Language and Topic Choice among Prolific and Non-Prolific Posters on an Arabic-English Website

R. Bianchi

***Abstract*—Mahjoob.com is a popular Jordan-based website featuring dozens of discussion forums in both English and Arabic. This paper explores the language and topic choices among the 1,261 posters that authored posts on mahjoob.com during a 14-month period. The results indicate that the top 10 prolific posters (i.e. those who have posted more than 1000 messages) have very different language and topic preferences to the rest of the posters. Prolific posters prefer to post using Arabic and to contribute to humor-related forums whereas non-prolific posters prefer to post in 3arabizi, a mixture of Arabic and English written in Latin script, and to a lesser extent, in English. These non-prolific posters tend to post to a variety of other topical forums besides the humour-related forums.**

***Index Terms*—Arabic, CMC, code choice, discussion forums.**

I. Introduction

This study presents findings from a doctoral study that investigated code and script choice on the popular Jordan-based website, mahjoob.com. The website is divided into Arabic-language and English-language sections and the data that informs the study was taken from a corpus of forum text messages downloaded from the English-language section of the website. At the time of data collection between March 2007 and May 2008, the English section featured some 41 topical forums and had 1,261 posters. The resulting corpus contains some 460,220 messages found within 21,626 discussion threads spread across the 41 topical forums. The English section of mahjoob.com was chosen for data collection because, in comparison to the Arabic section, it is notable for its highly multilingual and multiscriptal nature. Indeed, in addition to English, the English section also features a large number of messages written in Arabic-scripted Arabic and 3arabizi, a hybrid mixture of English and Arabic written in Latin script, which uses arithmographemes i.e., numerals as letters as in its name *3arabizi.* Other messages featured within the English section forums were written in Salafi English, a sort of Muslim English, in non-standard English, and in a mixture of Arabic and Latin script.

II. Literature Review

B. Danet and S. Herring (2007) provide an introduction to the emergent phenomenon of computer-mediated communication (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 3arabizi as opposed to either Classical Arabic or Modern Standard Arabic.

J. Androutsopoulos observes that "bilingual interaction is still a neglected issue in the study of the multilingual Internet" [1]. 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, His 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.

R. Wodak and S. Wright [2] 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[1]. 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 J. 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 [3], [4].

M. Warschauer, G. R. El Said, and A. Zohry examine

Manuscript received July 5, 2013; revised September 10, 2013.

R. Bianchi is with the Virginia Commonwealth University, Qatar (e-mail: rmbianchi@vcu.edu).

[1] A *seed post* refers to an opening post i.e. the initial post that starts off a given thread.

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…[5]

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.

D. Palfreyman and M. Al Khalil [6] 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.

B. Al Share [7] 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 whereas web forums are a form of asynchronous CMC and therefore afford participants more time and reflection in both production and reception of texts. 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, B. 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. 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 posters 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 almost no studies to date done on Latin-scripted Arabic in web forums.

B. Al Share [7] provides an orthographic description of CMC-based Latin-scripted Arabic among Jordanian web chatters similar to D. Palfreyman and M. Al Khalil's study in the UAE [6]. B. 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 B. 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 8]

Of particular relevance to the present study is the fact that in his illustration of contexts of diglossia, C. 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, C. 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, C. 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 [9]. Consequently, it is interesting to consider whether any carry over occurs from the face-to-face environment into the online environment. However, 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, B. Al Share [7] 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 [10], [11].

III. Data and Method

As mentioned above, the data were collected from the mahjoob.com website. Using a Perl script, all messages between March 2007 and May 2008 were downloaded and annotated into text file-based corpus. A second stage involved creating an SPSS database version of the corpus where each message, poster, thread, forum, etc. could be cross-tabulated

with one another and with several other variables such as language used, time of posting, poster location, etc.

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 message posters were consistent in their code use and whether they were similar to the average poster. Based on this, the research question underpinning this paper is: How are languages (aka codes) and topics distributed in terms of poster posting frequency within the mahjoob.com corpus? To address this, using SPSS, data was collected to determine the most prolific posters.

Initially, information on the most prolific posters 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 only ten prolific posters who had posted at least 1000 messages in the corpus. Consequently, these prolific posters were dubbed the "Top 10 posters". To explore further the possible impact of these prolific posters on code distribution in the corpus, the top 10 posters were grouped together in order to compare their code use patterns to the remaining 1,251 posters[2]. The regrouping of these posters entailed defining a new SPSS variable ("top10_authors") by recoding all messages posted by a top 10 poster with the value "1" and assigning the remaining messages the value "2". It was now possible to examine poster behaviour across both code choice and topic.

