Contemporary linguistic views of conversational Arabic describe constantly shifting “mixed” forms of the language, incorporating elements from Modern Standard Arabic (MSA) and several regional dialectal forms (Holes 2004, Shiri 2002). As part of an investigation of social dynamics in Massive Open Online Courses (MOOCs), we identified dialect shift in the threaded discussion forums of an Arabic Nanotechnology MOOC held by Coursera. While previous studies have looked at code-switching and language choice by Arabic speakers in online environments (e.g. Bianchi 2014, Kadir et al. 2012), we were specifically interested in the extent to which speakers show sensitivity in their dialect-relevant lexical choices to the displayed lexical choice tendencies of their interlocutors. We operationalized this sensitivity in terms of a shift in lexical usage away from the personal distributions of speakers to those displayed in two alternative interlocutive contexts, the message that started the thread and the immediately previous message. Our findings are twofold: First, although the language of the interactions was predominantly MSA, significant variation in lexical distributions was associated with geographic region, supporting the claim that differences in regional associations may be perceived between interlocutors. Second, speakers’ lexical choices exhibit a sensitivity to the lexical choice tendencies displayed by the proximal interlocutive context of their interactions.

We adopted a variationist approach, quantifying effects through a statistical hypothesis testing methodology. First, we constructed a set of variables representing each forum interaction’s dialect-relevant lexical choices. Specifically, these variables characterize the displayed dialectal bias within each post, the dialectal bias displayed by the interlocutive context of the post, and the personal dialectal bias of the post author themselves. The dialectal biases of the post and its discussion context were estimated lexically using the approach taken by Nguyen, et al. (2010) in their work on displayed ideological bias. The post author’s personal dialectal bias was estimated as the average dialectal bias across all posts by that author.

To evaluate whether regional lexical norms are reflected in a speaker’s lexical choice tendencies, we tested for variance of these tendencies with the speaker’s regional dialect. As a proxy for dialect, we used geographic dialect region information from each user’s IP address. Using an analysis of variance (ANOVA) model to separate post bias distributions using the nominal region variable, we find significant evidence of this effect (F(6,419) = 4.56, p < .0005). More specifically, we found the lexical choices of speakers in the Maghrebi and Levantine dialect regions were more biased towards MSA than those from speakers in other regions.

Next, we tested for lexical dialectal shift sensitive to interlocutive context. Using ANOVA models, we separate post bias distributions using the bias displayed by each of the two operationalizations of interlocutive context outlined above, controlling for a speaker’s normative tendencies using their personal bias estimate. In order to avoid artificially inflating the correlation between a user’s personal bias and the bias displayed in individual messages, we excluded messages from users with fewer than 10 posts altogether. This left 216 messages from 12 users for the analysis. Using the model above, we find the effect of the bias displayed more proximally in the immediately previous message was significant (F(1, 2013) = 3.86, p = .05, effect size .24) such that more bias was displayed by the speaker when more bias was displayed in the context. The data also provides weak evidence that the strength of this shift varies by region. However, the effect of the bias displayed more distally in the first post of the thread was not significant. Thus we find that while speakers’ lexical choices reflect regional norms, they exhibit a sensitivity to immediate interlocutive context.

Code-switching is a widely studied topic in sociolinguistics. We are specifically interested in diglossic communities in which code-switching between a lingua franca and a dialect is thought to play an important role in social positioning. Myers-Scotton’s (1993) ‘markedness’ model for code-switching in conversation proposes an inherent social motivation for code-switching, but the extent to which this model applies has also been a subject of debate (Auer 1998, Myers-Scotton 1998). This paper presents a sociolinguistic analysis of a corpus of computer-mediated communications among Arabic speakers from twenty countries, evaluating the extent to which the perceived ethnic identity of interlocutors has an effect on lexical choice in online discussions in an academic context in which the lingua franca would typically be regarded as the contextual norm.

