suggestion. Other observable utterance types were statements of self-disclosure, narratives, advice, descriptions, and compliments.

Overall, the topics associated with the frequent items in Code 3 (Arithmographemic Latin-scripted Arabic with English) were not very different from those in Code 2 (BNC English) or
even in Code 1 (Arabic-scripted Arabic), especially when Code 3 vernacular lexis is considered. Recurrent themes were songs, food, relationships and romance, participating in mahjoob.com, work and study, sports, computers, as well as more serious topics such as gender and social issues, Middle Eastern politics, and religion. Surprisingly, the Code 3 concordance lines were overwhelmingly English in composition, featuring relatively little Latin-scripted Arabic overall. And when present, Latin-scripted Arabic tended to be used to express ideas and cultural items that had no ready equivalents in English. In particular, as seen in the concordances of ALLAH and WALLAH, common expressions of well-wishing and blessing as well as exclamations were generally made in Latin-scripted Arabic and accompanied by smileys for enhanced emotional effect. Further, the presence of a relatively large number of smileys among its frequent words seems to underscore Code 3’s function for phatic communication and that it is relatively informal in content, containing several humorous and intimate elements.

Having now completed the analysis in this chapter, the final section provides a brief summary of the results.

6.4 Conclusions and summary

The following bulleted list provides a brief overview of the key findings of this chapter:

Overall similarities between Codes 1, 2, and 3:

- Considerable topical and lexical overlap between Codes 1, 2, and 3
- All three codes feature both formal and informal utterances including smileys
- A range of topics was found across all three codes especially local culture/issues
• Several semantic concepts are common to two or more of these codes: 'people' (Codes 1, 2, and 3), 'want' (Codes 1 and 2), 'know', 'think', 'love', 'time' (Codes 2 and 3), 'Allah' (Codes 1 and 3)

Key findings specific to each code:

A. Code 1 (Arabic-scripted Arabic):
   • Breaks conventional rule that Written Arabic should always be MSA or Classical Arabic
   • Reflects diglossia of offline contexts in the Arabic-speaking world:
     o Classical Arabic for Islam (e.g. quotations from Prophet Mohammed and Qur'an)
     o Vernacular forms for humorous content and/or involved phatic discourse
   • Relatively few smileys, suggesting more serious topics (in MSA)
   • Relatively few involved utterances, suggesting more informational content (in MSA)

B. Code 2 (BNC English):
   • Personalized lexis frequent: KNOW, THINK, LOVE, WANT, SEE, THANKS
   • More smileys and involved discourse than in Code 1, but less than in Code 3
   • Range of topics including taboo ones such as sex and sexual orientation, women's rights
   • Far less discussion of religion and Islam than in Code 1

C. Code 3 (Arithmographemic Latin-scripted Arabic with English):
   • Features more English lexis overall than Latin-scripted Arabic
   • Lexically and topically similar to Code 2, but contains more Arabic cultural references
   • Vernacular Arabic relatively sparse overall, but encountered frequently as conjunctions, discourse markers, expressions, exclamations, and proper Arabic nouns
   • Samples of Latin-scripted MSA are observable but rare
High smiley content compared to Codes 1 and 2, reflecting phatic function of Code 3

Like Code 2, features relatively little religious discussion

The macro-level topical tendencies of the three main codes that make up the bulk of the mahjoob.com corpus have been explored using random 100-line concordances based on the top 10 lexical words in Code 1 (Arabic-scripted Arabic), Code 2 (BNC English), and Code 3 (Arithmographemic Latin-scripted Arabic with English) respectively.

In brief, Code 1 (Arabic-scripted Arabic) exhibited the closest link to the topic of religion, especially Islam, as evidenced by several of its items and its numerous stylistically Classical and MSA utterances. Surprisingly, though, Code 1 also featured numerous vernacular forms, signalling a clear break from accepted practice when writing Arabic. Remarkably, Arabic diglossia between MSA/Classical Arabic as the high language and Vernacular Arabic as the low language seemed to be reproduced in CMC texts examined here. Indeed, Vernacular Arabic lexis was used for more mundane and frivolous topics, underscoring vernacular style as a common feature of humorous style while Classical/MSA lexis appeared primarily in religion-related lines. This seems to concur with Bentahila's (1983) findings based on spoken contexts about the functional and topical distribution of Classical Arabic and Vernacular Arabic in Morocco.

In contrast to Code 1 (Arabic-scripted Arabic), Code 2 (BNC English) featured a more diverse variety of topics ranging from hobbies to work and study, from computers to cooking, and from religion and politics to cars, with a range of styles from formal English grammar, spelling, and punctuation to informal Netspeak-style English. Interestingly, Code 2, in particular, revealed references to relatively taboo and sensitive topics such as homosexuality, sex, and women's
rights, perhaps indicating that such "Western" topics and issues were better expressed in a language other than Arabic.

Code 3 (Arithmographemic Latin-scripted Arabic with English) exhibited a similar range of style and was also topically closer to Code 2 with a diffuse range of topics overall. However, in contrast to Code 2, the frequent samples of Latin-scripted Arabic in Code 3 helped to draw a clear link between it and local Vernacular Arabic culture as typified by the frequently phatic use of Latin-scripted Arabic lexis such as ALLAH and WALLAH. As with Code 1, Code 3 vernacular use often betrayed humour and levity. That several items in Code 3 (Arithmographemic Latin-scripted Arabic with English) were identical to items from both Code 1 (Arabic-scripted Arabic) and Code 2 (BNC English) emphasized that it is a linguistically-mixed code (see McLellan, 2005; Smedley, 2006). Topically, the recurrent themes throughout all three codes mirrored to a large extent the topical content of the forums as discussed in Chapter 5.

As an aside, despite the clear risk of forum moderator censorship, the presence of utterances referring to generally sensitive and taboo topics within the Middle East such as jihad and political dissent, sex and dating, homosexuality, Christian-Muslim relations, and freedom of speech, underscores the fact that the mahjoob.com forums provide venues for participants to express viewpoints which they might not be able to do easily in offline public contexts. This also suggests that perhaps the diasporic membership may be encouraging or even driving the discussion of such topics.
The next chapter features a micro-level approach to the data, focusing on a subset of messages from a single thread featuring all of the main codes and several of the less frequently used codes. The purpose of this will be to demonstrate how mahjoob.com posters are able to perform identity work and index distinct identities via differentiated use of these codes in light of their stylistic differences as discussed in this chapter. This micro-level analysis of the strategic use of codes will establish that Mahjoobians too are well aware of the contrastive discursive and stylistic properties of these codes and are capable of exploiting these linguistic resources skilfully.
Chapter 7: Qualitative results and discussion

7.1 Overview
This chapter builds on the findings of Chapters 5 and 6 to illustrate how certain forum participants employ the codes on mahjoob.com strategically to perform identity work. Until now, the focus of the research has been on presenting macro-level quantitative analyses of the whole corpus, as well as examining the overall topic-related lexical patterns which distinguish the various codes in the corpus. These findings have been valuable in suggesting certain broad tendencies of code distribution and use on mahjoob.com. However, as Baker (2006) suggests, many of the most meaningful findings of a corpus approach are to be obtained by using more qualitative approaches after quantitative methods have yielded intriguing results or posed certain questions. In light of this, it is with the goal of investigating, at a deeper qualitative level, the macro-level quantitative patterns uncovered earlier in this study that the present chapter has been written. Specifically, this chapter addresses the final main research question of the study: 4) How do forum posters use the available script-code pairings to construct identities?

It must be noted at the outset that the findings of this chapter are not intended to be either conclusive or even fully generalizable beyond the immediate contexts in which they were obtained. Rather, it is hoped that the analysis of identity here will prove at the very least illustrative of the principle that strategic code-switching and orthographical choices play pivotal roles in identity-work among certain mahjoob.com forum participants.

Before presenting the results of the discourse analyses, and in order to contextualize the findings, a brief reference to identity-related code-switching and orthography theory covered in Chapters 1 and 2 is now provided. Once this has been discussed, the grounds for selecting the thread which
is the central focus of this chapter will be given. Finally, the analysis of the selected thread and authors will be presented, highlighting how different codes allow for the creation of distinct, oppositional personae by a select group of thread contributors.

7.2 Identity via code-switching, script-switching, and orthography

As mentioned in Chapter 1, identity is defined here as a social construct observable as distinct social personae created largely through strategic use of recognizable linguistically-realized styles (see Fairclough, 2003, p. 162), a notion akin to Och’s theory of indexicality (1987). In an online asynchronous written environment such as mahjoob.com these styles are enacted by forum authors through salient linguistic choices involving code, script, and use of paralinguistic features such as smileys in their written texts. Discourse analysis of these texts can then help delineate the range of locally-relevant stereotypes and styles created and projected by the authors in question (see K.M. Lee 2007; Smedley, 2006; Hinrichs, 2005; McLellan, 2005).

In this connection, several recent researchers have contended that written code-switching is more ideological in nature than spoken code-switching and underpinned by identity-construction motives. In parallel, others have argued that orthographical choices are neither autonomous nor neutral (Ferguson, 1959, p. 235). In conjunction with code-switching, therefore, the orthographic repertoires of the mahjoob.com participants are considered resources for identity-negotiation and construction (see Pavlenko and Blackledge, 2004, pp. 22-24). In this sense, script choice can also be considered a type of orthographical choice reflecting ideological motivations. That is, in addition to code-switching proper, different types of orthography, including script-switching, are used effectively to portray various styles. In fact, the decision to distinguish between two principal orthographically-distinct varieties of Latin-scripted Arabic in the corpus i.e. Code 3
(Arithmographemic Latin-scripted Arabic with English) and Code 10 (Non-arithmographemic Latin-scripted Arabic with English) was motivated by the possibility that such orthographic differences might betray identity-related processes (see Chapter 3). This chapter, in fact, furnishes evidence that text producers’ decisions to use either Code 3 or Code 10 seem to underscore identity-creation processes when their code use is contrasted within a given discussion thread.

7.3 The relevance of earlier findings in this study

To further contextualize the analysis presented in this chapter, it is worthwhile to recall briefly the earlier findings of this study. The results of Chapter 5 indicated that there were clear differences in the distribution and occurrence of the various codes in the corpus. Most notably, Code 1 (Arabic-scripted Arabic) was found to dominate in Humour-related, Poetry-related, and Local Culture-related forums, while Code 2 (BNC English) was found to predominate in Work/Study-related forums. In contrast, Code 3 (Arithmographemic Latin-scripted Arabic with English) was found to occur mostly in forums relating to personalized, non-professional, and familiar topics such as cooking, youth issues, and Jordan (e.g. *Kuluna Al Ordun* ‘We are all Jordan’ forum), and gender-related discussions. These findings were corroborated in Chapter 6, where it was shown that these codes also exhibit distinct tendencies in terms of the topics articulated through them as well as the levels of formality indexed by their use. For instance, further analyses of Code 1 lexis indicated that much of the Arabic-scripted Arabic content featured in the Religion and Palestine forums was likely to have been copied from other web sources where Arabic-scripted Arabic was found. Several Code 2 texts also seemed to have been imported from other sources. In contrast, Code 3 texts appeared to show the least amount of imported content. Further, in terms of informality, it was suggested that following messages in
Code 3 contained the highest number of smileys compared to both Code 1 and Code 2. Having gained this overall sense of how the linguistic codes in the corpus differ, it becomes easier to address the above research question since these codes can now be viewed as distinct linguistic resources available to authors in order to create different personae (see Lee, 2007; Su, 2003). To accomplish this, Fairclough’s concept of discursively-constructed styles will be relied upon (2003, pp. 159-163), though, admittedly, the approach here is heuristic in nature (see Hinrichs, 2005).

7.4 Structure of the analysis
In this chapter, first, an overview of the major topical theme of the selected thread is traced. Next, the decision to focus on a subset of messages within the selected thread is discussed. Once this has been done, a contrastive analysis is presented of the language use of four major thread contributors, who represent opposing positions in the debate that dominates the selected subset of thread messages. The main contributors’ divergent use of codes illustrates how code and orthography can help posters to project different and oppositional personae. But before presenting the results of this analysis, a few words are in order regarding the selection process that led to the detailed investigation of a particular discussion thread and its main contributors.

7.5 Selection of the thread for analysis
With a corpus of some 37 million words, it is virtually impossible to trace all the salient identity-related styles and personae featured in the data. Therefore, in the interests of both space and manageability, I decided to analyse a single thread, which satisfied three important criteria: 1) linguistic diversity, 2) length, and 3) unconventionality. Linguistic diversity was an important selection factor because I wanted to be able to contrast the use of different linguistic codes in interaction. Length was deemed important because a longer thread would allow for an
opportunity to see if linguistic patterns are sustained over of a broader series of interactions. Unconventionality was deemed desirable because of the likelihood that a linguistically atypical thread might reveal the strategic use of codes to achieve identity-related goals. In other words, if topic largely correlates with code choice, then other motives such as a desire to project a specific identity might help to account for when a poster deviates from expected code use patterns.