The examination of code distribution patterns across poster type offered 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. Thus, in order to ensure that the 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 variable poster frequency had a 5% or less likelihood of having occurred by chance. The Chi square test of significance revealed that all differences between prolific and non-prolific posters were in fact significant.

IV. Findings

The top 10 most prolific posters were indeed found to be different from the non-prolific posters in terms of their preferred topical forums to post in and their choice of code as (see the paneled bar charts in Fig. 1 below). In order to clarify the data within the chart, it will be useful to highlight how the variables are organized. The Y-axes in each chart show the percentage of following messages whereas the X-axes show the eight overarching topics that posters can choose to post within: 1) Humour, 2) Poetry, 3) Work/Study, 4) Family/Friends, 5) Local Culture, 6) Hobbies, 7) Gender/Age-related, and 8) General Discussion/Topics. The numbered colour segments of each bar refer to the codes in which poster can post their messages as follows: No. 1 - Arabic-scripted Arabic (blue), No. 2 – BNC English, No. 3 – 3arabizi (beige), No. 4 – Mixed Latin and Arabic script (purple), No. 10 – Salafi English (yellow), and No. 14 – Non-BNC English (red). It is important to note that Fig. 1 presents 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 needs to be mentioned that messages composed by the top 10 prolific posters account for a full 20% of all following messages in the entire corpus.

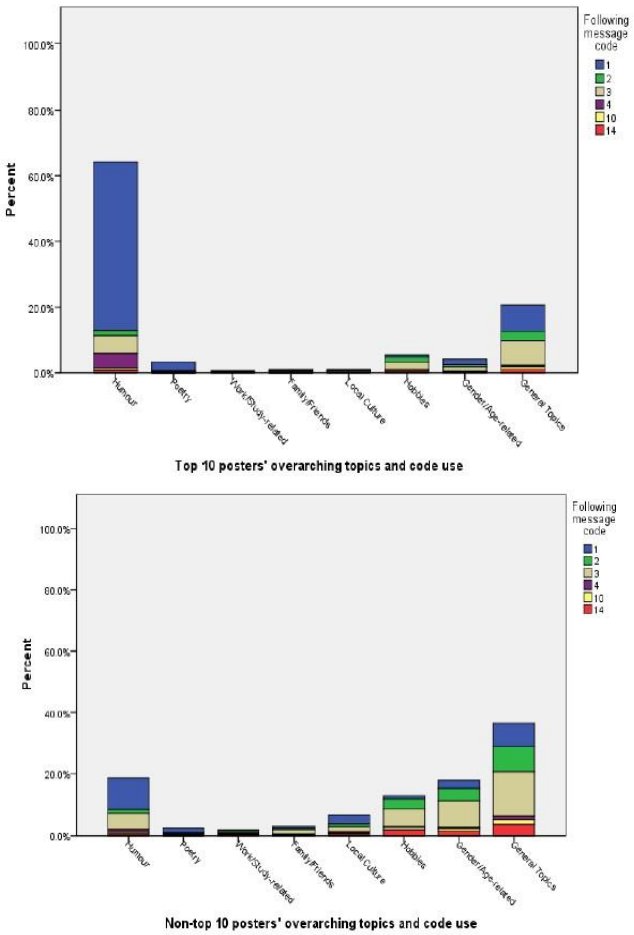

Fig. 1. Top 10 posters vs. non-Top 10 posters.

Before considering differences, it is worth noting that 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

[2] The impact of such prolific posters could not be overlooked since it was determined that the top 10 posters alone accounted for roughly 20% of all forum messages in the corpus.

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 Arabic (Code 1). This is not surprising given their tendency to post to Humour and Poetry-related forums which have been shown to be connected to Arabic in the corpus[3]. However, top posters also appear to use Code 1 relatively more often for General Topic messages at about the same rate that they use 3arabizi (Code 3) for these. In contrast, non-top posters tend to use BNC English (Code 2) and 3arabizi far more often. Indeed, for Hobby forums, Gender/Age-related forums, and General Discussion forums, the non-top posters prefer 3arabizi and, to a lesser extent, BNC English.

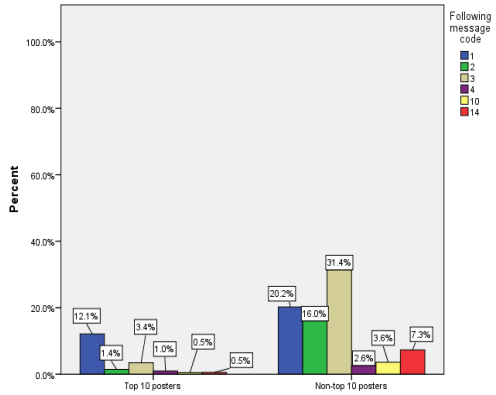

Fig. 2. Top 10 poster vs. non-top 10 poster code use.

