Code-Switching as a Social Act: The Case of Arabic Wikipedia Talk Pages

Michael Miller Yoder, Shruti Rijhwani, Carolyn Penstein Rosé, Lori Levin
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA
{yoder,srijhwan,cprose,lsl}@cs.cmu.edu

Abstract

Code-switching has been found to have social motivations in addition to syntactic constraints. In this work, we explore the social effect of code-switching in an online community. We present a task from the Arabic Wikipedia to capture language choice, in this case code-switching between Arabic and other languages, as a predictor of social influence in collaborative editing. We find that code-switching is positively associated with Wikipedia editor success, particularly borrowing technical language on pages with topics less directly related to Arabic-speaking regions.

1 Introduction

Code-switching, mixing words from multiple languages in conversation, is common in multilingual communities. This phenomenon has been studied by linguists for nearly half a century (Auer, 2013), and syntactic models of code-switching are still in development (Gardner-Chloros, 2009).

Alternating between languages can also be considered a conversational act with communicative function (Auer, 2013). Code-switching has been found to convey social and interactional meaning in a variety of contexts (Alvarez-Cáccamo, 1990; Blom and Gumperz, 1972; Bassiouney, 2006), though its role in online communities has largely been unexplored. Studying the relationship between social variables and code-switching (CS) can give insight into the role of CS as a pragmatic tool of multilingual speakers.

We offer a quantitative look at how CS functions as a sociolinguistic choice in the editing community around the Arabic Wikipedia, an online encyclopedia which anyone can edit. Our focus is on *talk pages*, where Wikipedia editors discuss article improvements, coordinate work and resolve disagreements on the content they edit (Ferschke, 2014). Relationships between linguistic and social meanings are indirect and difficult to operationalize (Nguyen et al., 2016; Ochs, 1992), but Wikipedia offers an opportunity to quantify social influence in the collaborative task of editing articles. We use code-switching features from editors' talk page contributions to predict the proportion of those users' edits that have lasting impact on the article, a measure of social influence.

We formulate three hypotheses about the social effect of CS on Arabic Wikipedia talk pages. Though other hypotheses are possible, these three are motivated by the sociolinguistic concept of *markedness* (Myers-Scotton, 1998), which attaches social meaning to talk that deviates from conversational expectations. We use markedness as a theoretical lens to assess community norms and social value placed on language choices on Arabic Wikipedia talk pages.

Hypothesis 1. Code-switching may function without clear social meaning (Auer, 2013) and simply be the accepted norm on Arabic Wikipedia talk pages. This could mean that users do not especially notice code-switching or that it is noticed but has no clear effect.

Hypothesis 2. Code-switching marks a Wikipedia user as an outsider who does not follow the Arabic conversational norm (Myers-Scotton, 1998). Code-switching has a negative effect on an editor's acceptance.

Hypothesis 3. Languages other than Arabic, such as English, may carry some sort of value in certain settings (Safi-Stagni, 1991). Code-switching could demonstrate a level of expertise or world knowledge and have a positive effect on the acceptance of an editor's contributions.

Proceedings of the Second Workshop on Natural Language Processing and Computational Social Science, pages 73–82,
Vancouver, Canada, August 3, 2017. ©2017 Association for Computational Linguistics

To determine which of these hypotheses is a more likely explanation for CS in this context, we construct a publicly released dataset that pairs discussion between Wikipedia editors with a measure of editor success in article edits.

We find a positive correlation between the presence of CS in the discussion and editor success, which supports Hypothesis 3. CS features also improve a linear regression model over a reasonable unigram baseline in predicting editor success.

An analysis of an annotated sample of our dataset suggests the possible value the Arabic Wikipedia editing community places on CS for technical language on articles unrelated to Arabic history, people, and culture.

2 Related Work

Code-switching was first linguistically studied to find systems of syntactic and morphological constraints on its use. Myers-Scotton (1995) proposed a CS framework in which grammatical structure is supplied by a dominant "matrix" language, while content morphemes can be drawn from an "embedded" language (Bassiouney, 2009). In contrast, MacSwan (2000) argues against the existence of nearly any universal syntactic constraints on CS.

Sociolinguists take interest in CS as a property of language related to social interaction. Gumperz (1982) proposes a distinction between CS based on factors internal to a conversation and on connotations a language carries across contexts.