In order to identify a thread for follow-up analysis, the SPSS version of the corpus was examined to locate atypical threads in terms of code use within following messages. First, Figure 7.1 (this is identical to Figure 5.7 above) was examined to identify the most common patterns of code use in longer threads. Surveying this figure, it is becomes clear that the codes most commonly used for following messages in longer threads are Code 1 (Arabic-scripted Arabic) at 37.5% followed by Code 3 (Arithmographemic Latin-scripted Arabic with English) at 34%, and Code 2 (BNC English) at 14.3%:
A second step involved generating a table of all longer threads to locate any which were somehow atypical in terms of code use. As a result, it was determined that there was only one longer thread in the entire corpus that did not have a majority of its following messages composed in Code 1 (Arabic-scripted Arabic). This linguistically atypical discussion thread was Thread 206940 in the *Religion* forum. The most frequently used code in Thread 206940 was in fact Code 10 (Non-arithmographemic Latin-scripted Arabic with English), a relatively rare code, which accounts for a mere 4.1% of all following messages in the entire corpus (see Chapter 5). Yet within Thread 206940, Code 10 messages accounted for 106 messages i.e. 33% of all the messages in the entire thread. Clearly, then, this thread was somehow different from other longer threads. Thread 206940 was also found to be quite heterogeneous linguistically: among its 317 text-based messages it featured messages not only in Code 10 but also in Codes 1, 2, 3, 4, 6, 8, 12, 13, and 14, signifying that it featured ten of the 14 linguistic codes found in the corpus.
Further, this thread was unconventional in another sense: it contained a relatively higher amount of Latin-scripted Arabic and Latin-scripted English than one might normally expect in a thread in the *Religion* forum, which is noted for its extensive use of Code 1 (Arabic-scripted Arabic) (see Chapters 5 and 6).

In terms of analysis, since the focus here was to locate evidence for identity construction, it was decided to concentrate on the most prolific contributors to the thread\(^\text{111}\), who, incidentally, also appeared to exhibit salient code use behaviours when viewed in juxtaposition. In other words, I was interested in examining how certain posters positioned themselves with regard to other posters in the thread through selective use of code. To delineate such behaviour, specific lexical, grammatical, and orthographic features in the posted messages of main authors were contrasted to see if these features helped the posters to index any broader styles associated with the codes. Main authors were identified as those who had posted more than three times to the discussion thread.

To gain a broad sense of its discussion and topic, the entire thread containing 322 messages was copied from the website into a Microsoft Word document and printed. At this point, the thread was read and re-read several times. In this way, the first 75 messages were found to revolve around a specific delimited discussion topic: whether or not it is permissible for Muslims to criticize the Muslim rulers of Saudi Arabia in light of apparent un-Islamic actions, which these rulers are alleged to have committed. In the interests of space within the thesis, this subset of 75

\(^{111}\) Although the focus in this chapter is on main authors, reference to other authors will be made occasionally where their own code use seems to reflect code use patterns similar to those of the main authors.
messages, containing a total of 14,046 words, was selected for further detailed discursive analysis (see Appendix G), exploring the interactions between frequently-posting authors within these messages in terms of what they revealed about identity-creation processes.

Procedurally, the analysis entailed reading the 75 messages several times and determining the number of authors and frequencies of their messages within these messages as well as their code use patterns. These frequencies were tabulated using WordSmith 5.0 (see Table 7.1 below). Such data helped to identify the authors that contributed most often to Messages 1 to 75, who then became the objects of more detailed analysis. In the next section, the thread discussion occurring between Messages 1 and 75 is described, followed by a presentation of the main authors and their acts of identity via strategic code use.

7.6 Messages 1-75 in Thread 206940, “masha2a allah, masha2a allah”

The entire discussion thread contains a total of 322 messages, 317 of which are textual messages, making it one of the longest threads in the entire corpus. Like the corpus itself, it features texts written in Code 1 (Arabic-scripted Arabic), Code 2 (BNC English), and Code 3 (Arithmographemic Latin-scripted Arabic with English) as well as several other codes such as

---

112 Regarding the name of the thread itself, “masha2a allah masha2a allah”, this is written in Code 3 (Arithmographemic Latin-scripted Arabic with English) and translates literally as, “what God (has) willed, what God (has) willed”. This phrase is a very common formulaic expression used to invoke God’s continued blessings for favourable circumstances and to ward off the ‘evil eye’, i.e. to protect the speaker or someone dear to them from being afflicted by a curse or experiencing any misfortune. However, the use of ‘masha2a allah’ in this context appears to be ironic because the thread initiator, Kharoof Tayeh proceeds to describe a situation that is anything but favourable in his point of view i.e. the existence of corrupt governmental authorities in Saudi Arabia.
As mentioned above, the overall topic of the thread segment from Messages 1 to 75 is whether or not it is legitimate in Islam to criticize Muslim rulers, especially the current rulers of Saudi Arabia. To put this into perspective, consider that the Kingdom of Saudi Arabia is generally considered to be a modern-day example of an absolute monarchy (see Moaddel, 2007; Walden Publishing Ltd., 2006). Essentially, power resides primarily with the Saudi royal family though in recent years small steps have been taken toward democratization (see Kechichian, 2001; Sakr, 2007). These steps have included allowing for the direct election of municipal representatives by popular vote (Sakr, 2007).

Constitutionally, Saudi Arabia is an Islamic monarchy and the basis for its legal system is Shari’ah law, a system that derives its ultimate legitimacy from the Qur’an (see Kechichian, 2001, pp. 1-2; Vogel, 2000, p. 170), and from the Sunnah, the model for living righteously based on the Prophet Mohammed’s example as outlined in the Hadith, the body of traditions or sayings attributed to the Prophet Mohammed (Vogel, 2000, p. 4). The method for applying Islamic teachings to legal circumstances i.e. jurisprudence is known as fiqh (lit. ‘understanding’) and is derived from past and present pronouncements of Islamic jurists (Vogel, 2000, p. 5). In theory, the role of the Ulamaa\textsuperscript{113} (lit. ‘learned ones’), the body of Islamic scholars cum clergy, is very

\textsuperscript{113} Within Thread 206940, it is worth noting that Ulamaa is spelled several different ways with or without capitalization: In Code 10 (Non-arithmographemetic Latin-scripted Arabic with English) texts, it is written either as
important in an Islamic state such as Saudi Arabia because these religious experts set the standards for determining which actions and behaviours are *halal* (‘lawful’) or *haram* (‘unlawful’). While Muslim rulers have the last say in governance, they should defer to the spiritual wisdom and insight of the Ulaama for guidance whilst governing a Muslim state (Vogel, 2000, p. 173).

Beyond the Ulaama, groups of concerned citizenry often organize themselves into quasi-official organizations with the express intent of upholding the Islamic character and dignity of society in Muslim countries such as Saudi Arabia and Indonesia (see An-Na`im, 2008, p. 256) and in locales with significant Muslim communities such as northern Nigeria. In Saudi Arabia, this group is known officially under the name هيئة الأمر بالمعروف والنهي عن المنكر *hay'at al-'amr bil-ma'rūf wa nahi 'an al-munkar* (lit. “The Organization of the Cause for Propriety and the Forbidding of Sins”) (Lacroix, 2004), however, it is popularly referred to simply as الهيئة *al-hay'ah* (lit. “the organization”). In practical terms, groups like the Hay’ah take it upon themselves to encourage, and even enforce, compliance with Shari’ah law in the lives of individual citizens, using harassment and violence if necessary (see An-Na`im, 2008, p. 256;

---

114 This lengthy name derives directly from the following Qur’anic verse as cited in Cook (2001, p. 13): “‘Let there be one community of you (wa-l-takun minkum ummatan), calling to good, and commanding right and forbidding wrong (wa-yamurūna bil-ma'rūfi wa-yanhawna ‘ani l-munkar); those are the prosperers’ (Q3:104)”. Cook points out that there are seven other similar verses throughout the rest of the Qur’an. These verses provide their supporters with the religious justification for the creation of such community-based morality policing forces.
Baker, 2007, p. 126). Baker notes that these groups often enjoy at least tacit support and approval from religious authorities and certain politicians (2007, p. 126).

To put the above into context, consider that, in Thread 206940, the key issue at hand is whether the Hay’ah can and should advise members of the Saudi government and royal family about their own Islamic practice. Anti-establishment posters contend that Hay’ah members focus their efforts on correcting the behaviour of ordinary citizens while remaining silent about gross violations of Islamic law committed by both individuals from, and the collective of, the Saudi royal family and ruling establishment. The debate that ensues highlights some very intriguing patterns of language use especially as a means of projecting either an Anti- or a Pro-establishment stance. The linguistically salient exchanges between the main posters on these two opposing sides and how such exchanges help their authors to project distinct personae form the crux of the analysis in this chapter, presented over the next few sections.

7.6.1 Anti-establishment authors vs. Pro-establishment authors
As mentioned above, the main objects of debate in the thread are the Saudi royal family, the Ulamaa (i.e. Islamic scholars), and the Hay’ah of Saudi Arabia, and their respective moral integrity and legitimacy. In terms of stance, Messages 1-75 can be categorized into two broad, opposing themes: 1) Anti-establishment messages that criticize the Saudi rulers for un-Islamic behaviour and the Saudi religious authorities (both Ulamaa and Hay’ah) for not speaking out against these same rulers and 2) Pro-establishment messages that refute the Anti-establishment

---

115 A search for the Arabic language equivalent of “hay’ah” on Youtube yields several videoclips purporting to show groups of Hay’ah members harassing women and men in public places.

116 A decision was taken to conclude the analysis at message 75 since a new topic starts at message 76, where a critique of the secular nature of the Tunisian government is launched by a poster who is new to the thread.
arguments by emphasizing the religious duty of Muslims to obey their Muslim rulers regardless of their behaviour, citing counter-arguments to support the legitimacy and Islamic integrity of both the Saudi rulers and the Hay’ah and Ulamaa. Among these two opposing stances, three argumentation strategies are evident: 1) citing Islamic traditions and Muslim scholars to support their viewpoints, 2) accusing the opposing side of misinterpreting Islam, and 3) *ad hominem* attacks on the morality and/or scholarly credibility of opposing posters.

Within the first 75 messages of the thread, there were seven contributors who appear to all be male based on their user IDs as well as their references to themselves and others as “brothers”. Table 7.1 below presents each of these authors along with a tabulation of the number and code content of their respective. For comparison, Table 7.2 presents the overall code usage patterns of these same posters across the entire corpus. Included in these tables is a column indicating each author’s stance as either “Anti” (Anti-establishment) or “Pro” (Pro-establishment):
Before delving into the analysis, a couple of patterns regarding author stance and code use are worth noting. Looking at Table 7.1, it is interesting to observe that the three Pro-establishment authors, Snipe_aac, Muslim4, and SamiulIslam completely avoid Code 3 (Arithmographemic Latin-scripted Arabic with English), but have strong preferences for either Code 1 (Arabic-scripted Arabic) e.g. Muslim4 with 10 out of 11 of his posts written in that code, or Code 10 (Non-arithmographemic Latin-scripted Arabic with English). Regarding Code 10, SamiulIslam composed all three of his posts to Thread 206940 in it while Snipe_aac composed 16 out of 31 (i.e. over 50%) of his messages in this code.

Table 7.1: Code use among the authors within the first 75 messages of Thread 206940

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Kharoof Tayeh</td>
<td>Anti</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>7</td>
</tr>
<tr>
<td>Guillotine</td>
<td>Anti</td>
<td>10</td>
<td>4</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>19</td>
</tr>
<tr>
<td>3affash</td>
<td>Anti</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Muslim4</td>
<td>Pro</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>11</td>
</tr>
<tr>
<td>Snipe_aac</td>
<td>Pro</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>16</td>
<td>0</td>
<td>3</td>
<td>2</td>
<td>5</td>
<td>0</td>
<td>31</td>
</tr>
<tr>
<td>argumentive36</td>
<td>Anti</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>SamiulIslam</td>
<td>Pro</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 7.2: Code use among Thread 206940 authors across the entire corpus

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Kharoof Tayeh</td>
<td>Anti</td>
<td>0</td>
<td>0</td>
<td>83</td>
<td>443</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>42</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>48</td>
</tr>
<tr>
<td>Guillotine</td>
<td>Anti</td>
<td>1676</td>
<td>931</td>
<td>278</td>
<td>87</td>
<td>9</td>
<td>80</td>
<td>4</td>
<td>28</td>
<td>2</td>
<td>40</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>306</td>
<td>126</td>
<td>3576</td>
</tr>
<tr>
<td>3affash</td>
<td>Anti</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>13</td>
</tr>
<tr>
<td>Muslim4</td>
<td>Pro</td>
<td>2288</td>
<td>19</td>
<td>51</td>
<td>4</td>
<td>2</td>
<td>47</td>
<td>7</td>
<td>3</td>
<td>0</td>
<td>19</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>8</td>
<td>113</td>
<td>2564</td>
</tr>
<tr>
<td>Snipe_aac</td>
<td>Pro</td>
<td>47</td>
<td>35</td>
<td>3</td>
<td>4</td>
<td>0</td>
<td>11</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>316</td>
<td>5</td>
<td>31</td>
<td>18</td>
<td>139</td>
<td>4</td>
<td>604</td>
</tr>
<tr>
<td>argumentive36</td>
<td>Anti</td>
<td>5</td>
<td>948</td>
<td>324</td>
<td>13</td>
<td>6</td>
<td>6</td>
<td>1</td>
<td>15</td>
<td>13</td>
<td>60</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>1013</td>
<td>17</td>
<td>2424</td>
</tr>
<tr>
<td>SamiulIslam</td>
<td>Pro</td>
<td>0</td>
<td>15</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>42</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>23</td>
<td>0</td>
<td>80</td>
</tr>
</tbody>
</table>

| Code Totals |       | 4017   | 2033   | 1103   | 108    | 17     | 144    | 13     | 47     | 15     | 521    | 9      | 36     | 14     | 1541   | 266   | 9884  |

Before delving into the analysis, a couple of patterns regarding author stance and code use are worth noting. Looking at Table 7.1, it is interesting to observe that the three Pro-establishment authors, Snipe_aac, Muslim4, and SamiulIslam completely avoid Code 3 (Arithmographemic Latin-scripted Arabic with English), but have strong preferences for either Code 1 (Arabic-scripted Arabic) e.g. Muslim4 with 10 out of 11 of his posts written in that code, or Code 10 (Non-arithmographemic Latin-scripted Arabic with English). Regarding Code 10, SamiulIslam composed all three of his posts to Thread 206940 in it while Snipe_aac composed 16 out of 31 (i.e. over 50%) of his messages in this code.

