Carnegie Mellon University

Code-switching on Arabic Wikipedia talk pages

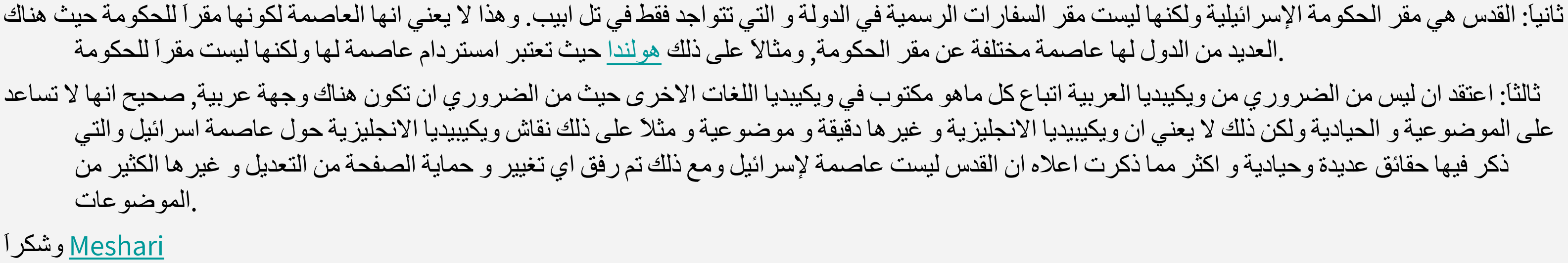
Michael Miller Yoder
23 June 2017

www.cmu.edu

Motivation

Arabic Wikipedia talk pages

Arabic Wikipedia talk pages



ثانياً: القدس هي مقر الحكومة الإسرائيلية ولكنها ليست مقر السفارات الرسمية في الدولة و التي تتواجد فقط في تل ابيب. وهذا لا يعني انها العاصمة لكونها مقراً للحكومة حيث هناك
العديد من الدول لها عاصمة مختلفة عن مقر الحكومة, ومثالاً على ذلك هولندا حيث تعتبر امستردام عاصمة لها ولكنها ليست مقراً للحكومة.

ثالثاً: اعتقد ان ليس من الضروري من ويكيبيديا العربية اتباع كل ماهو مكتوب في ويكيبيديا اللغات الاخرى حيث من الضروري ان تكون هناك وجهة عربية, صحيح انها لا تساعد
على الموضوعية و الحيادية ولكن ذلك لا يعني ان ويكيبيديا الانجليزية و غيرها دقيقة و موضوعية و مثلاً على ذلك نقاش ويكيبيديا الانجليزية حول عاصمة اسرائيل والتي
ذكر فيها حقائق عديدة وحيادية و اكثر مما ذكرت اعلاه ان القدس ليست عاصمة لإسرائيل ومع ذلك تم رفق اي تغيير و حماية الصفحة من التعديل و غيرها الكثير من
الموضوعات.

وشكراً Meshari

Arabic Wikipedia talk pages

Sorry for writing in English. My Arabic is not good enough. It makes no sense to put a map of \Palestine on the whole territory of Mandatorial Palestine, and no map of Israel. THis is certainly not the objective truth

Arabic Wikipedia talk pages

...there is no space for subjectivity and typical arab responses of "occupation" and down-right erasure of Jewish history in Israel. I don't have an arabic keyboard so i can't type in arabic

--Yoav Ben Zakai

you dont seem to be able to read arabic, or you havent read the article and the history section!!

wala ya habibi? maa ta'mil assumptions, ana bahki arabi,wa baqrah arabi..tab hala ana shoo bahki...seeni wala yabani?

--Yoav Ben Zakai

Arabic Wikipedia talk pages

من كتاب "ستة أيام كيف شكلت حرب 1967 الشرق الأوسط" للمؤلف Jeremy Bowen
"The Egyptian army was collapsing so quickly that no one noticed an expeditionary force of 1250 men had been sent by the ruler of Kuwait..."

Arabic Wikipedia talk pages

النص الإنجليزي ؟؟ ما هذا !!سلاسل المحارف د and it has a multi-threaded fs.
It is NOT protable. لا أعرف المعني التقني الدقيق لهذا المصطلح، ولكن ليس متعدد سلاسل المحارف
فالمصطلح يعني threads وليس strings بكل تأكيد! (أعرفها

Arabic Wikipedia talk pages

Portrail Adrar
Je voulais demander votre autorisation pour prendre quelques paragraphes sur le Maroc
ترجمة الرسالة: أرجوك أن تقوم بذلك (و المقصود أقتباس مقالات من موقع أدرار) ، فمبدئي بالأساس هو تعميم المعلومة و المعرفة بالمجان.
Dear Mr. Webmaster of the Portrail Adrar,
I want to ask for your authorization to take some paragraphs about Morocco

Motivating questions

- What are the social motivations (if there are any) for code-switching on Wikipedia talk pages?

- What social effect does code-switching have on Wikipedia talk pages?