We frame our understanding of the social effect of CS on *markedness theory*, which posits that *marked* linguistic choices deviate from understood norms for speakers in certain situations and thus carry social significance (Myers-Scotton, 1998). This emphasis on conversational norms is rooted in Grice's maxims, which give guidelines for expectations in conversation and a framework for social meaning attached to deviations from those norms (Grice, 1975). Note that we are not attempting to prove or disprove markedness theory or Grice's maxims, but instead are using them to understand meaning in interaction and to more fully explain natural language data.

We assume a community norm of Arabic on the Arabic Wikipedia and expect CS to be marked and have some sort of social effect. However, Myers-Scotton (1998) allows the possibility of contexts where CS is itself unmarked; this would also be possible in our case.

Recent computational analyses of style, metaphor, framing and politeness have investigated how language is used to achieve social goals in online communities (Danescu-Niculescu-Mizil et al., 2012, 2013; Jang et al., 2016; Tsur et al., 2015). We examine CS in a similar fashion. Interactional, discourse-level features are context-specific, and the relationship between social and linguistic features is fluid and often difficult to computationalize (Nguyen et al., 2016). Code-switching may not carry clear social meaning at all in a given context (Auer, 2013), much less a predictable signal. Our work enters this conversation by exploring the effect of code-switching on social influence in an online community.

The NLP community has largely studied code-switching apart from its social context. Much work has focused on word-level CS language identification, encouraged by shared tasks (Solorio et al., 2014; Molina et al., 2016). Others have worked to predict code-switch points from preceding text. Solorio and Liu (2008) predict code-switch points with features including the previous *n*-grams' identified language, POS tag, and location in constituent parses in both languages. Piergallini et al. (2016) tackle the same task in combination with language identification on a Swahili-English online forum dataset. They note the possibility of using discourse structure and social variables for predicting code-switch points.

Interest in computational models of the social and pragmatic nature of code-switching is growing. Begum et al. (2016) present an annotation scheme for the pragmatic functions of Hindi-English code-switched tweets, which includes reinforcement, sarcasm, reported speech, and changes from narration to evaluation. Rudra et al. (2016) study language preference for the expression of sentiment among Hindi-English multilinguals, finding that speakers more commonly use Hindi to express negative sentiment and English for positive sentiment on Twitter.

3 Code-Switching on Arabic Wikipedia Talk Pages

Though many language Wikipedias contain code-switching on their talk pages, we select the Arabic Wikipedia for the variation we observe and previous Arabic CS work in NLP (Solorio et al., 2014; Elfardy et al., 2014).

Talk page	Text	English translation
GNU/Linux	سلاسل المحارف and it has a multi-threaded fs.	... a string used, and it has a multi-threaded fs.
Oran, Algeria	Salam, Les missions principales du centre sont: la recherche...	Greetings, the main missions of the center are: research...
Said Aouita	hafid hassan ana fakhour الصفحة الهدف b3outa	Target page [name] I am proud to...
Lebanon	Sorry for talking english I notice you use the image...	Sorry for talking english I notice you use the image...

Table 1: Observations of non-Arabic text in Arabic Wikipedia talk pages

Terms and definitions for code-switching and code-mixing across studies vary considerably (Gardner-Chloros, 2009). Since we are interested in all deviation from the likely norm of Arabic, we accept any instances of switching between languages in a conversation as code-switching. We also include "script-switching", since we assume most editors can use Arabic characters and there may be social significance attached to writing Arabic in Latin script (something called Arabizi), especially since such language is usually dialectal (Darwish, 2014).

Table 1 presents a few motivating examples of language variety in Arabic Wikipedia talk pages. Most CS we see is Arabic-English, but there are examples of French and Arabizi, the romanized Arabic seen in the third example in Table 1.

We also note apologies for using English, including a longer exchange on the *Israel* talk page where an editor is confronted about language choice and replies in Arabizi:

> Editor 1: ...downright erasure of Jewish history in Israel. I don't have an arabic keyboard so i can't type in arabic
>
> Editor 2: you dont seem to be able to read arabic, or you havent read the article and the history section!!
>
> Editor 1: wala ya habibi? maa ta'mil assumptions, ana bahki arabi,wa baqrah arabi (trans. *Hey, don't make assumptions, I speak Arabic, and read Arabic*)

This example suggests that choice of language explicitly matters in some Wikipedia talk page contexts. Editor 1 feels compelled to explain why they are not typing in Arabic, an acknowledgment of the community norm of offering contributions in Arabic. In the second speaker's reply, not using Arabic is leveled as grounds for not being a responsible editor. If Editor 1 is not successful, this interaction suggests Hypothesis 2, where not using Arabic negatively marks an editor as an outsider. Editor 1's response in Arabizi is another language choice with social implications, especially that it is in Levantine Arabic dialect and not in Modern Standard Arabic like the article.