In Table 7.2, main authors i.e. those who post more than three messages are indicated in bold. Specifically, these are Kharoof Tayeh, Guillotine, Muslim4, and Snipe_aac. Also, note that ‘C 1’ refers to Code 1 and so on.
In contrast, the four Anti-establishment authors: Kharoof Tayeh, Guillotine, 3affash, and argumentive36 are notable for their preference for Code 3 (Arithmographemic Latin-scripted Arabic with English): Kharoof Tayeh uses it for all seven of his posts, argumentive36 uses it for 2 out of 3 of his posts, and 3affash uses it for his single post to the discussion. Of all seven posters here, Guillotine, from the Anti-establishment side, is clearly the most multilingual in this thread: He writes over 50% of his 19 posts in Code 1 (Arabic-scripted Arabic), followed by over 20% in Code 2 (BNC English), then 15% in Code 3, and one post each in Code 14 (Non-BNC English) and Code 15 (Graphic content-only). In general, though, Anti-establishment authors have a strong preference for Code 3 while Pro-establishment authors avoid it entirely, tending to use Code 10 (Non-arithmographemic Latin-scripted Arabic with English) or Code 1 instead.

When Table 7.1 is compared to Table 7.2 (see Muslim4), several differences in code use pattern become apparent. Most noticeably, authors that tend to use only one code exclusively such as Kharoof Tayeh (Code 3) and Muslim4 (Code 1) are found to use other codes in other threads. These differences in code use patterns among specific authors will be explored later in this chapter.

Now that these overall code use trends within Thread 206940 have been discussed, it will be useful to consider specific cases of code use here with the goal of establishing how such use supports the creation of distinct identities among the thread’s contributors. In this regard, the analysis will now focus on the varying code use of the main contributors to the first 75 messages: Kharoof Tayeh, Snipe_aac, Muslim4, and Guillotine.
To provide a broader picture of how contrasting code use serves to create separate, oppositional personae here, first, the confrontational exchanges between Kharoof Tayeh from the Anti-establishment camp and Snipe_aac from the Pro-establishment camp will be compared linguistically. Then, the posts of Muslim4 from the Pro-establishment camp will be explored, followed by those of Guillotine from the Anti-establishment camp. As a matter of course, exchanges between Kharoof Tayeh and Muslim4 or between Snipe_aac and Guillotine will also be considered where necessary. Additionally, the four main authors’ code use patterns outside of Thread 206940 (see Tables 7.2 and 7.3) will be touched upon where such patterns shed further light on the identity-creation processes at work in the thread. Finally, reference will occasionally be made to the posts of the other authors in the thread: 3affash, argumentive36, and SamiulIslam when this is pertinent to the analysis of identity. Having established the structure of the analysis, the next section will begin with an investigation of Kharoof Tayeh’s postings, which are then contrasted with those of Snipe_aac, highlighting how each poster’s differential code use and orthography contribute to their projection of distinct and oppositional personae.

7.6.2 Anti-establishment author: Kharoof Tayeh

Kharoof Tayeh starts the discussion thread by citing an article from an online newspaper in Arabic that describes how large sums of money have been given to Western nations post-9/11.

The original article cited in Kharoof Tayeh’s post appears to have come from the website of Al-aswaq al-arabiyah, (lit. “the Arab Markets), a subsidiary website of the Al-Arabiya news website owned and operated by the Al-Arabiya News Channel (http://www.alarabiya.net/english/). In brief, the article claims that in the wake of 9/11 GCC countries contributed 300 billion dollars to the US, 100 billion dollars to Europe, and 60 billion dollars to Asian countries. In his message, Kharoof Tayeh seeks to underscore the irony of this contribution of “gulf money”...
He follows this imported text with a commentary composed in Code 3 (Latin-scripted Arabic with English):

Message 1:

1  wbéejo bégoooláy énno elhai2a bıtenteqêd élmashayêkê.  
   And they come and tell me that the organization criticizes the Sheikhs?

2  6âyêb...does anyone know the phone number for the hai2a?  
   Fine then organization

3  I'd like to inform them that 300 billions of gulf money is helping the economy of 
   "their enemy" as they claim. 😅

Here Kharoof Tayeh, whose user ID suggestively means “sheep who has gone astray”, starts his commentary in Latin-scripted Arabic in what amounts to an emphatic statement expressing incredulity at the assertion that the hai2a (religious authorities) actually reprimand the Sheikhs i.e. the rulers of Saudi Arabia (and perhaps other Arabian Gulf Cooperation Community (GCC) rulers). In Line 2, Kharoof Tayeh continues in Latin-scripted Arabic: he uses a common Arabic discourse marker, 6âyêb that approximates the English expression ‘fine then,’ conveying frustration or exasperation, before switching to Latin-scripted English for the remainder of his to Western countries which are widely regarded in the region as being “enemy” countries for their support of Israel against the Palestinians.

119 The convention used in this chapter for displaying messages written in part in Latin-scripted Arabic and part in English will be to translate these elements into English beneath the original word in the message. Message Line numbers are also added for ease of reference.

120 Here the use of the whistling smiley underscores Kharoof Tayeh’s position that the religious authorities feign ignorance when it comes to Gulf countries sending billions of dollars to anti-Islamic Western economies such as the US, the UK, and the rest of Europe who recognize and support Israel while waging war against Muslim countries such as Iraq and Afghanistan as alluded by Kharoof Tayeh in Message 1.
commentary. The only other Latin-scripted Arabic word used after this point is the word *hai2a* at the end of Line 2, which has no ready equivalent in English.

Kharoof Tayeh’s juxtaposition of linguistic codes here creates a contrast. The cited Arabic-scripted Arabic text provides the pretext for the discussion that follows. But as an imported text, Kharoof Tayeh is able to index a number of qualities through it. First, the article is written in Standard Arabic, which lends it an air of authority and formality. The credibility of the article is also bolstered by the fact that its original source is the Al Arabiya News Channel website, a top-rated news channel across the Arabic-speaking world (Feuilherade, 2003). Hence, the audience is presented with content that is to be taken as credible. Also, the fact that Kharoof Tayeh starts off the thread by quoting an article written in Arabic as opposed to English may be salient because it seems to frame the debate from a non-Western perspective. Indeed, starting off with a Western-leaning newspaper article written in English might have been considered less objective or less sympathetic to Islam by the other forum participants. Note also Kharoof Tayeh’s use of the phrase “their enemy” in scare quotes in referring to Western countries in his commentary. This highlights his position that Gulf countries are aligned with the West despite professions to the contrary. Below, further examples will be presented of Kharoof Tayeh’s frustration with his perception of Gulf Arab governments’ duplicitous relations with the West.

Immediately after citing the article written in Modern Standard Arabic, at Line 1 of the message above, Kharoof Tayeh starts his own commentary in Arithmographemic Latin-scripted Arabic (Jordanian Vernacular). This achieves a couple of objectives. First, it identifies Kharoof Tayeh as a Jordanian Arab. Second, by taking on features of informal language through use of Code 3 (Arithmographemic Latin-scripted Arabic with English) lexis, Kharoof Tayeh transforms the text
via a genre chain (see Fairclough, 2003, pp. 31-32, 66) into the language of the ‘common people’ on the website.\textsuperscript{121} As a result, Kharoof Tayeh construes himself as a concerned Mahjoobian, a populist voicing popular sentiment toward the authorities.

Kharoof Tayeh’s switch from Arithmographemic Latin-scripted Arabic to English in Line 2 is interesting because, on the surface, it solicits information from other forum members. But, as becomes clear later in the thread, the question was rhetorical and Kharoof Tayeh admits that he has no intention of calling the religious authorities because he feels that they will not take his criticisms seriously. Nevertheless, with the current post, Kharoof Tayeh attempts to portray bravado by confronting the religious authorities, something that Snipe_aac will challenge (see below).

In terms of identity, Kharoof Tayeh’s use of Arithmographemic Latin-scripted Arabic along with English within the same message positions him as a Vernacular Arabic-English bilingual poster typical of the mahjoob.com website, where discourse markers such as \textit{6ayyeb} are commonplace (a variant form of \textit{6ayyeb}, \textit{6ab} occurs at position 1451 in the top 2000 token wordlist). The use of the smiley alongside Arithmographemic Latin-scripted Arabic is another tactic that serves to index an informal style and to identify Kharoof Tayeh as a typical Mahjoobian. Note the irony

\textsuperscript{121}Fairclough’s notion of ‘genre chains’ refers to the process whereby texts are reproduced discursively for various social purposes (Fairclough, 2003, p. 31). He cites the example of an interview transcript between an academic researcher and a business manager. A genre chain in this case might be that the initial text is transformed into an academic report, which is then transformed into a handbook for educational managers, and finally, into a policy document for organizations (ibid., p. 32). In the case of the Al-Arabiya article, Kharoof Tayeh takes part in transforming this text into a populist manifesto for challenging the political and religious authorities in the GCC.
and humour injected by the use of the whistling emoticon at the end of Line 3, conveying the idea that something unjust is being overlooked by the authorities. Even in his use of English, Kharoof Tayeh conveys informality by using so-called texting orthography items like ‘u r’ for ‘you are’ (see Crystal, 2001) as seen in this extract from his Message 62 addressed to Snipe_aac:

I'm just asking simple questions and u r posting long long posts about different 30lama2 (Religious) scholars and their opinions. [emphasis mine]

Now consider the first half of Kharoof Tayeh’s Message 6, his second post in the thread:

Message 6:

1. two questions:

2. where is the hai2a, and it's stand regarding sending 300 billions, significant part (religious) organization

3. of it belong to saudi's nationals fe belaad alferenjeh?Is wearing in European countries

4. "improper" hijab more distructive than sending 300 billions, to our islam? headscarf

5. Do we really need to call the hai2a hot line to let them know that this

6. money is not serving islam? do they even read newspapers? Have the

7. hai2a seen tapes of reem bent elwaleed's wedding and what was Reem (daughter of) Al-Waleed's

8. going on it it? Do they not see that having 14 expensive cars in

9. sheekhah 7afsah 2al so3ood's garage is israaf..w la israaf feldeen? Sheikha Hafisa Al-Saud's excess... and there is no excess in religion?

Here, Kharoof Tayeh continues the use of Latin-scripted Arabic and English. English dominates, but Latin-scripted Arabic is used referentially to describe certain members of the Saudi royal family, the amr bel ma3roof and nahi 3an elmonkar i.e. religious authorities (see Line 10 below), and certain religious terms such as hijab (Line 4) and israaf (Line 9). On the one hand,
these referential code-switching examples of Latin-scripted Arabic might simply highlight Kharoof Tayeh’s lack of knowledge of the correct spelling or naming conventions in English for these referents, though *hijab* does in fact conform to Standard English spelling. On the other hand, however, they may in fact betray Kharoof’s purposeful attempt to reaffirm his identity as a Mahjoobian layperson.

In contrast to the layperson style assumed by Kharoof Tayeh, the last phrase in Line 9, *w la israaf feldeen* (lit. “and there (is) no excess in religion”) is striking because it is a Classical Arabic construction, which appears to be a Qur’anic quotation and mimics Qur’anic style, but this phrase does not even occur in the Qur’an. By employing such pseudo-Qur’anic phrasing, despite claiming lay status, Kharoof Tayeh is actually attempting to strengthen his condemnation of the excesses of the ruling family of Saudi Arabia, appropriating the style and authority of the Holy Qur’an.

Also of interest is Kharoof Tayeh’s use of the following Latin-scripted Arabic phrase: *belaad alferenjeh* (lit. ‘Frankish countries’ i.e. European countries). This phrase evokes the medieval crusades waged by Frankish kings and has racist connotations. By using this phrase as opposed to a more neutral phrase such as *belaad aluroobia* (European countries) Kharoof Tayeh indexes an anti-Western stance. In parallel to this, he emphasizes his solidarity with, and ownership of, his religion with the English phrase “our islam” (Line 4).

---

122 In Arithmographemic Latin-scripted Arabic, *hijab* would be rendered ‘7ijab’, ‘7ejab’ or even ‘7jab’, forms attested elsewhere in the corpus.