V. Conclusion

To sum up, top posters contribute to Humour and Poetry-related forums more often and make use of 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 Arabic use is accounted for by a small group of posters, posting in a limited range of forums. Indeed, it is interesting to note that Arabic-language Joke Zone following messages alone account for 66.7% of all Arabic messages in the entire corpus. In this connection, it is also worth mentioning that the top posters, who are only ten in total, account for a full 37.5% of all Arabic following messages in the entire corpus whereas the remaining 1,251 posters account for the remaining 62.5% of Arabic following messages. In fact, these top 10 poster Arabic 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 Fig. 2 above). As mentioned earlier, 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.

Thus, in the mahjoob.com corpus there are clear differences between prolific and non-prolific posters. Interestingly, the relatively narrow linguistic and topical focus of these posters means that, while they account for a large proportion of the content of the web forums, they are responsible for a great deal of the Arabic language as well as and the humorous and poetic content of the web forums. In contrast, 3arabizi and English are more popular with the average poster within the forums, suggesting that, despite its relative novelty and informality, the hybrid language of 3arabizi is a viable means of communication for a majority of posters on mahjoob.com.

[3] 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.

References

[1] J. Androutsopoulos, "Language choice and code switching in german-based diasporic web forums," in *The Multilingual Internet: Language, Culture, and Communication Online*, B. Danet and S.C. Herring, Editors, 2007, Oxford University Press: Oxford.

[2] R. Wodak and S. Wright, "The European Union in Cyberspace: Democratic participation via online multilingual discussion boards," in *The Multilingual Internet: Language, Culture, and Communication Online*, B. Danet and S.C. Herring, Eds., Oxford: Oxford University. 2007.

[3] J. Paolillo. "How much multilingualism? Language diversity on the Internet," in *The Multilingual Internet: Language, Culture, and Communication Online*, B. Danet and S. C. Herring, Eds., 2007, Oxford: Oxford University.

[4] R. Bianchi. "Revolution or fad? Latinized Arabic vernacular," in *the 11th TESOL Arabia Conference Proceedings*, Dubai: TESOL Arabia. 2006.

[5] M. Warschauer, G. R. E. Said, and A. Zohry. (2002). Language choice online: Globalization and identity in Egypt. Journal of Computer-Mediated Communication. [Online]. 7(4). Available: http://www. jcmc.indiana.edu/vol7/issue4/warschauer.html

[6] D. Palfreyman and M. Al Khalil. (2003). A Funky Language for Teenzz to Use: Representing Gulf Arabic in Instant Messaging. Journal of Computer-Mediated Communication. [Online]. 9(1). Available: http://www. jcmc.indiana.edu/vol9/issue1/palfreyman.html

[7] B. A. Share, "A sociolinguistic analysis of Jordanian Netspeak (JNS)," M.A. thesis, Jordan University of Science & Technology: Amman. Jordan, 2005.

[8] M. Sebba, "Spelling and society: the culture and politics of orthography around the world," Cambridge: Cambridge University Press, 2007.

[9] C. C. Ferguson, "'Diglossia' 1959," in *Language and Social Context: Selected Readings*, P. Giglioli, Ed. London: Penguin, 1985, pp. 232-251.

[10] Y. Nishimura. (2003). Linguistic Innovations and Interactional Features of Casual Online Communication in Japanese. *Journal of Computer-Mediated Communication*. [Online]. 9(1). Available: http://www. jcmc.indiana.edu/vol9/issue1/nishimura.html

[11] L. Hinrichs, "Jamaican Creole on the Internet: Forms and functions of an oral language in computer-mediated communication," PhD thesis, Freiburg University Freiburg: Germany, 2005.

Robert M. Bianchi was born in Toronto, Canada on April 7, 1972. Bianchi received his PhD in Applied Linguistics from Lancaster University, UK in 2012. Bianchi's research field is sociolinguistics, specifically discourse analysis of multilingual texts in online contexts.

He has worked in English language teaching and higher education for over 15 years. Currently, he serves as Assisant Professor of English at Virginia Commonwealth University in Qatar in Doha, Qatar. He has also taught English and French in Japan at Bertlitz and in Canada at the York University English Language Institute (YUELI). He also taught in Oman at the Sur College of Education, in the UAE at the UAE University, and in Qatar at the College of the North Atlantic-Qatar. He has published several articles on Arabic-English code choice and bilingualism in TESOL Arabia, BAAL, and Acta Linguistica Asiatica. His current research interests revolve around the status of Arabic in Qatar.