Does this demonstrate enough knowledge of Arabic for status as a contributor? What social effect does writing in English on the talk page have when Arabic is an assumed norm? What effects do other multilingual choices have in other contexts? These questions motivate our study.

4 Data and Task

To capture the social effect of code-switching, we choose a task predicting social influence from CS features in discussion. In the context of the Arabic Wikipedia, we measure social success by the proportion of a Wikipedia user's edits that remain in the article's content after a discussion ends (Priedhorsky et al., 2007) and hypothesize that CS may be associated with this measure.

To set up this task, we pair discussions containing CS from Arabic to other languages with simultaneous article edits, which we use to define individual editor success. Our dataset[1] consists of 5259 instances in which an editor interacts with other editors in a talk page discussion thread and achieves some degree of influence on the associated article page. Statistics for our dataset can be seen in Table 2; a more detailed description of the dataset construction follows.

[1] https://github.com/michaelmilleryoder/wikipedia-codeswitching-data

Number of editor-thread pairs (instances)	5259
Number of code-switching instances	786
Number of discussion threads	2103
Number of talk pages	1031
Number of editors	917

Table 2: Code-switching discussion dataset

4.1 Dataset Construction

Each Wikipedia article has an associated talk page, though many are empty. We begin with all talk pages and article revisions (versions) in the Arabic Wikipedia from a 10 October 2016 data dump.

We use the Java Wikipedia Library (Ferschke et al., 2011) to remove much of the Mediawiki markup on article revisions, and segment the talk pages into posts using talk page revision history and paragraph breaks. Posts under the same heading are organized into discussion *threads*.

There must be sufficient interaction on a talk page thread to measure social effect, so we remove threads with only one participant. To identify CS, we further restrict threads to contain at least one post with at least 3 words with all Latin characters. This filtering leaves 2103 threads remaining out of the original 10,116 (20.8%). Note that the majority of text within these threads are in Arabic, but at least one post within the thread has CS.

In our dataset, we organize each instance as a specific editor's concatenated text in the entire thread (all their posts), along with the combination of all other editors' text as separate features.

4.2 Language Identification

We find a diversity of language on Arabic Wikipedia talk pages not written in the Arabic script, including English, French, Hebrew, Turkish, Chinese and even a few words written in the Tifinagh and Syriac scripts.

To initially survey the distribution of languages, we run all spans of tokens without Arabic characters (and that are not wholly punctuation) through `langid.py` (Lui and Baldwin, 2012), a language identification tool that can detect 97 languages. It is trained in a supervised fashion with Naive Bayes on byte n-grams, using cross-domain training data. `langid.py` finds 66 languages present within the dataset, but a qualitative analysis finds that named entities and noise in the dataset (special characters, usernames that passed through our preprocessing, and Wikipedia-specific material) confuse the language identifier.

This qualitative analysis and our later annotation of a sample finds that the vast majority (estimated 94%) of CS is to English, with some scattered French, Hebrew and other languages.

4.3 Editor Success Scores

Following the example of Priedhorsky et al. (2007), we assess the impact of editors based on the longevity of the edits they make. We define a success score s for each editor in a specific discussion. This score is the proportion of their edits–words deleted and words added–that remain 1 day after the discussion ends. Note that this score only reflects changes in word frequencies, and does not take word re-ordering into account.