123 An important contradiction is worth noting here. Despite Kharoof Tayeh’s clear antipathy toward the West implied by his condemnation of the support of GCC governments for Western economies, Kharoof actually uses
In the second half Message 6, Kharooft Tayeh’s identification with Islam continues:

10 Guys..fellow muslims..I'm not on a campaign against amr belma3roof and nahi 3an
   Cause for Propriety and Forbidding of

11 elmonkar?..at some point of my life, I was a lost soul swimming in alcohol and filth..and
   Sins

12 because this amr belma3roof and nahi 3an elmonkar (in america not saudi arabia) I was
   Cause for Propriety and Forbidding of Sins

13 saved by fellow muslims and 7amdlah back to praying, fasting, serving islam to my best
   Praise (be) to God

14 ability. So yes, we need amr belma3roof and nahi 3an elmonkar.but what we don't need is
   Cause for Propriety and Forbidding of Sins

15 people who use doublestandard when doing this amr belma3roof. It is being implemented
   Cause for Propriety

16 on the weak..but not the strong...and the hai2a needs to go back and read the hadith
   religious organization holy tradition

17 "men afdal aljehad 3end allah, kalemato 7aqq 3end sul6an ja2er" aw kama qal
   "among the best holy war for God and his word is true is against a wrongful sovereign" or as has said

18 rasool allah PBUH. I call on the hai2a...7ayya 3aljehad in kontom sadeqeem. 😜
   the Messenger of God PBUH 124 religious authorities...hasten to holy war if you are sincere. offersflower

Here, Kharooft Tayeh employs Arithmographemic Latin-scripted Arabic again to quote Islamic religious traditions in Classical Arabic as seen in Lines 17 and 18, reminiscent of his mimicked Qur’anic style in Line 9. By doing so, he is once more attempting to establish the authority of his argument for action against the rulers based on religious injunctions from the past. Once again,

English, the dominant Western language, extensively. However, his use of English will be seen to contrast starkly with posters such as Muslim4 whose code use is explored below in this chapter. Regardless, the appropriation of English as a linguistic medium to voice anti-Western sentiments cannot be overlooked here.

124 The acronym PBUH stands for ‘peace be upon him’. This phrase translates the Arabic ‘alaihi as-salām which is habitually used when mentioning the names of recognized prophets in Islam such as Moses, Jesus, and especially, Mohammed (see Chapter 6) as an invocation of God’s blessings upon their revered persons. The convention has carried over into English and is frequently used by Muslim writers whenever they refer to any of the prophets of Islam in their writings.
Kharoof Tayeh positions himself as a Muslim who is familiar with the original Arabic language of the holy writings of his faith. Simultaneously, Kharoof Tayeh is issuing a rhetorical challenge to the religious authorities in Latin-scripted Arabic to take on their rulers by appropriating the voice of the Prophet Mohammed himself, further strengthening Kharoof Tayeh’s projected style of a conscientious and observant Muslim, who was, according to his own narrative, led back to Islam by these same authorities working in America (see Lines 10-14). At the same time, Kharoof Tayeh uses an emoticon of friendship at the end of the message in Line 18, which comes across as an invitation and conciliatory gesture for the hai2a (and perhaps the other posters as well) to join his side of the cause.

A word on the use of script here: Kharoof Tayeh’s decision to post Classical Arabic quotes written in Latin-scripted Arabic has two paradoxical effects on the audience: 1) it gives his post a sense of informality and familiarity and 2) it demonstrates that he knows the language of the Qur’an. In this way, while Kharoof Tayeh is not presenting himself as an academic as Muslim4 or Snipe_aac do (see below), he does manage to portray himself as a concerned lay Muslim, who is conversant with the Qur’an, and is speaking on behalf of the ‘people’ and of their religion and not on behalf of the ‘establishment’. Kharoof Tayeh continues in this way to position himself as a Muslim populist in his subsequent posts, alternating between Arithmographemic Latin-scripted Arabic and English. Nevertheless, Kharoof Tayeh’s initial choice to import a news website article written in Arabic-scripted Arabic is noteworthy in that it helps to demonstrate that he can at least read Modern Standard Arabic.

Much later, in Message 63, Kharoof Tayeh’s concluding message to the debate, he strives to re-establish goodwill in the thread, which has become quite heated and combative by this point, by
thanking his opponents, using Jordanian Vernacular Arithmographemic Latin-scripted Arabic as opposed to the Classical Arabic he employed earlier in his calls for action against the rulers:

Message 63:

1  yen6eeko el3afyeh..magassarto
   (God)  bless you all       thank-you all        hatsoff

2  mosh megtene3 fe kalamkom..bas allah yejzeeko elkhair..magassarto.
   I am not convinced       by your words ..  but may God grant you all    blessings     thank-you all

This message again positions Kharoof Tayeh as a layman who, despite philosophical differences, still bears no ill-will to participants that do not share his views. The use of male-sounding Arithmographemic Latin-scripted Jordanian Vernacular Arabic125 here without any English qualifies Kharoof Tayeh primarily as a Jordanian Arab male and indexes kinship across the ideological divide with the other posters, who are presumably also ethnic Arab males. Note also the use of the hatsoff smiley126 in Line 1 to convey a sense of play and to soften the tone of debate (see Su, 2007). Such linguistic behaviour will be seen to be highly salient when compared to that of other posters in the debate such as Snipe_aac.

125 See Section 2.3.2 for a discussion of the strategic use of /g/ in place of /ʔ/ as a marker of masculinity among Jordanian males in both face-to-face and online contexts (see also Al Share, 2005; Holes, 1995).

126 The so-called hatsoff smiley 😊 is actually an animated smiley which, when viewed in a web browser, shows the happy face figure removing his hat respectfully. As a symbol it serves to show respect and appreciation to the posters. Kharoof Tayeh is eager to appear to maintain a sense of goodwill toward the Pro-establisments posters even though he continues to disagree with them.
7.6.3 Pro-establishment author: Snipe_aac

Kharoof Tayeh’s linguistic behaviour reflected a desire to project an Anti-establishment Muslim populist identity. A Pro-establishment author, Snipe_aac’s linguistic behaviour creates a very different persona from Kharoof Tayeh’s. To begin with, consider Message 7 (below), which is Snipe_aac’s first post to the thread, and replies to Kharoof Tayeh’s Message 1 at the beginning of Thread 206940:

Message 7:

1. You seem to harbour much hatred for the Hay'ah.  

2. If you really have a complaint and are sincere:

3. This is the phone numbers for the ra'ees:

4. 8386600 – 8362200

5. And this is for the Mudeer:

6. 8345408

7. His assistant:

8. 8365617

9. If this doesn’t work for you, I'll try give you Shaykh 'Abdul-Azeez Aal ash-Shaykh,

10. the grand mufti's number. If you cannot get in touch with him, I'll give you one of

11. the numbers to the members of the Hay'a Kibaar al 'Ulemaa in Saudia.

In this first message by Snipe_aac, the most prolific contributor to both the entire Thread 206940 and to the first 75 messages of the thread\textsuperscript{127}, there are clear linguistic differences with Kharoof

\textsuperscript{127} Snipe_aac contributes a total of 31 posts out of 75 in this sub-set of Thread 206940’s messages.
Tayeh’s posts. From the beginning, Snipe_aac establishes himself as a Pro-establishment, Muslim academic. In order to achieve this, he periodically uses relatively formal English vocabulary and structure: “you seem to harbour much hatred for the Hay’ah” (Line 1 above), although note his use of contractions in Lines 9 and 10. This semi-formalized style while writing in English is maintained throughout all of his subsequent posts such as Message 8 below:

Message 8:

1. What we learn when we read the books of Salaf concerning the affair of advising, *Pious ancestors*
2. such as Ibn Rajab’s book on the topic, we find that they always speak of advising
3. before refuting. And this is a known methodology of the ‘Ulemaa. *Religious Scholars*

Particularly noticeable, however, is Snipe_aac’s deliberate and consistent use of Code 10 (Non-arithmographemic Latin-scripted Arabic with English). For instance, compare his ubiquitous usage of *hay’ah* to Kharoof Tayeh’s use of *hay2a*. Indeed, Snipe_aac employs Non-arithmographemic Latin-scripted Arabic exclusively to transliterate Classical Arabic or Standard Arabic quotes and terms. This explains why 16 out of 31 of his posts in this debate are marked as written in Code 10 (Non-arithmographemic Latin-scripted Arabic). And even Snipe_aac’s three Code 12 (Non-arithmographemic Latin-scripted Arabic with English and some Arabic script) messages and his five Code 14 (Non-BNC English) messages in the debate reveal the same orthographic tendency to avoid use of arithmographemics altogether while transliterating Arabic items.128 In fact, Snipe_aac’s avoidance of arithmographemics is so systematic that, of his total

128 Snipe_aac’s number of Non-arithmographemic Latin-scripted Arabic messages is even higher when his messages in Code 14 (Non-BNC English) are considered. All of these contain Non-arithmographemic Latin-scripted Arabic items which were so rare (i.e. occurred less than five times) in the corpus that they did not become part of the wordlist used to identify texts as belonging to Code 10 (Non-arithmographemic Latin-scripted Arabic with English)
592 messages found in the corpus (see Table 7.2 above), not one of these is written in Code 3 (Arithmographemic Latin-scripted Arabic with English), despite the fact that Code 3 is the most widely used code for composing following messages in the entire corpus, representing 35% of the total (see Section 5.3.4). In contrast, 316 of Snipe_aac’s 592 messages (i.e. over 50%) are written in Code 10, a relatively obscure code that accounts for only 4% of all following messages in the rest of the corpus (see Section 5.3.4). Clearly, then, Snipe_aac’s strong preference for Code 10 distinguishes him from a large number of Mahjoobians including Kharoof Tayeh, who, as was seen above, positioned himself as a typical Mahjoobian by using Code 3 so extensively.

Another striking contrast with Kharoof Tayeh is Snipe_aac’s almost total avoidance of markedly vernacular forms of Arabic in his posts with the exception of formulaic expressions common to both Classical Arabic and Vernacular Arabic such as BarakAllaahu feek ‘God bless you’ (Message 72). Viewed together, Snipe_aac’s strategic use of Non-arithmographemic Latin-scripted Arabic (and smileys), his avoidance of Vernacular Arabic, and his preference for formalized English combine to create a very salient persona, that of a serious and conscientious proponent of Islam. In fact, Snipe_aac’s method of non-arithmographemic transliteration mirrors the style found in the numerous texts written ostensibly by Salafist Islamic scholars129, which he (see Chapter 3). Examples of these items from his Code 14 messages are: Khaarijiyyah (dissenter movement), Sadaqaat (acts of benevolence) found in his Message 10.

129 The Salafist movement is closely linked to the Wahhabist ideology, the branch of Islam currently espoused by the Saudi government and other GCC states such as Qatar, Bahrain, the UAE, and Kuwait (see Lane, 2008, pp. 227-228). Its name derives from the salaf al-sāliH (lit. ‘the pious ancestors’), the agglomerate of early Sunni Muslim scholar saints. Its proponents advocate a return to a pristine Islam as practiced in the days of the Prophet Mohammed and immediately afterwards. Lane notes that Salafists are a widely divergent group in terms of ideology and political activism (2008, p. 228). To illustrate, although Bin Laden and his followers claim to be Salafists, their advocacy of
imports into his messages in order to back up his arguments (see Messages 11, 12, 13, 16, 18, 20, 30, 54, 55, and 56). By emulating the writing style of these Salafist authors in his own writing, Snipe_aac is attempting to identify with these Islamic scholars and to position himself among them (see Butler, 1990; Rampton, 1999, 2007).

Now consider another code-switching message of Snipe_aac written in Code 12 (Non-arithmographemic Latin-scripted Arabic with English and some Arabic script):

Message 52:

1. About 'Aydh al-Qarnee:

2. Shaykh Ahmad an-Najmee hafidhahullaah said:

3. Shaykh 'Abdullaah al-Bukhaaree said:

4. As for Safar al-Hawaalee, the one who had the nerve to label Imaam al-Albaanee

5. with Irjaa:

6. Amongst those who refuted his book Dhaahiraatul Irjaa was Shaykh 'Alee Hasan and

7. Shaykh Khaalid al-Anbaree and many others.

8. Shaykh Muqbil said about him that he's from the Khawaarij of our time.

Here, in addition to his usual Non-arithmographemic Latin-scripted Arabic with English, Snipe_aac employs Arabic-scripted Arabic. In so doing, he reaffirms his image as a violent uprising against corrupt Muslim rulers and their followers is condemned by other Salafists such as Snipe_aac (see Message 53), who argue for political acquiescence in the face of Muslim tyranny.
knowledgeable Islamic academic in several ways. First, he quotes directly from the works of Islamic scholars in the original, copying and pasting their Arabic-scripted Arabic texts into his post. Also, in terms of argumentation, Snipe_aac uses the statements of scholars to refute other scholars, a technique that also establishes the breadth of his scholarship\(^\text{130}\). Again, his Latin-scripted Arabic serves to transliterate Classical Arabic formulaic expressions such as “God preserve him” (Line 2), which are used perfunctorily in Arabic whenever respected living persons such as esteemed religious experts are mentioned. Snipe_aac also employs specialized Islamic theological terms such as *Irjaa* (Line 5) and *Khawaarij* (Line 8) (commonly referred to as ‘Kharajites’ in Western academia)\(^\text{131}\). Orthographically, the use of the apostrophe and of the dipthong “ee” for the personal name ‘Alee (Line 6) is noteworthy since this orthography contrasts both with the conventional English spelling “Ali” and with the Arithmographemic Latin-scripted Arabic spelling “3ali”. Such orthography parallels Snipe_aac’s re-spelling of other proper names from the Muslim world e.g. *Afghaanistaan* and *‘Iraaq* (see Message 23), and *Allaah* (see Messages 15, 16, 28, 54) found elsewhere in his thread messages. The overall effect of such re-spellings is twofold: First, Snipe_aac re-appropriates these terms for use by Muslims

\(^{130}\) Indeed, the Anti-establishment poster, Guillotine acknowledges Snipe_aac as a hard working student and then attacks his objectivity by alleging that Snipe_aac’s religious studies are funded directly by the Saudi government (see Message 65).