Social Dynamics in Online Education: Linguistic Variation in an Arabic Massive Open Online Course

Keith A. Maki
Language Technologies Institute
Carnegie Mellon University
kmaki@cs.cmu.edu

Carolyn P. Rosé
Department of Education in Technology and Science
Technion Israel Institute of Technology
ccprose@cs.cmu.edu

Miriam I. Barak
Carnegie Mellon University
kmaki@cs.cmu.edu

Language

Code-switching also provides weak evidence that the strength of the bias context in which the 'bias' (MSA) and several regional dialectal forms (Holes 2004, Shiri 2002) was found the lexical choices of speakers in the Maghrebi and Levantine dialect regions. We performed a linguistic analysis of a corpus of computer-mediated communications among Arabic speakers from twenty countries, evaluating the extent to which the perceived ethnic identity of interlocutors has an effect on lexical choice in online discussions in an academic context in which the lingua franca would typically be regarded as the contextual norm.

Contemporary linguistic views of conversational Arabic describe constantly shifting “mixed” forms of the language, incorporating elements from Modern Standard Arabic (MSA) and several regional dialectal forms (Holes 2004, Shiri 2002). As part of an investigation of social dynamics in Massive Open Online Courses (MOOCs), we identified dialect shift in the threaded discussion forums of an Arabic Nanotechnology MOOC held by Coursera. While previous studies have looked at code-switching and language choice by Arabic speakers in online environments (e.g. Bianchi 2014, Kadir et al. 2012), we were specifically interested in the extent to which speakers show sensitivity in their dialect-relevant lexical choices to the displayed lexical choice tendencies of their interlocutors. We operationalized this sensitivity in terms of a shift in lexical usage away from the personal distributions of speakers to those displayed in two alternative interlocutive contexts, the message that started the thread and the immediately previous message. Our findings are twofold: First, although the language of the interactions was predominantly MSA, significant variation in lexical distributions was associated with geographic region, supporting the claim that differences in regional associations may be perceived between interlocutors. Second, speakers’ lexical choices exhibit a sensitivity to the lexical choice tendencies displayed by the proximal interlocutive context of their interactions.

We adopted a variationist approach, quantifying effects through a statistical hypothesis testing methodology. First, we constructed a set of variables representing each forum interaction’s dialect-relevant lexical choices. Specifically, these variables characterize the displayed dialectal bias within each post, the dialectal bias displayed by the interlocutive context of the post, and the personal dialectal bias of the post author themselves. The dialectal biases of the post and its discussion context were estimated lexically using the approach taken by Nguyen, et al. (2010) in their work on displayed ideological bias. The post author’s personal dialectal bias was estimated as the average dialectal bias across all posts by that author.

To evaluate whether regional lexical norms are reflected in a speaker’s lexical choice tendencies, we tested for variance of these tendencies with the speaker’s regional dialect. As a proxy for dialect, we used geographic dialect region information from each user’s IP address. Using an analysis of variance (ANOVA) model to separate post bias distributions using the nominal region variable, we find significant evidence of this effect (F(6,419) = 4.56, p < .0005). More specifically, we found the lexical choices of speakers in the Maghrebi and Levantine dialect regions were more biased towards MSA than those from speakers in other regions.

Next, we tested for lexical dialectal shift sensitive to interlocutive context. Using ANOVA models, we separate post bias distributions using the bias displayed by each of the two operationalizations of interlocutive context outlined above, controlling for a speaker’s normative tendencies using their personal bias estimate. In order to avoid artificially inflating the correlation between a user’s personal bias and the bias displayed in individual messages, we excluded messages from users with fewer than 10 posts altogether. This left 216 messages from 12 users for the analysis. Using the model above, we find the effect of the bias displayed more proximally in the immediately previous message was significant (F(1, 2013) = 3.86, p = .05, effect size .24) such that more bias was displayed by the speaker when more bias was displayed in the context. The data also provides weak evidence that the strength of this shift varies by region. However, the effect of the bias displayed more distally in the first post of the thread was not significant. Thus we find that while speakers’ lexical choices reflect regional norms, they exhibit a sensitivity to immediate interlocutive context.