Formally, we consider each edit $\mathbf{e}$ as a vector of word frequency changes, both positive (additions) and negative (deletions) for each word type. For an example in English, an edit that changed one instance of *suggested* to *insinuated*, as well as adding *old* might be represented as a set {'suggested': -1, 'insinuated': +1, 'old': +1'}. Let vector $\mathbf{c}$ be the changes in word frequencies from that edit to the final revision in the session. This change vector represents how many tokens that an editor deleted were put back and how many tokens the editor added were afterward deleted. Let $||\mathbf{e}||$ be the sum of the absolute values of word frequency changes of the edit and $||\mathbf{c}||$ be the sum of the absolute values of word frequency changes from the edit to the final revision. The score s of a particular Wikipedia editor u in thread t across edits $\{\mathbf{e}_1, \mathbf{e}_2, ..., \mathbf{e}_n\}$ made by that editor in that thread is:

$$s(u,t) = 1 - \frac{\sum_{i=1}^{n} ||\mathbf{c}_i||}{\sum_{i=1}^{n} ||\mathbf{e}_i||}$$

Each editor's score is the proportion of tokens they changed that remain changed, so $s \in [0, 1]$.

In a qualitative evaluation, this editor score formulation was found to accurately reflect an editor's impact on the revision of the article after the discussion.

5 Experiments and Results

Our goal is capturing the relationship between CS on talk pages and the success of editors on article pages. We consider the presence of CS in an editor's text, as well as other CS features to study the variation among types of CS (section 5.1).

We evaluate the effect of CS features on editor score in two ways. We first evaluate the association between CS and editor success with statistical measures (section 5.2). Then, we test the strength of this association by using CS features in a predictive model of editor success (section 5.3).

5.1 Features

We select code-switching features that we expect to vary in deviation from a community expectation of Arabic, a concept motivated by markedness theory (Myers-Scotton, 1998). Each datapoint separates the text contributed by one specific editor in a thread from all other text in the thread, and features (listed below) are extracted from both the editor's and all other editors' text. We examine Latin characters in particular since non-Latin and non-Arabic scripts are negligible in the corpus, and restricting to Latin characters reduces noise from nonlinguistic symbols and rare punctuation that otherwise are detected.

- **Presence of CS:** whether the text contains non-Arabic content, operationalized as three or more tokens longer than one character in all Latin characters.

- **Proportion of non-Arabic words:** the proportion of non-Arabic content, operationalized as the proportion of words in all Latin characters.

- **Proportion of code-switch points.** To capture how frequently an editor switches languages, each word boundary is counted as a potential code-switch point from Arabic to another language or vice versa. This feature is the number of actual switch points between languages, normalized by the number of word boundaries.

- **Presence of CS and quotes.** We naively capture quoting in non-Arabic languages by determining if there are more than three words in all Latin characters and two double-quotation (`"`) marks.

- **Proportion of non-Arabic named entities.** Named entities written in scripts other than Arabic are quite frequent in our dataset and may carry less social significance than other types of CS. We operationalize this feature as the proportion of words in all Latin characters that are capitalized.

- **Apologies.** We are particularly interested in apologizing for using a language other than Arabic, as this recognizes deviation from an Arabic community norm. We naively assume that any apology is likely to be about language use, and so extract use of the word *sorry* or any version of the lemma *apolog*. However, there are too few examples of this feature even in English, the most frequent non-Arabic language used, to meaningfully compare its relation to editor score.

- **Presence of specific languages.** We extract separate features for the presence of specific languages automatically identified with `langid.py` (see section 4.2), as well as the proportion of all words that are identified as that specific language. Most likely due to noise in automatic identification and the overwhelming presence of English, these features do not improve regression performance or relate in statistically significant ways to editor success, so we do not consider them further.

We also separately consider unigrams longer than 1 letter that are completely in Arabic script or completely in Latin characters.

As nonlinguistic features, we include the number of editor turns and other turns. Both were found to have very weak negative correlation with editor success and were not considered further.

Note that named entities and full sentences in non-Arabic characters are included in our CS features. Since we want to explore as many possible effects as possible, our aim is to broadly capture any use of terms outside of an assumed Arabic norm. Thus our definition of "code-switching" is loose, including what may simply be considered borrowing words or writing a talk page post all in one language in a conversation that includes multiple languages.

5.2 Statistical Evaluation

In order to evaluate the relationship between CS and social influence, we use statistical tests of association between CS features and the editor success score. For binary features, we simply measure the difference in editor score means between instances for which a feature is `TRUE` and instances where a features is `FALSE`. For continuous features, we measure the correlation between that feature and the editor score.