\(^{131}\) The Kharajites from *khawārij* (lit. “those who leave” the fold of Islam) were a group of Muslims who lived shortly after the death of Prophet Mohammed. They are remembered as having challenged the succession of rightly-guided caliphs after Mohammed and brought about Islam’s first internal religious wars. As such the Kharajites are vilified throughout Islamic history for having ruptured the unity of early Islam. Snipe_aac repeatedly mentions that the Anti-establishment criticisms and calls to holy war against the rulers are reminiscent of the actions of the Kharajites, essentially accusing the Anti-establishment authors of spiritual treason and apostasy.
for Muslims (as opposed to non-Muslims). As a result, he also seems to reflect a kind of orthographic purism,\textsuperscript{132} which may parallel a kind of religious purism as advocated by a Pro-establishment branch of the Salafist movement\textsuperscript{133} that Snipe_aac clearly supports. These points both seem to concur with evidence that such spellings reflect prevalent orthography among Salafist proponents when they write in English, a group who simultaneously attempt to reclaim authority in interpreting Islam correctly while advocating a return to its original purity as they conceive it (Lane, 2008, pp. 227-228) (see Message 60, Lines 1 and 2 below).\textsuperscript{134} Thus, by elevating his discourse above the widespread ‘slang’ arithmographemic variants exemplified in Kharooof Tayeh’s messages, Snipe_aac at once demonstrates his commitment to the (one and

\textsuperscript{132}Through the course of this research several articles were discovered on the Web featuring debates about proper transliteration of Arabic and Islamic terms into English written ostensibly by Muslims (e.g. http://www.soundvision.com/Info/education/muslimenglish.asp) and non-Muslims (e.g. http://micahtillman.wordpress.com/2007/09/25/anglo-islamic-spelling/).

\textsuperscript{133}In his work on varied ethno-cultural responses to globalization, Lane (2008) cites modern Salafism as an example: “The most general approach to Islamic fundamentalism is that of Salafism. Salafis idealize an uncorrupted, pure religious community know as the Salaf. The first three generations of Islam (Arabic: Salaf): Muhammad, his companions, and the followers of the companions from the earliest generations of Muslims, express the essential Islamic life. Their Sunnah, or practice, together with The Koran, presents an infallible guide to life, as innovations are forbidden. ‘Salafism’ depicts a school of thought that takes the pious ancestors (Salaf) of the patristic period of early Islam as exemplary models. Most puritanical groups in the Muslim world are Salafi in orientation. Salafism is not a sect per se but describes a simplified version of Islam, in which adherents follow a few commands and practices. Wahhabism (Saudi Arabia) is a particular orientation within Salafism, close to the Hanbali School” (Lane, 2008, p. 227).

\textsuperscript{134}Or, as the sahihalbukhari.com website puts it, “the creed and manhaj (methodology) of the salaf us-saalih – pure and clear” (sahihalbukhari.com, 2010).
only) ‘correct’ representation of his religion as much as to the correct practice of his religion, which forms the backbone of his whole argument. In support of this, consider Snipe_aac’s reply to Message 59 from Kharoof Tayeh:

Message 60:

Quote: Originally Posted by Kharoof Tayeh

loool muslim dude..inshalla..allah yehdeena kollna ameen yarab.

God willing.. may God guide us all, amen, O Lord.

samiulislam

so ur saying that there is one form of islam, and I seem to be missing this one? Don't worry, I'm not saying ennak betkafferni..no ma3ath allah. Where is this one islam is being practiced?

Snipe dude

maho ya 2akhi ur bombarding us with all this "cut and paste" of shyookh ana

By the way, Brother religious teachers I

masme3et feehom. I'm just asking simple questions and u r posting long long posts have (never) heard of

about different 3olama2 and their opinions. Forgive me brother

religious scholars

Forgive me brother
What akhoona Samiulislam said is that there are many interpretations of Islaam and only ONE is valid. And that is the interpretation of the salaf as Saalih. As what i’m bombarding you with, is information pertaining to this matter of the deen. If you seek the truth, this is what you want to look for, the explanations of the Ulemaa. Not a simple answer but an understanding with daleel. You cannot publically speak against any of the muslim rulers. The proof of this is found in the fataawa. If you desire to correct the rulers, advise them in private. [Emphasis mine]
‘akhook’ lit. ‘your brother’ referring to himself in Quote Line 9). Kharoof Tayeh also uses informal Netspeak forms (e.g. ‘u’ for ‘you’ and ‘ur’ for ‘you’re’ and ‘u r’ for ‘you are’ see Quote Lines 3, 7, and 8). He even pokes fun at Snipe_aac’s apparent erudition by saying that Snipe_aac has “cut and paste” from Sheikhs, i.e. religious leaders he has never even heard (Quote Lines 7 and 8). Again, Kharoof Tayeh’s vernacular Arithmographemic Latin-scripted Arabic here grounds him firmly among lay people on the website such as the posters in the General Forum and the Male Chef forum. Even the fact that he asks permission-seeking questions of the Pro-establishment supporters might be interpreted as feigning a subordinate, non-initiate status vis-à-vis the clerical authority indexed by the Pro-establishment posters. In stark contrast to Kharoof Tayeh’s levity, Snipe_aac ignores his opponent’s humour and maintains a serious tone. Moreover, when addressing Kharoof Tayeh, Snipe_aac avoids kinship terms and consistently uses the stylistically direct and formal pronoun ‘you’. In terms of stance, Snipe_aac reiterates his position as being on the side of the Ulemaa by referring Kharoof Tayeh back to their writings for guidance (Lines 3, 4, and 5).

In summary, Snipe_aac has separated himself stylistically from Kharoof Tayeh in several ways: through his use of non-arithmographemic Latin-scripted Arabic, through formalized English and an avoidance of smileys, and through importing Arabic-scripted Arabic texts and “Salafi/Wahhabi” English-style from various Salafist websites. As a result, Snipe_aac manages to project the persona of a serious scholar/student of Islam, as opposed to a lay person, a firm supporter of the political cum religious establishment of Saudi Arabia. At this point, an exploration of the linguistic behaviour of another Pro-establishment author, Muslim4 will serve to provide another intriguing example of strategic use of a particular code in order to project a clearly distinct persona.
7.6.4 Pro-establishment author: Muslim4

On the Pro-establishment side, linguistic consistency helps Muslim4 project a salient identity as it did Kharoof Tayeh on the Anti-establishment. But whereas Kharoof Tayeh consistently uses Code 3 (Arithmographemetic Latin-scripted Arabic with English), Muslim4 is noteworthy for preferring Code 1 (Arabic-scripted Arabic),135 which he uses for 10 of his 11 posts in Thread 206940. Such consistent Code 1 use identifies Muslim4 as someone who has likely studied Islam formally. Indeed, even Muslim4’s style of writing in Arabic-scripted Arabic is remarkable because of its high level of formality, further indexing erudition and contrasting sharply with the informality of Kharoof Tayeh’s Code 3. In fact, Muslim4’s first post to the thread in response to Kharoof Tayeh’s initial attack on the Hay’ah (see Message 1 above) features decidedly formal literary Arabic forms:

Message 5 (last two sentences):

وصد قني عندما تعود إلى الأردن وتخرج في المساء وتطل برأسك من النافذة وتُشاهد شاباً وفتاة يزنون في السيارة أمام منزلك ستعلم أهمية الهيئة

Translation:

And believe me, when you return to Jordan and venture out at night and lift your head up to the (car) window and witness a youth with a maiden fornicating in the car in front of your residence, you shall come to know the importance of the Hay’ah.

---

135 It should be borne in mind that across the Arabic-speaking Middle East and North Africa, Islamic religion as a field of study is taught primarily in Modern Standard Arabic while subjects such as Engineering and Medicine are often taught in English (see Hussein & Zughoul, 1993; Maamouri, 1998).
In this excerpt, several words and phrases are formal: تُشاهد شابا tushāhid shābban ‘you witness a youth’ in Line 1 (inflected in the indefinite singular accusative case) replaces the more casual تشوف شاب tushūf shabb ‘you see a young guy’, أمام منزلك amām manzalika ‘in front of your residence’ (Line 2) is found instead of the colloquially more common قدم بيتك gidaam baytak ‘in front of your house’, and ستتعلم sata3lam ‘you shall come to know’ (Line 2) is employed instead of the more colloquial لاح تعرف la7 t3araf ‘you’ll know’. Again, such a stylized formal Arabic writing style sets Muslim4 apart from most of the other posters in the thread (except for Guillotine, see below) and establishes him as a sophisticated and educated Arabic language user. This level of formality, albeit in a different language, emphasizes Muslim4’s solidarity with Snipe_aac, whose choice of expressions, e.g. “you seem to harbour much hatred for the ha’ya”, and avoidance of smileys and arithmographemes also reflected a formal style (see Section 7.6.3 above). Further, regarding formality, it is telling that, in contrast to the Anti-establishment posters, i.e. Kharoof Tayeh and Guillotine, Muslim4 avoids the use of smileys entirely in his posts to the thread. In this way, he again resembles other Pro-establishment posters such as Snipe_aac and SamiullIslam. Indeed, the overall absence of smileys lends an air of seriousness to the Pro-establishment authors’ messages when contrasted with the Anti-establishment authors’ messages.

In addition to his formal style, within four of his 11 posts, Muslim4 imports Arabic-scripted Arabic texts from a variety of sources in order to defend the Hay’ah and bolster his contention that it is wrong to challenge the authorities (see Messages 5, 24, 25, and 39). Such repeated...
importation of texts is reminiscent of Snipe_aac’s posts. For instance, Muslim4’s second post, Message 5, consists of a lengthy news article written entirely in Modern Standard Arabic copied from the website islamtoday.net. The article relates the killing of Saudi national who was in possession of drugs and firearms by a member of the Hay’ah during a raid on his home. The gist of the article is that while such private citizen vigilantism is controversial, it may in fact be justified due to the inability of the Saudi police force to respond in a timely fashion to such criminality. The other imported texts employed by Muslim4 in his posts include quotations from the Qur’an as well as references to Hadith, all given in the original Arabic-scripted Arabic. Again, such Arabic-scripted Arabic quotes coming from religious texts that were originally written in Arabic (as opposed to English translations) lend religious credibility, authority, and authenticity to Muslim4 as a Muslim initiate137, attributes which Guillotine will also attempt to portray in similar ways, as will be seen below.

Additionally, the pervasive use of Code 1 (BNC English) by Muslim4 within this debate helps to identify him with the Arabic-language-dominant posters on mahjoob.com. His use of Standard Arabic without smileys situates him among more serious posters such as those who frequent the Religion forum, where over 44% of all following messages were found to be written in Code 1 (see Chapter 5).

However, it is worth noting that Muslim4 does in fact deviate from his use of Code 1 (Latin-scripted Arabic) within the debate. Muslim4’s only message not written in Code 1 (Arabic-scripted Arabic) is his last contribution to the debate, Message 61. This message is particularly salient because it is written in Code 15 (Graphic content-only), indicating that it contains either

137 Also, by assuming the user ID Muslim4, Muslim4 is certainly attempting to highlight his affinity for Islam.
only imported material or graphic material and features no author-composed text whatsoever. Typically, Code 15 messages contain smileys without any text, URL links to other sites, rapidshare files for downloading, or imported images and pictures. However, Muslim4’s sole Code 15 message is different because it contains none of the above. Instead, Message 61 features two imported quotations (these were copied into the message as text boxes, i.e. pictures), the first of which comes from Message 59 written by Kharoof Tayeh, and contains a comment and question about advising rulers. The second quotation, which Muslim4 cleverly uses as reply to Kharoof Tayeh’s question, is actually an imported quotation from a segment of Message 54 originally written by Snipe_aac in Code 10 (Non-arithmographemic Latin-scripted Arabic with English). To illustrate, here is the entire message starting with Kharoof Tayeh’s quoted question and ending with Muslim4’s quotation of Snipe_aac’s Message 54 segment as an answer to Kharoof Tayeh’s question:

Message 61:

Quote (question imported from Kharoof Tayeh’s Message 59):

1. loool muslim dude..inshalla..allah yehdeena kollna ameen yarab. samiulislam so ur saying that God willing..God guide us all amen, O Lord
2. there is one form of islam, and I seem to be missing this one? Don't worry, I'm not saying ennak That you
3. betkafferni..no ma3ath allah. Where is this one islam is being practiced? Snipe dude maho ya Are calling me an infidel..no God forbid by the way, hey
4. zakh ur bombarding us with all this "cut and paste" of shyookh ana masme3et feehom. I'm just Brother shaykhs I've (never) heard of
5. asking simple questions and u r posting long long posts about different 3olama2 and their scholars
6. opinions. 😊 er7am akhook ya2akhi :D ya3ni momken tjawbo Hugginfriend God bless your brother, Brother I mean can you (guys) reply
7. ajwebeh basee6ah be yes or no or even "depends on" type of answers. Can I speak agaionst 2al With simple answers with the Al-
8. sa3ood based on what we hear and see of their actions (not generalizing, but specific omara2)?
Saud family

Does islam allow me to do that?

Quote (imported segment from the last part of Snipe_aac’s Message 54):

And there is a great difference between the ameer or the haakim about which you desire to speak out against is actually in front of you and between him being absent. Since, all of the rejections that have been reported from the Salaf, all of them took place in front of the ameer (pious) Ancestors or the haakim himself. Hence, the difference is that when he is present he is able to defend himself, and explain his viewpoint, and he could actually be right and we (the ones who criticise) could actually be wrong. Hence, if you are eager for goodness, then go to him, and face him, and advise him in that which is between you and him.