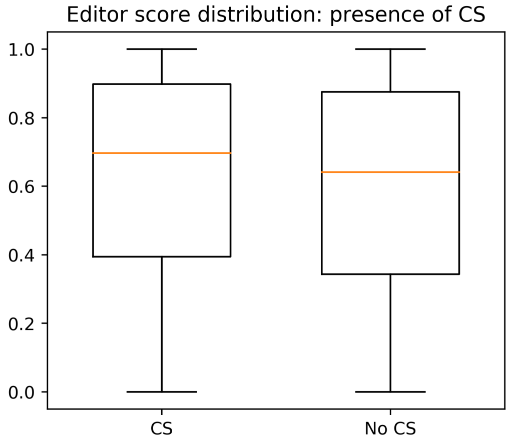

Figure 1: Editor score distributions of instances with and without CS in the editor's text. The difference between means is significant $p < 0.01$.

We find a positive association between the presence of CS and editor success. The presence of CS has a significantly positive effect on editor score, a mean score of 0.628 with CS and 0.593 without ($p < 0.01$ using student's t-test). Distributions for the presence of CS in editor text are in Figure 1. Hypothesis 1, the possibility of no social influence, is unlikely given this statistical evidence of effect on editor success, and instead, Hypothesis 3's claim of a positive social effect is supported.

The presence of CS with quotes also has a marginally significant positive effect on editor score. The mean score of instances with CS and quotes was 0.637 and 0.596 without ($p \approx 0.03$).

The strongest correlation among continuous features is the proportion of switches, which still only weakly correlates with editor success, $r = 0.058$ ($p < 0.0001$).

5.3 Predictive Modeling

We also use editor score as an outcome variable for a linear regression classifier, which we evaluate using 10-fold cross-validation in scikit-learn (Pedregosa et al., 2011). Support vector machine regression yields similar trends.

CS features are more predictive of editor scores than unigrams with feature selection and tf-idf weight (1000 features, selected by mutual information). Results for the classifier are described in Table 3, reporting root mean squared error.

Performance decreases with unigrams and CS features from the text of discussion participants other than the scored editor (*editor+other* vs. *editor-only*), so only the editor's code-switching has an effect.

Feature set	LinReg
Editor-only	
unigrams	0.350
Arabic unigrams	0.350
Latin unigrams	0.319*
CS	**0.315***
unigrams+CS	0.349
Editor+others	
unigrams	0.341
CS	**0.315***
unigrams+CS	0.341

Table 3: RMSE in editor success prediction. Unigrams are restricted with feature selection to 1000. Scores marked with an * are significantly different ($p < 0.01$) from editor-only unigrams. CS are code-switching features. *Editor+others* includes features from the scored editor and others in the discussion thread.

The most informative CS feature for the linear regression classifier is the proportion of code-switch points, while the CS features are included in the top 10 most informative features for the unigrams+CS feature set.

In a further experiment, we aggressively select unigram features with tf-idf weight based on mutual information down to just 10. This restricted group of unigram features reaches the prediction performance of CS features (RMSE of 0.315). However, the unigram features are difficult to interpret; it is unclear why they index social influence (Table 4). The focus of this paper is to evaluate the relationship of CS to a measure of social influence; we leave model development toward better prediction performance to future work.

We also examine the effect of Arabic unigrams (top 1000 features selected) and Latin unigrams (no feature selection). The performance of Arabic unigrams matches that of all unigrams, but Latin unigrams perform significantly better. This reinforces language choice as relevant to social influence in this context.

6 Discussion

The social influence of CS may depend on context, and we examine different types of CS and variation in article topic as reasonable influencing factors. We randomly sample 100 instances of the

Arabic word	English gloss
عدة	several
من	from
بع	after
هذه	this
على	on it
في	in
غير	but
أو	or
أن	that
لها	to it (fem.)

Table 4: Top unigram features. Linear regression with these 10 features (after tf-idf feature selection using mutual information) reaches the performance of CS features, but these features are much less interpretable.

CS type	**%**	**Editor success score (mean)**
All	*100*	*0.631*
Named entities	36	0.539
Technical	26	0.818
Single words	9	0.714
Phrases	8	0.323
Challenges	7	0.724
Quotations	6	0.394
Translations	2	0.825
Other	6	0.722

Table 5: CS types distribution in our annotated set

data that contain non-Arabic words for manual annotation of CS type and article topic.

6.1 CS Type

The annotator (one of the authors) noted the language of CS as well as the possible reasons why CS was used in those instances, using the annotation framework by Begum et al. (2016) as a reference for structural and semantic functions of CS.

The distribution of these *CS types* are listed in Table 5. Most prominently, the dataset contains a significant percentage of instances with **named entities** written in non-Arabic script. These instances include both Western and Arabic names. For example, *Howard Stern* or *Ibn an-Nafīs*. It is interesting that several named entities are written in Latin script within a large conversation in Arabic, even though names are often freely transliterated over scripts. This could be because certain names are more familiar in their Latin form (like *CNN*).

Article type	**Editor success score (mean)**
Technical	0.796*
Non-technical	0.553
Arabic	0.537
Non-Arabic	0.747*

Table 6: Mean editor success scores across article topics. * indicates significance $p < 0.01$

Using **English technical terms** is also commonly seen when Wikipedia articles of a technical nature are discussed (*Cytoplasm* and *vertebrates*, for instance). These are examples of topic-related CS (Barredo, 1997; Begum et al., 2016). Such code-switched technical words are likely used when there is no commonly used Arabic equivalent. We see a high mean editor success score when technical terms are code-switched. Most of the instances in which this CS type occurred were threads about articles not specific to Arabic-speaking regions and came from topics like science or world history. The strong editor success is in support of Hypothesis 3, which suggests that deflection from the norm of Arabic might be useful in particular scenarios, non-Arabic-specific technical topics in this case.

We also see instances of the **quotation function** and the **translation function** of CS (Begum et al., 2016). The former occurs primarily when the discussion involves quoting parts of the corresponding English Wikipedia page or relevant English news articles and the latter translates Arabic words and phrases to another language. In the Wikipedia context, these functions likely serve to ease explanation of article content edits, and complement the discussion which is predominantly in Arabic.

More specific to Wikipedia is the **challenge** CS type. These are instances where, within Arabic text, phrases in non-Arabic languages are used to debate or contest the content edits being discussed. For example, *there may be some errors that need to be addressed* and *the image is wrong*. Some of these instances are in the Narrative-Evaluative form of CS, which contains a language-switch between stating the fact (the suggested content edit in our case) and an opinion about the fact (Begum et al., 2016).

Apart from these types, CS with other single

Talk page	Text	English translation	Editor outcome
Endorphin	The physiological importance of the beta-endorphin ...	The physiological importance of the beta-endorphin ...	successful
Cybernetics	في ال open loop نعطي النظام القيمة...	In the open loop, we give the system the value...	successful
Egypt	وال دي ان اى هو ما لخصه الدكتور كيتا... wrote that “There is no scientific reason...”	the DNA is summed up by Dr. Keita, who wrote that "There is no scientific reason... "	unsuccessful
Yazidism	وعدم تقبله ان “Malak Ta’us has often been identified by outsiders with the Judeo-Christian figure of Satan”	not accepting that “Malak Ta’us has often been identified by outsiders with the Judeo-Christian figure of Satan”	unsuccessful

Table 7: Code-switching examples from effective and ineffective editors

non-Arabic words and phrases account for around 16% of the annotated sample. These generally consist of common English words like *had been good* and *sorry*, similar to the tag-switching structural form (Begum et al., 2016).

Although CS with English is far more prominent than other languages (94% of the instances), we also see French, Hebrew and Arabizi used in the dataset. The ‘Other’ instances in Table 5 refer to CS that did not have an interpretable function (Wikipedia-specific terms, for instance).

6.2 Article Topic

We used DBpedia (Lehmann et al., 2014) to get Wikipedia categories for each article. For our selected sample of 100 instances, the annotator verified these categories and judged whether the article was of a technical subject or not, as well as whether the article was centered around content from Arabic-speaking regions. Articles on general topics or topics not specifically related to Arabic history, language and culture were annotated as ‘non-Arabic’.

CS on pages about non-Arabic topics is on average much more successful than on Arabic-related topics (Table 6). CS on pages with a technical subject is also more successful on average than on pages with other topics.

These findings are supported by a qualitative analysis of example Arabic-English discussion contributions with CS. Using medical terms in English on talk pages for articles on *Endorphin* and *Cancer* was associated with success, as was using English technical terms on the talk page for *Cybernetics* (see Table 7).