Perhaps the most striking feature of Muslim4’s message here is his deliberate attempt to reply to Kharoof Tayeh’s question posed in Code 3 (Arithmographemic Latin-scripted Arabic with English) without having to compose an answer in Arithmographemic Latin-scripted Arabic or English himself. By purposely avoiding the use of English (or Latin script for that matter), Muslim4 appears to maintain his persona as an Arabic-dominant and educated individual. One might be deceived into believing that Muslim4 is actually unable to write in English, which then prompts him to avoid using it. Nevertheless, the very fact that Muslim4 is able to reply to Kharoof Tayeh’s message by choosing a logical and appropriate quotation from among Snipe_aac’s English-language messages demonstrates Muslim4’s command of English. Even more interesting is the observation that outside of Thread 206940, Muslim4 does in fact use Code 2 (BNC English) 19 times in the corpus and Code 3 (Arithmographemic Latin-scripted Arabic with English) 52 times (see Table 7.2 above). Though these numbers are dwarfed by his use of Code 1 (Arabic-scripted Arabic) in the corpus (2258 messages), they do, nevertheless,
indicate Muslim4’s obvious ability to use Code 2 and Code 3. Consider, for instance, the seed message below (Message 1) written by Raskolnikov and posted to the Girls Talk forum under the Thread 232362 entitled, “Hijab Survey?” followed by Muslim4’s reply to it (Message 5):

Message 1 (seed message posted by Raskolnikov)

1 Speaking of the current hot Hijab issue, at least in this community, and of course out
2 of curiosity, do you wear Hijab, is it a "proper" one, based on your own standard? or
3 just a Hijab?
4 The poll is invisible, and no need to comment unless you feel the opposite.

Message 5 (replying message posted by Muslim4):

1 i dont wear hijab
2 runningaway

Here Muslim4 elects to reply to a post written by Raskolnikov in an ostensibly female-only forum on a question which is clearly directed at female Mahjoobians i.e. whether or not they wear the hijab (Islamic head scarf). Yet Muslim4’s reply, written in Code 2 (BNC English) is intended to be taken comically. First, as a male Muslim, he is obviously under no religious obligation to wear the hijab, so his comment “I don’t wear hijab” appears to have flouted Grice’s Maxim of Relation, which states ‘Be relevant’ (Grice, 1975). Second, his use of the runningaway smiley\textsuperscript{138} indicates that he is aware of this flouting and the potential response it may receive.

\textsuperscript{138} Like the hatsoff smiley mentioned above, the runningaway smiley is actually an animated smiley and shows the blue figure running away to the right when viewed directly in web-browser. Thus, a reader would read Muslim4’s comment in Line 1 and then observe the smiley figure running away.
Third, his English style is highly informal, Netspeak-like in fact, ignoring capitalization and apostrophe use rules. Such informality contrasts strongly with his Code 1 (Arabic-scripted Arabic use) in Thread 206940. That Muslim4’s comment was taken humorously by other posters becomes apparent when two other authors choose to reply to it: In Message 6, Mlabas_3_looz jokingly comments to Muslim4 “Allah yehdeek” (may God guide you) while in Message 12, Fiasco has the following one-word comment for him:

“KAAAAAAAAAAAAAAAAAAAAAAAAAFERRRRRRR” (lit. ‘infidel’). The capitalized flaming style of this apparent condemnation of Muslim4 for not wearing his hijab renders the comment all the more playful and humorous.

The above example of Muslim4’s use of Code 2 (BNC English) elsewhere in the corpus, highlights his ability to use codes strategically. Within a heated debate discussing an issue as serious as insurrection against the political establishment in Saudi Arabia found in the Religion forum, Muslim4 opts to project a persona that is formal, educated, religious, and linguistically-aligned with the Standard Arabic of Islam through his own use of Code 1 (Arabic-scripted Arabic), downplaying his own ability to write in English. In such a serious debate, there is no place for frivolity. However, in the Girls Talk forum, Muslim4 is more than happy to construct himself as possessing a more humorous identity, using Netspeak-like Code 2 and employing the type of smiley that he utterly avoided in Thread 206940 above.

Having considered Muslim4’s projection of a Pro-establishment identity through his use of Code 1 (Arabic-scripted Arabic), it will be valuable to compare such consistent use to that of Guillotine, the remaining main contributor to the debate. This will be the focus of the next section.
7.6.5 Anti-establishment author: Guillotine

Guillotine, from the Anti-establishment camp, presents patterns of code use that diverge from the other posters discussed thus far. As mentioned earlier, of all contributors to the debate, he is clearly the most multilingual, and this applies to his linguistic behaviour both inside Thread 206940 and across the rest of the corpus (see Table 7.2 above). First, consider that Guillotine’s first post to the thread, Message 2, comes immediately after Kharoof Tayeh’s seed message, which contained the imported article in Arabic-scripted Arabic accusing Gulf leaders of supporting Western governments inimical to Islam (see Section 7.6.2 above). Guillotine’s Message 2 consists of a single sentence written in Code 1 (Arabic-scripted Arabic) and is, on the surface, an invocation of blessing on the soul of the late King Faisal of Saudi Arabia: 

\[
\text{ الله يرحم الملك فيصل (lit. ‘God have mercy (on the soul) of King Faisal’).}
\]

But in reality, this message conveys the sentiment that King Faisal would be outraged if he were alive today to see how his country is being governed. Next, Guillotine’s second post to the thread, Message 14, is written in Code 3 (Arithmographemic Latin-scripted Arabic with English) and is addressed to Snipe_aac:

Message 14:

1. Bo3bo3, I know where do you stand and where are you coming from on this. The only
   \textit{Monster}

2. thing I can say regarding this issue is that Al Hay'a don't have the authority yet to control
   the organization

3. the acts of the king for example and the royal family, they can be rejected easily if they

4. interfer, and they can easily be brought to each one of them to go home and no Hay'a any

5. more. No one can tell still if in the future they will be able to do so or to be unraveled.
Notice the use of the term *Bo3bo3* (‘monster’) in Line 1, which Guillotine employs to address Snipe_aac. Used as a vocative, this term is a playful epithet that might be translated as “Hey, rascal” or “Hey, trouble”. By using this colloquial term written with arithmographemics, Guillotine adds a playful, yet subtly provocative dynamic to the exchange. Such casual usage is reminiscent of Kharoof Tayeh’s use of Code 3 discussed earlier and serves to identify Guillotine as a typical Mahjoobian. And indeed, Guillotine’s next post to the debate, Message 17 is written in Code 15 (Graphic-only content) and features only the *foolishsmacksownhead* smiley\(^{139}\) without any other accompanying text. Viewed in the context of the thread, Guillotine’s Message 17 conveys incredulity and frustration at Snipe_aac’s Messages 15 and 16 immediately preceding it, which feature lengthy imported texts about how past criticism of Muslim rulers led to disunity and bloodshed within Islam. Regardless of the intent of this message, the simple use of such a smiley in this heated debate, positions Guillotine alongside Kharoof Tayeh on the Anti-establishment side of the debate and reinforces his image as more of an average Mahjoobian than the Pro-establishment posters who avoid using smileys and Code 3 (Arithmographemic Latin-scripted Arabic with English) entirely during the debate.

Having established himself early on as a typical Mahjoobian and populist in line with Kharoof Tayeh, Guillotine exhibits a wider range of orthography and code choice in his subsequent posts, which varies depending on whom he addresses. For instance, when replying directly to Snipe_aac, who uses Latin-scripted codes such as Code 10 (Arithmographemic Latin-scripted Arabic with English), Code 14 (Non-BNC English), and Code 2 (BNC English), Guillotine most

\(^{139}\) The *foolishsmacksownhead* smiley 😱 is an animated smiley that displays the figure smacking his own head and closing his eyes as in shock or disbelief.
often uses Latin-scripted codes in kind such as Code 2, Code 14, and Code 3 (Arithmographemic Latin-scripted Arabic with English) (see Messages 14, 19, 22, 26, 32, 41, and 75). Even within these posts, at points, Guillotine actually approximates the Non-arithmographemic Latin-scripted Arabic orthography espoused by Snipe_aac by selectively omitting the arithmographeme “3” or replacing it with the Non-arithmographemic “\’”, e.g. Guillotine’s use of unho instead of 3anho (lit. ‘about/with him’) in Message 26, and ’ala huqq instead of 3ala 7uqq (lit. ‘on the truth’ i.e. ‘on the right path’) in Message 75.140

Similarly, when interacting with Muslim4, who almost always uses Code 1 (Arabic-scripted Arabic) in the debate, Guillotine also uses Code 1 exclusively (see Messages 29 and 34). In one way, this type of linguistic accommodation may be seen as a sign of respect for Muslim4’s own code use preferences. But perhaps, a second, more plausible reason for Guillotine’s Code 1 use with Muslim4 is that he wishes to establish himself as a sophisticated and fluent Arabic user on par with Muslim4. This is evidenced by the fact that, like Muslim4, Guillotine also uses stylistically formal Standard Arabic in several of his posts (see Messages 29, 34, 38, 44, 45, 65, 66, 67, 68, 69, 74). The same may also hold true for Guillotine’s use of non-arithmographemic forms with Snipe_aac. Thus, a sort of linguistic one-upmanship is established here whereby Guillotine seeks to outdo his opponents by proving that he can use their language equally well and is even more grounded in Islam than they are. Given the competitive nature of the debate, this is hardly surprising.

140 But notice within this same message that Guillotine uses the arithmographemic “3” as in 3alem Jaleel (lit. “grand scholar”) instead of the non-arithmographemic style favoured by Snipe_aac, ‘Alim Jaleel.
Further evidence of Guillotine’s attempt to outdo his opponents by establishing himself as more knowledgeable and credible in interpreting Islam and (therefore more justified in his calls for challenging the Saudi establishment) can be seen in both the frequency and content of his messages composed in Code 1 (Arabic-scripted Arabic). Typically, Guillotine composes Code 1 messages with imported Arabic-scripted Arabic texts from various Islamic sources such as the Qur’an and several Hadith and uses them to defend his Anti-establishment stance (see Messages 38, 44, 45, 65, and 67-69). By citing Arabic-scripted Arabic texts straight from the original religious works as opposed to relying on the interpretation of scholars translated into English as Snipe_aac generally does, Guillotine seeks to assert that his own interpretations of calls to action against corrupt rulers and Ulamaa are more valid and in line with the original Islam. In other words, Guillotine seeks to usurp the authority from scholars to interpret texts such as the Qur’an and Hadith for him. Thus, he is directly challenging the Pro-establishment Salafist stance put forth by Snipe_aac and Muslim4 (and SamiulIslam, see Message 60) that the only way to interpret Islam and its teachings correctly is by relying on the interpretations of the Salaf (lit. ‘(pious) ancestors’) and the scholars who follow them. By quoting Qur’anic and Hadith passages directly in Arabic-scripted Arabic (i.e. not in translation as Snipe_aac tends to do) and interpreting them, Guillotine is de facto establishing himself as a qualified interpreter of Islam in his own right. In fact, in Message 65\textsuperscript{141} addressed to Snipe_aac, Guillotine voices such sentiments:

Message 65 (translated from Arabic-scripted Arabic):

\textsuperscript{141} Due to the length of the original message written in Code 1 (Arabic-scripted Arabic), only the translation is given here. The original message is available in Appendix G.
With all due respect, Brother Snipe, since you are the hard working student who is trying to attain (divine) knowledge, you cannot simply discredit any scholars as you fancy and then re-embrace them and disown them again whenever it pleases you to do so. Surely, there is a way for me to say to you now, “Fear God for your own sake and for the sake of the Muslim scholars who have spoken the truth which did not accord with your fancy nor the fancy of the rulers’ scholars. Indeed, you have shared, clarified, and provided evidence and proofs that suited you. Yes, you provided these while repudiating what I offered you from the Hadith, preferring your scholars over the Messenger of God, peace be upon him, and over his holy companions. My question now is direct, and if you consider it a private matter, then do not answer it: “What is the motivation behind your desperate defence of the ruler and his kin which forces you to denounce Shaykh Aidh Al-Qarni and then to hold him in esteem again according to your wishes and whims? Is the Saudi government supporting the funding of your research and studies?”

---

142 The Sahābah (lit. ‘companions’) of Prophet Mohammed were his earliest followers and contemporaries such as his wives, members of his family, closest friends, and successors. Their recollections of the words and sayings of the Prophet Mohammed form the basis of the several thousand Hadiths (holy traditions) which supplement, and provide applications of, the teachings of the Qur’an.

143 Aidh Abdullah ibn Al-Qarni is a contemporary Muslim scholar and author (see Islamicthinkers.com, 2010). During the debate he is variously cited as a reliable religious authority and repudiated by several authors.
This is a stinging post in which Guillotine levels three accusations against Snipe_aac, namely, that he whimsically upholds and rejects scholars as it suits his argument, that he prefers the teachings of the pro-ruler scholars over those of the Prophet Mohammed and his disciples, and finally, that his religious studies are funded by the Saudi authorities, which would help to explain Snipe_aac’s strong Pro-establishment bias. That it is written in a formal style of Standard Arabic serves to heighten its severity and sternness. Again, by using Code 1 (Arabic-scripted Arabic) here, Guillotine achieves three mutually-supportive goals: 1) he indexes Islamic authenticity by writing eloquently in Standard Arabic, the language most closely associated with of Islam, 2) he outdoes Snipe_aac linguistically, who, apart from quoting Arabic-scripted Arabic occasionally, never composes full messages in Code 1 in the debate, and 3) he demonstrates his respect for, and alignment with, the Arabic sayings of the Prophet Mohammed and his holy companions over the English translations of these and of other scholars cited by Snipe_aac throughout the debate.