However, unsuccessful editors who switch to English seem to do so on pages whose subjects are more directly related to Western Asian and North African culture. For example, we find unsuccessful CS on the page about *Yazdanism*, a religion indigenous to Mesopotamia and on the *Egypt* page about the ancestry of the Egyptian population (see Table 7). Hypothesis 2’s claim of CS as an ‘outsider’ effect may be supported in these contexts.

7 Conclusion and Future Work

We present a task and dataset to study the social effect of CS in the context of an online collaborative community, as well as an analysis of how sociolinguistic theory about deviation from conversational norms in CS can explain this data. We find that CS on Arabic Wikipedia talk pages is associated with making successful article edits, a measure of social influence. This finding supports a social interpretation of CS as a positive marker in this community, especially when the subject matter is technical or relates to non-Arabic topics.

Hypothesis 3 is most clearly supported by the positive association of CS with editor influence. Hypothesis 1, the lack of relationship between CS and social meaning, is unlikely given the effects we see on social influence. Hypothesis 2, a negative evaluation of CS as deviating from an Arabic norm, could explain the effect of CS in some contexts we observe, such as pages with topics related to Arabic culture.

In future work, norms specific to pages, users, languages and topics could be quantitatively explored and could nuance our measures of the markedness of editor contributions from those norms. Our dataset could also be used to analyze other factors contributing to editor success, such as speech acts, politeness, or conversational roles.

Further, this framework could easily be expanded to a broader multi-lingual analysis across Wikipedias of different languages, or even dialectal analysis within the Arabic Wikipedia. Different community norms about language choice on talk pages could yield different correlations with social influence.

Acknowledgments

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship program under Grant No. DGE-1252522, as well as by NSF Grant No. IIS-1546393. This work was also partially supported by the Defense Advanced Research Projects Agency (DARPA) Information Innovation Office (I2O) under the Low Resource Languages for Emergent Incidents (LORELEI) program issued by DARPA/I2O under Contract No. HR0011-15-C-0114. The views expressed are those of the authors and do not reflect the official policies or positions of the National Science Foundation, Department of Defense or the U.S. Government.

References

Celso Alvarez-Cáccamo. 1990. Rethinking conversational code-switching: Codes, speech varieties, and contextualization. In *Annual Meeting of the Berkeley Linguistics Society*. volume 16, pages 3–16.

Peter Auer. 2013. *Code-Switching in Conversation: Language, Interaction and Identity*. Taylor & Francis.

Inma Munoa Barredo. 1997. Pragmatic functions of code-switching among basque-spanish bilinguals. *Retrieved on October* 26:2011.

Reem Bassiouney. 2006. *Functions of code switching in Egypt: Evidence from monologues*, volume 46. Brill.

Reem Bassiouney. 2009. *Arabic Sociolinguistics*. Edinburgh University Press.

Rafiya Begum, Kalika Bali, Monojit Choudhury, Koustav Rudra, and Niloy Ganguly. 2016. Functions of Code-Switching in Tweets: An Annotation Scheme and Some Initial Experiments. In *LREC*. i, pages 1644–1650.

Jan-Petter Blom and John J. Gumperz. 1972. Social meaning in linguistic structures: code-switching in Northern Norway. *Directions in Sociolinguistics: The Ethnography of Communication* pages 407–434.

Cristian Danescu-Niculescu-Mizil, Lillian Lee, Bo Pang, and Jon Kleinberg. 2012. Echoes of power: Language effects and power differences in social interaction. *Proceedings of the 21st international conference on World Wide Web - WWW '12* page 699. https://doi.org/10.1145/2187836.2187931.

Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. A computational approach to politeness with application to social factors. *The 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013)* .

Kareem Darwish. 2014. Arabizi Detection and Conversion to Arabic. In *ANLP 2014*.

Heba Elfardy, Mohamed Al-Badrashiny, and Mona Diab. 2014. Aida: Identifying code switching in informal arabic text. *EMNLP 2014* page 94.

Oliver Ferschke. 2014. *The Quality of Content in Open Online Collaboration Platforms: Approaches to NLP-supported Information Quality Management in Wikipedia*. Ph.D. thesis, Technische Universität, Darmstadt.

Oliver Ferschke, Torsten Zesch, and Iryna Gurevych. 2011. Wikipedia revision toolkit: Efficiently accessing wikipedia's edit history. In *Proceedings of the ACL-HLT 2011 System Demonstrations*. Association for Computational Linguistics, Portland, Oregon, pages 97–102.