To summarize, initially, Guillotine uses linguistic markers such as Code 3 (Arithmographemic Latin-scripted Arabic with English) and smileys in order to position himself as a concerned Mahjoobian, much like Kharoof Tayeh. Then, Guillotine takes on a more competitive and adversarial persona by reflecting back to his key opponents, Snipe_aac and Muslim4, their own code choices. In this way, Guillotine attempts to establish his own erudition and credibility in light of the arguments the Pro-establishment posters make, backed up by various imported texts. Next, while distinguishing himself clearly from the Pro-establishment Salafists, Guillotine actually attempts to outmanoeuvre them by citing Qur’anic verses and Hadith directly to support his argument and discredit corrupt rulers (see Messages 44 and 45) and scholars (see Message 67). In this way, Guillotine shows himself to be one of the most nuanced authors linguistically,
portraying a complex identity of a learned Muslim who, nevertheless, identifies with his co-religionists who are outraged by the corruption of the establishment.

7.7 Conclusions and summary

To put all these findings into context, Table 7.3 below summarizes the key similarities and differences between the four main authors discussed here. Following this, a bulleted list of the main findings of this chapter is also given.
Table 7.3: Four main authors in Thread 206940 contrasted: Kharoof Tayeh, Snipe_aac, Guillotine, and Muslim4

<table>
<thead>
<tr>
<th>Author ID</th>
<th>Stance/Views</th>
<th>Code Use</th>
<th>Stylistic features</th>
<th>Projected Identities</th>
</tr>
</thead>
</table>
| Kharoof Tayeh | **Anti-establishment:** Gulf rulers, religious leaders, and Hay'ah members are corrupt; Hay'ah leaders should challenge the authorities when they contravene Islam | **- Within thread 206940:** Code 3 exclusively, no Code 1 or Code 2  
**- Across corpus:** Uses Code 3 most; never composes messages in Arabic script himself | - uses smileys and humour  
- uses Arabic kinship terms | - Typical mahjoobian  
- Young male Jordanian  
- disgruntled populist  
- observant Muslim layperson |
| Snipe_aac | **Pro-establishment:** questioning Muslim authorities openly is sinful and leads to sectarianism; religious and Hay'ah leaders can and do advise the leaders in private about how to be better Muslims | **- Within thread 206940:** Uses Code 10 mostly, then Codes 14, 6, and 12. Never uses Code 3.  
**- Across corpus:** Uses Code 10 most, only uses Code 3 three times out 604 posts, uses Code 1 occasionally | - lack of humour  
- formal English  
- no smileys  
- almost no kinship terms | - mainstream Muslim scholar  
- English-dominant Salafist  
- Believer in the divine right of the State to rule |
| Guillotine | **Anti-establishment:** Gulf rulers, religious leaders and Hay'ah members are corrupt; it is morally-binding for Muslims to rise against any tyrannical leader even if he is a Muslim | **- Within thread 206940:** Uses Code 1 mostly, then, Codes 2 (x4), Code 3 (x3), and others.  
**- Across corpus:** Uses Code 1 mostly, followed by Codes 14, 2, 3 | - uses Code 3 and smileys  
- uses Code 1 with Muslim4  
- uses Code 1 and Code 2 with Snipe_aac  
- serious, combative tone but uses smileys occasionally | - knowledgeable Muslim  
- disgruntled populist  
- Well-read bilingual |
| Muslim4 | **Pro-establishment:** The Hay'ah do a great job of protecting society; Muslims are never justified in disobeying a Muslim ruler | **- Within thread 206940:** Uses Code 1 almost exclusively, copies and uses only 1 Code 2 message  
**- Across corpus:** Uses Code 1 mainly, but also Codes 2 and 3 | - uses only Code 1 without smileys  
- highly formal and serious tone (despite his use of Code 2 and humour in other forums) | - Arabic-educated  
- Mainstream Muslim scholar  
- Arabic-dominant Salafist  
- Believer in the divine right of the State to rule |
Main Findings:

1. Author-related findings:
   A. Anti-establishment authors:
      - favour Code 3 (Arithmographemic Latin-scripted Arabic with English)
      - behave like average mahjoobians, using Code 3, smileys, and humour
      - use Classical Arabic lexicogrammar to bolster their arguments about Islam
      - advocate challenging the authorities
   
   B. Pro-establishment authors:
      - favour Codes 1 (Arabic-scripted Arabic) and/or Code 10 (Non-arithmographemic Latin-scripted Arabic with English) for serious debate, but Code 2 and smileys for humour
      - copy Salafist style e.g. Code 10 (Non-arithmographemic Latin-scripted Arabic with English)
      - use numerous quotations in Code 1 and Code 10 to back up their points
      - avoid using smileys and humour

2. Code and code use findings:
   - Code 10 closely resembles Salafist English writing style
   - Code 3 indexes familiarity, humour, and populism
   - Code 1 indexes Islamic erudition and authenticity
   - Some authors select the code of their interlocutors e.g. Guillotine, while other refuse to accommodate e.g. Snipe_aac
   - Some authors avoid use of a given script altogether (e.g. Kharoof Tayeh avoiding Arabic)

3. Identity-related findings
   - Some authors purposely avoid using certain codes despite knowing these (e.g. Muslim4 avoiding Code 2 in Thread 206940) in order to portray a distinct identity
   - Some authors use the codes of their interlocutors to show themselves as intellectual and linguistic equals to these interlocutors (e.g. Guillotine)
The analysis of the four authors presented in this chapter has shown that, in a given thread on a specific topic, the language types encountered in the corpus can in fact be employed by different posters in very different ways to project distinct identities. In this thread at least, the use of Classical Arabic lexicogrammar appears to index Islamic religious authority and moral correctness and superiority, regardless of script. Interestingly, though, script choice seems to inject either a formal and standard feel to the Arabic, or an informal, subversive element to it such as when Kharoof Tayeh uses Arithmographemic Latin-scripted Arabic to transliterate various Qur’anic verses and Hadith. Even given a specific script choice, different orthographic representations appear to have different effects. On the one hand, a rendering of religious terms using Code 10 (Non-arithmographemic Latin-scripted Arabic with English) coincided with the Salafist, Pro-establishment message content of posters such as Snipe_aac and SamiulIslam. On the other hand, use of Code 3 (Arithmographemic Latin-scripted Arabic with English), which is unknown within the scholarly tradition, but indisputably common for articulating the popular discourse on the rest of the website, seemed to parallel Kharoof Tayeh’s attempt to distance himself from the political and religious establishment (and Salafist movement) while aligning and identifying himself with the website majority. In the end, the effect of these linguistically-differentiated styles is to leave the reader with the impression of two very different characters: the law-abiding, Muslim scholar and conservative who uses Code 10 (Non-arithmographemic Latin-scripted Arabic) or Code 1 (Arabic-
scripted Arabic) vs. the disgruntled Muslim populist citizen and anarchist who identifies with other Mahjoobians through use of Code 3 (Arithmographemic Latin-scripted Arabic with English). Over-generalizations, though, should be avoided, as it was seen that Guillotine managed to cross linguistically (see Rampton, 1999), passing himself off as an authority on Islam by composing Code 1 messages and imitating Code 10-type orthography.

As noted earlier, for the sake of brevity, a decision was taken to concentrate on a single thread within a single forum on the website. However, it is almost certain that several other linguistically-realized styles and identities can be traced discursively on the website. In this regard, the original Labovian identity features such as age, sex, class, etc. may actually help provide me with axes for defining them. For instance, the specific themes of each forum, in a similar fashion to the Religion forum, are likely to provide posters with written contexts in which to construct various identities and locally-recognizable character types such as online gamer and hacker, Jordanian Christian, Palestinian activist, housewife with young children, etc. Whatever these many distinctive styles may be, it is clear from the examples shown here that for the purpose of investing contrastive styles in their posts, Mahjoobians have at their disposal a range of linguistic systems i.e. the scripts, lexicogrammars, and orthographies which are linguistic currencies on the mahjoob.com website.

In summary, this chapter began with a discussion of the relationship between code-switching and identity as well as orthography and identity, highlighting that several
studies have shown that strategic linguistic behaviour did often reveal identity-creation motives. Next, the criteria for data selection and the methodology for the analysis of the data involving discourse analysis were presented. The thread selected for analysis, Thread 206940 was described along with the contextual background necessary to make sense of the analysis. Finally, the identity-related code use behaviours of the four main posters to the subset of Thread 206940 were examined and compared. The analysis appeared to support the hypothesis that differential code use on the part of certain authors can help them to project distinct identities in line with their own stated ideologies and beliefs. Nevertheless, it is important to be aware of one obvious limitation of this analysis: there is always a risk of misinterpreting the authors’ motives and intentions as well as the effects of their texts upon their addressees when clear textual evidence is not present to substantiate such motives, intentions, and effects.

In Chapter 8, the concluding chapter of this study, the various findings regarding the discovery, distribution, topical focus, and identity-related values of the distinct codes in the corpus will be examined in conjunction.
Chapter 8: Conclusions

8.1 Overview
This chapter presents a summary of the specific findings of the four research questions presented in Chapters 3, 5, 6, and 7 respectively, highlighting the original contributions of the thesis to field of sociolinguistics. After this, reflections on the thesis are offered especially concerning its limitations, unexpected results, and lessons learned from it. Finally, the implications of the research and future directions are discussed including potential expanded use of the mahjoob.com corpus.

8.2 Summary of research findings
To put the research findings into perspective, it will be helpful to consider the four original research questions again:

1) What are the various types of code and script-switching featured on the English website of Mahjoob.com?

2) How are script-code pairings distributed across certain types of postings in terms of
   a) text type (title, first message, following messages, quoted content)
   b) overarching forum topic
   c) thread length
   d) author posting frequency

3) What do frequent lexical items in the main script-code pairings in the corpus reveal about the topical content and stylistic features of those pairings?

4) How do forum posters use the available script-code pairings to construct identities?
Briefly, these questions were ordered in such a way as to provide a logical investigative progression, i.e. each had to be answered before the next could be addressed: 1) Defining the primary linguistic objects i.e. the script-code pairings (codes) of the study present within the corpus, 2) determining the frequency and distribution of these linguistic codes in the corpus in terms of text type, topic, author type, and thread length, thereby identifying the main codes in the corpus, 3) analysing the main codes to determine their broad topical focus and stylistic features through concordances of their respective top 10 most frequent lexis, and 4) in light of the broad topical and stylistic patterns associated with each code, investigating micro-level instances of code choices among forum contributors to identify the strategic use of codes to construct distinct identities.

The following bulleted list summarizes the most salient findings with regard to each of these questions:

1. The code and script switching found on the English website of mahjoob.com:
   - The most common form of script-switching is the use of Latin script with arithmographemics to transliterate (mainly) Vernacular Arabic
   - the English-website contains more Code 3 (Arithmographemic Latin-scripted Arabic with English) and Code 1 (Arabic-scripted Arabic) messages than it does Code 2 (BNC English-only messages), challenging the dominance of English on the internet
   - messages featuring Arabic-script and Latin-script are surprisingly rare suggesting that changing between scripts may be inconvenient and/or that many posters are mono-scriptal
   - numerous messages contain only imported content or non-linguistic content e.g. Codes 15 and 16.