Penelope Gardner-Chloros. 2009. *Code-switching*. Cambridge University Press.

H. Paul Grice. 1975. Logic and conversation. In Peter Cole and Jerry L. Morgan, editors, *Syntax and Semantics: Speech Acts*, Academic Press, New York, pages 41–58.

John J. Gumperz. 1982. *Discourse Strategies*. Studies in Interactional Socio. Cambridge University Press.

Hyeju Jang, Yohan Jo, Qinlan Shen, Michael Miller, Seungwhan Moon, and Carolyn Rose. 2016. Metaphor Detection with Topic Transition, Emotion and Cognition in Context. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* 1:216–225.

Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer, and Chris Bizer. 2014. DBpedia - a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web Journal* .

Marco Lui and Timothy Baldwin. 2012. langid.py: An off-the-shelf language identification tool. *Proceedings of the ACL 2012 System Demonstrations* (July):25–30.

Jeff MacSwan. 2000. The architecture of the bilingual language faculty: evidence from intrasentential code switching. *Bilingualism: Language and Cognition* 3(1):37–54. https://doi.org/10.1017/S1366728900000122.

Giovanni Molina, Rey-Villamizar, Thamar Solorio, Fahad AlGhamdi, Mahmoud Gohneim, Abdelati Hawwari, and Mona Diab. 2016. Overview for the Second Shared Task on Language Identification in Code-Switched Data. *Proceedings of The Second Workshop on Computational Approaches to Code Switching, held in conjunction with EMNLP 2016.* pages 62–72.

Carol Myers-Scotton. 1995. *Social Motivations for Codeswitching: Evidence from Africa*. Oxford studies in language contact. Clarendon Press.

Carol Myers-Scotton. 1998. *Codes and Consequences: Choosing Linguistic Varieties*. Oxford University Press.

Dong Nguyen, A. Seza Doğruöz, Carolyn P. Rosé, and Franciska de Jong. 2016. Computational sociolinguistics: A survey. *Computational Linguistics* 42(3):537–593. https://doi.org/10.1016/j.jksus.2015.08.001.

Elinor Ochs. 1992. Indexing Gender. In Alessandro Duranti and Charles Goodwin, editors, *Rethinking context: Language as an interactive phenomenon*, Cambridge University Press, chapter 14, pages 335–358.

Fabian Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12:2825–2830.

Mario Piergallini, Rouzbeh Shirvani, Gauri S Gautam, and Mohamed Chouikha. 2016. Word-Level Language Identification and Predicting Codeswitching Points in Swahili-English Language Data pages 21–29.

Reid Priedhorsky, Jilin Chen, Shyong Tony K Lam, Katherine Panciera, Loren Terveen, and John Riedl. 2007. Creating, destroying, and restoring value in wikipedia. *Proceedings of the 2007 international ACM conference on supporting group work - GROUP '07* page 259. https://doi.org/10.1145/1316624.1316663.

Koustav Rudra, Shruti Rijhwani, Rafiya Begum, Kalika Bali, Monojit Choudhury, and Niloy Ganguly. 2016. Understanding Language Preference for Expression of Opinion and Sentiment: What do Hindi-English Speakers do on Twitter? pages 1131–1141.

Sabah Safi-Stagni. 1991. *Agrammatism in Arabic*. Perspectives on Arabic Linguistics. John Benjamins Publishing Company.

Thamar Solorio, Elizabeth Blair, Suraj Maharjan, Steven Bethard, Mona Diab, Mahmoud Gohneim, Abdelati Hawwari, Fahad AlGhamdi, Julia Hirschberg, Alison Chang, and Pascale Fung. 2014. Overview for the First Shared Task on Language Identification in Code-Switched Data. *Proceedings of The First Workshop on Computational Approaches to Code Switching, held in conjunction with EMNLP 2014.* pages 62–72.

Thamar Solorio and Yang Liu. 2008. Learning to predict code-switch points. *EMNLP '08 Proceedings of the Conference on Empirical Methods in Natural Language Processing* pages 973–981. https://doi.org/10.16373/j.cnki.ahr.150049.

Oren Tsur, Dan Calacci, and David Lazer. 2015. A Frame of Mind: Using Statistical Models for Detection of Framing and Agenda Setting Campaigns. *ACL* pages 1629–1638.