2. The distribution patterns of the codes:
A. Text type findings:

- thread titles are composed in Code 2 and Code 1 rather than Code 3, perhaps because Codes 1 and 2 are more informational in function
- The mono-scriptal dominate the corpus especially Codes 1, 2, and 3 and to a far lesser extent Codes 10 (Non-arithmographemic Latin-scripted Arabic with English) and 14 (Non-BNC English), suggesting that script-switching is not widely used
- Of the three main codes, Code 3 is the most used code in following messages while Code 2 is the least used, suggesting differences in function

B. Overarching topics:

- Code 1 is used for Poetry, Humour, and Local culture forums
- Code 2 is used for Work/Study related forums
- Code 3 is used for General discussion forums

C. Thread length:

- Longer threads are composed in Code 1, shorter threads in Code 3

D. Posting frequency

- Top 10 posters compose in Code 1, non-top posters in Code 3
- Top 10 posters have a tremendous impact on code use, accounting for 20% of all following messages and 12% of all Code 1 messages
- Top 10 posters mostly post in Joke Zone and Poetry-related forums

3. The topical and stylistic features of Codes 1, 2, and 3 as indicated by top 10 lexis

- Code 1 is topically and stylistically-dichotomous, where Arabic vernacular is used for Humorous content and phatic functions while MSA/Classical Arabic is used for religion, highly reminiscent of Arabic diglossia
- Code 2 is topically diverse and contains references to taboo topics
- Code 3 is mainly English linguistically, and its Latin-scripted forms are usually discourse markers, expressions, and Arab cultural references
- Code 3 is more phatic than Codes 1 and 2, with frequent use of smileys
4. The use of codes to construct identities

- Code 3 indexes solidarity with other mahjoobians and young Arabs
- Code 1 indexes scholarship and authenticity
- Code 10 indexes Salafist ideology
- Posters choose which codes to use based on topic, level of formality, and interlocutor

Most salient overall findings:
- the English-website is highly heterogeneous and script-switching, though rare, does occur in fact occur
- codes overlap in topical content and function, but Arabic, English, and Arithmographemic Latin-scripted Arabic (3arabizi) do have distinct functional uses
- 3arabizi is the most common code in the corpus, but is mainly English with a smaller amount of actual Latin-scripted Arabic
- mahjoob.com provides a space for Arabic bilinguals to engage in a wide range of topics including politically-charged and taboo ones such as political and religious dissent, sex, and sexual orientation
- Top 10 posters contribute considerably to the website, but have code and topical preferences very different from non-top 10 posters, preferring Code1 (Arabic-scripted Arabic) and Humour and Poetry-related forums

It is worthwhile to discuss these findings in more detail. First, the considerable linguistic heterogeneity of the corpus was clearly established, highlighting the overall dominance of the mixed code, Code 3 (Arithmographemic Latin-scripted Arabic with English) in the corpus followed by Code 1 (Arabic-scripted Arabic), and then Code 2 (BNC English). Evidently, mono-scriptal codes are by far the more popular choice among the website posters, a hypothesis posited by Al Share (2007) that she attributes to the inconvenience
involved in changing keyboard sets while composing messages based on the results of her
own study of IRC involving Jordanian Netspeak and English (see Al Share, 2005), an
assumption that has been hard to test on a large-scale basis until now. In this regard,
certain codes were found to associate more clearly with certain forums such as Code 1
with poetry, humour, and religion-related forums. Cursory observation of these forums
suggested that this was so because, on several occasions, content in these forums may
have been copied and pasted from other Arabic-scripted Arabic websites on the Internet.
As a corollary, Code 2-dominant forums such as *Health and Science* perhaps reflected a
similar process of large scale importation of texts such as health-related news articles (see
files in “health_and_science” folder in Appendix B). Consequently, Code 3-dominant
forums appear to be those that feature the most original content composed primarily by
Mahjoobians themselves.

Regarding the findings related to the second research question, a number of observations
can be made here: First, the 14 textual codes of the corpus differ in both quantity and
distribution across the corpus. Second, mono-scriptal codes are by far the most common
type of code in the corpus, especially Code 1 (Arabic-scripted Arabic), Code 2 (BNC
English), and Code 3 (Arithmographemic Latin-scripted Arabic with English). Third,
topic has a certain relationship to code choice where Humour, Poetry, and Local Culture-
related forums reveal more Code 1 use, while Work/Study forums favour Code 2, and
General Discussion, Age/Gender-related, Friends/Family-related, and Hobby-related
forums exhibit more Code 3 use. In this regard, to a certain degree, the findings seem to
parallel Bentahila’s study (1983) highlighting the functional distribution of
Classical/Modern Standard Arabic, Vernacular Moroccan Arabic, and French in Morocco where Classical and Standard Arabic are used for religious and classical literary domains, Vernacular Arabic for intimate and informal settings, and French for professional and academic purposes. Lastly, in terms of mahjoob.com authors’ code use patterns, prolific authors favour Code 1 over both Code 2 and Code 3 which are preferred by the bulk of non-prolific authors for composing their following messages. This concurs with the finding that prolific authors post very frequently to the Joke Zone forum in Code 1, which in itself may betray frequent importation of humorous material composed in Arabic-scripted Arabic from other humorous websites, though such an assertion is difficult to test given the problems associated with determining ultimate authorship on the web (Hoffmann, 2007).

The findings stemming from the third research question revealed that Code 1 (Arabic-scripted Arabic), Code 2 (BNC English), Code 3 (Arithmographemic Latin-scripted Arabic with English) while exhibiting degrees of topical overlap, were clearly distinct in several ways. For instance, of the three codes, Code 1 seemed to be most closely connected to the topic of religion and to exhibit a formal and non-involved style. On the other hand, Codes 2 and 3 were much more involved in style and featured a much broader range of topics which paralleled the major topics of the website’s forums. Yet despite certain topical and stylistic similarities between these two codes, in contrast to Code 2, Code 3 was most clearly connected to Vernacular Arabic and local Arab culture, featuring a relatively light, humorous, and playful style as evidenced by the preponderance of smileys found in the concordances of its top 10 frequent lexical items. The awareness of
such discursive function and stylistic differences between the three major codes of the website helped provide the necessary background knowledge in order to carry out the micro-level analysis of identity-creation as required by the fourth main research question of the study. Indeed, the step-by-step investigative process guided progressively by the first, second, and third research questions made it possible to then examine whether the distinct linguistic codes in the corpus could be used strategically in support of identity-construction. The results in this regard were positive. Specifically, it was found that, whether written in Arabic-scripted Arabic or Latin-scripted English (see Muslim4’s and Snipe_aac’s respective uses of Arabic and English to further the same arguments in Chapter 7), stylistically-formal utterances as marked by distinct lexis and orthography were used systematically to establish erudite and Pro-establishment identities that contrasted rather sharply with the vernacular and informal lexis and orthography of the Anti-establishment posters. In this regard, recalling Fishman’s categorization of multilingual contexts (see Section 2.2.1), perhaps due to the relative novelty and flexibility of its sociolinguistic context, mahjoob.com appears to provide an example of bilingualism without diglossia insofar as its main codes often overlap in their topics and discursive functions despite observable trends for specific codes to be used for certain topics, levels of formality, and communicative functions (see Chapters 5 and 6).

Having reviewed the major findings of the thesis, it is now opportune to reflect on the various limitations of this work as well as lessons learned from the investigative process followed here.
8.3 Reflections on the thesis

The originality of the thesis in terms of content lies in the fact that, to date, there appear to have been no studies done on script and code-switching involving Arabic and English within discussion forums, and certainly not with a corpus of almost half a million messages. In this sense, the data have helped to provide yet another angle from which to investigate Arabic-English code-switching and specifically, script-switching involving the Arabic and Latin scripts in particular. Nevertheless, despite the interesting findings noted above, it is not warranted to generalize these beyond the data presented here. This is because code-switching is a situated practice, reflecting the peculiarities of a given context. For instance, it may be the case that a large number of Mahjoobians are based outside the Arab world and consequently have to use Latin script even though they would prefer to use Arabic script. Thus, the lack of direct feedback from forum posters on their code use represents a major drawback of this study, though such author feedback on code use would naturally reflect self-perceptions of the authors' behaviour rather than their actual behaviour.

Another limitation, from a logistical point of view, is the fact that I had to rely extensively upon another researcher (Sebastian Hoffmann) who was skilled in Perl script and text editor applications. This meant that the corpus in this study was more the fruit of collaboration than of individual design and effort. While such collaboration is not negative in and of itself, the vast amounts of collaborative time and effort required to produce the corpus in its present state (which is still far from flawless) suggests that such a time and labour-intensive methodology may not be suitable to a majority of researchers.
As with the technological aspects of the thesis, I also have to acknowledge certain linguistic and cultural limitations on my part. For one, although I have lived in the Middle East for over a decade now and have been in close contact with Arabs of the Levant for much longer than this, I am admittedly not a native speaker of Arabic nor a member of the cultural in-groups featured in this study. As such, both my linguistic and cultural perspectives and insights are certainly more limited than if I had been born and raised in a Jordanian or Palestinian Arabic-speaking milieu. I have done my utmost to address these potential liabilities by having recourse to colleagues and assistants from the target cultural group in order to render assistance in interpreting and translating the data correctly. Nevertheless, due to my non-initiate status, I have doubtless been oblivious to certain aspects of the data that might have been equally interesting to investigate. Having said this, because of my own limitations, the possibility of making this corpus available to other researchers, especially those who are native Arabic-speakers, is an especially exciting one, as such scholars may actually be able to uncover even more compelling research questions than the ones addressed here.

In terms of the findings themselves, one key discovery is particularly intriguing: the widespread use of Code 3 (Arithmographemic Latin-scripted Arabic with English) when compared with the more conventional code-script pairings represented by Code 1 (Arabic-scripted Arabic) and Code 2 (BNC English). If Code 3 texts tend to be authored by posters themselves as opposed to being imported from other sources as appears to often be the case with the other two codes, then from an educational perspective, the potential value of Latin-scripted Arabic, as a pedagogical aide (see Maamouri, 1998) in
the development of literacy (even bi-literacy) should not be ignored. Regardless, from a sociolinguistic perspective, the fact that Code 3 was found to be the most common code for composing following messages in the corpus, suggests the vitality of a form of language which is at once innovative, hybrid, unplanned, unofficial, and unsupported. In this light, it is fascinating to consider whether a grassroots, youth-driven Code 3 i.e. 3arabizi/Arabish movement across the Arabic-speaking world might be capable of challenging the primacy of Modern Standard Arabic in a way akin to Demotic Greek’s supporters’ struggle against, and ultimate triumph over, Katharevousa’s supporters, or, much earlier, the triumph of neo-Latin vernaculars such as French and Italian over Church Latin across medieval and early Renaissance Catholic Europe. Perhaps just as intriguing as the linguistic ramifications of the potential spread of 3arabizi is the social upheaval which its spread implies, namely, the apparent wresting away of linguistic hegemony from established centres of learning such as government-sanctioned schools, universities, academies, and seminaries where Arabic is highly prescribed and regulated and the appropriation of this hegemony by youth within relatively unregulated spaces (see Sebba, 2007) such as web forums.

8.4 Implications of the research and future directions

Despite the time and effort required to build the mahjoob.com corpus, or perhaps because of it, the corpus should certainly continue to be investigated. On the one hand, granted the relatively large size of the corpus and the variety of its data, a virtual plethora of new research questions have emerged. Moreover, given the existence of dual corpora, one for WordSmith 5.0 and the other for SPSS, based on the same original data set of messages, the opportunities for analysis and triangulation are considerable. At several points, I had
to curb the urge to delve into many ad hoc research questions that presented themselves. To illustrate, consider that code-switching phenomena have been traditionally viewed as either situational (speaker and context-related) or metaphorical (e.g. topic and discourse-related). This study has clearly focused on the latter. Nevertheless, using the SPSS-version of the corpus, it would theoretically be possible to carry out several statistical tests of correlation between the code (or script for matter) of the first message in a thread and that of the second message to see whether initial code choice conditioned in some way subsequent code choice. Similarly, several tests could be run to determine code choice over time across a group of authors; i.e. do new authors post in different codes than more established authors? Other demographics could be investigated: how many authors are multilingual/multiscriptal?

In the post-9/11 era, a time when Western society seeks to better understand the Middle East and Arab culture in order to avoid the so-called clash of civilizations, the large array of texts in the mahjoob.com corpora offers a veritable treasure chest of data to explore by combining corpus analysis and discourse analysis methods. In this regard, several potential questions remain unexplored on a wide range of social and cultural issues many of which have been broached elsewhere without the benefit of such a large data set:

1. What are prevalent attitudes toward Islamism and Jihad among Mahjoobians?
2. How is Arab culture defined and differentiated from Western culture?
3. How are gender and sexual orientation constructed in the data set?
4. How are key figures such as Osama Bin Laden, Tony Blair, Saddam Hussein, George W. Bush represented in texts?
5. What kinds of discussions are had concerning hijab?

Beyond this list of questions, at the risk of sounding clichéd, the opportunities for further research involving the mahjoob.com corpus are virtually endless based on its sheer volume of texts coupled with its two distinct formats which are each relatively easy-to-search using WordSmith 5.0 and SPSS respectively.

8.5 Concluding comments

This research has sought to investigate areas to which scant attention had been paid in the field of sociolinguistics until relatively recently: written code-switching and online code and script choice in support of identity creation. In this regard, this work has highlighted several important aspects of such linguistic behaviour. First, humans as complex linguistic beings are capable of appropriating pre-existing linguistic practices and re-moulding them to suit their own communicative and identity-related purposes. In so doing, they are able to create highly innovative linguistic varieties such as Arithmographemic Latin-scripted Arabic which then acquire their own stylistic and communicative values and functions separate from their putative linguistic forbears, i.e. Arabic-scripted Arabic and Latin-scripted English. Then, in similarly creative fashion, language users further innovate by strategically taking up (or shunning) these linguistic varieties and their features in order to create distinct and salient identities which are readily recognizable to their interlocutors. All this occurs in a medium devoid of physical voice and sound. In some sense, one might be led to believe that the physical world is somehow being reproduced in the virtual world by authors where orthography takes the place of accent. However, the possibility for script-switching adds an element of style.
which has no parallel in the spoken world. As a result, it is clear that the online medium allows for certain nuances of identity-expression which are unavailable in face-to-face environments. In a certain sense, Mahjoob.com forums appear to represent an online community practice, an amalgam of influences and inputs that represent a different kind of reality, no matter how similar they are to face-to-face reality, resembling as they do face-to-face clubs, pubs, and meeting rooms. Similarly, the written codes of mahjoob.com, exemplified best by the hybrid code of 3arabizi/Arabish defined here as Code 3 (Arithmographemic Latin-scripted Arabic with English), are an amalgam of influences that, while deriving much input and structure from face-to-face forms of language, are clearly distinct from these in both form and function. Indeed, the online identities created by Mahjoobians through their skilful use of written codes, though certainly inspired by real-life character types, are nevertheless only made possible by virtue of the online medium in which they are situated. Thus, mahjoob.com, perhaps above all else, is a reminder of the fascinatingly complex interplay between language user and medium, and may be a harbinger of potentially greater technologically-influenced language use and identity creation in the future.
References