Code-switching

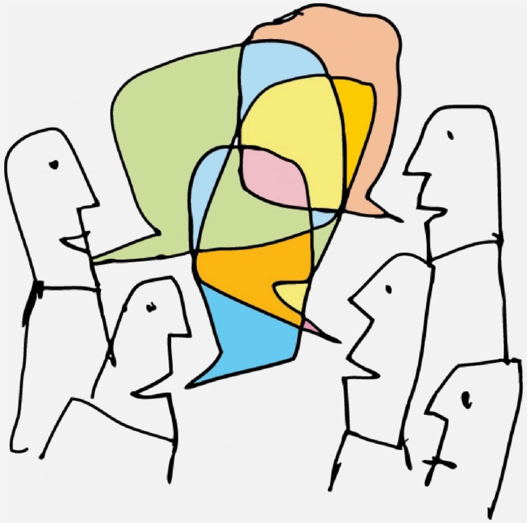
Social motivations

Social motivations

- Different identities, roles, voices (Heller 1988)

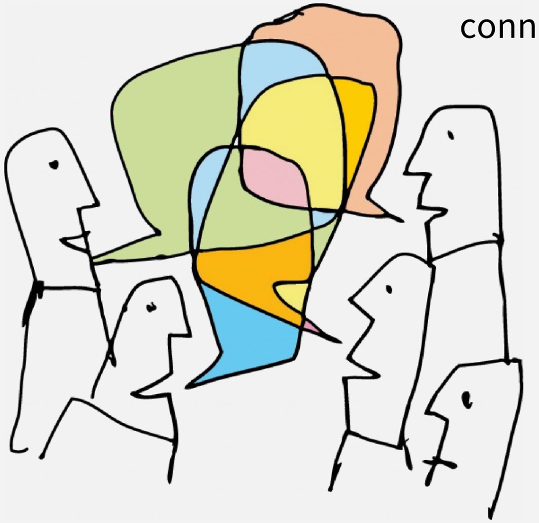
Social motivations

- Different identities, roles, voices (Heller 1988)
- Different factors (Thomason and Kaufmann 1988)



Social motivations

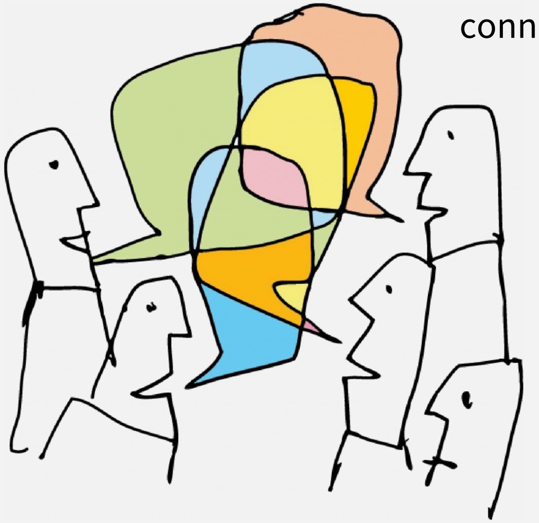
- Different identities, roles, voices (Heller 1988)
- Different factors (Thomason and Kaufmann 1988)



connotations with language

Social motivations

- Different identities, roles, voices (Heller 1988)
- Different factors (Thomason and Kaufmann 1988)

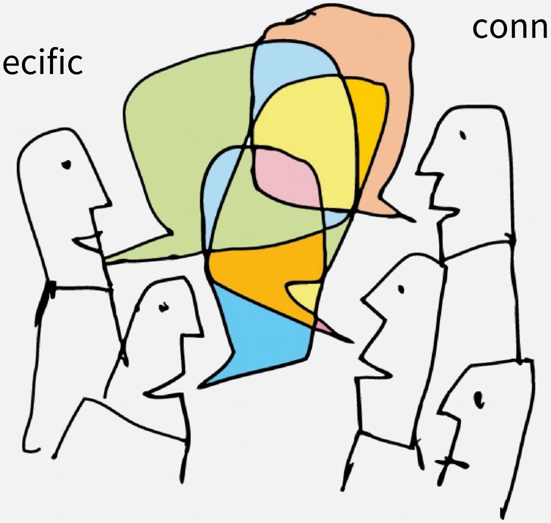


connotations with language

speaker-specific

Social motivations

- Different identities, roles, voices (Heller 1988)
- Different factors (Thomason and Kaufmann 1988)



conversation-specific

connotations with language

speaker-specific

Social motivations

Markedness theory (Myers-Scotton 1998)

Social motivations

Markedness theory (Myers-Scotton 1998)

- Assumes a conversational "norm"

Social motivations

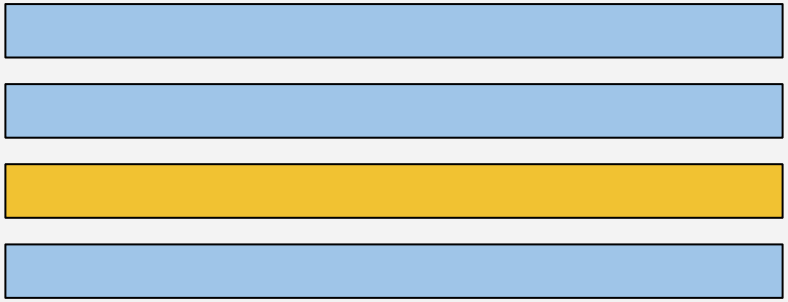
Markedness theory (Myers-Scotton 1998)

- Assumes a conversational "norm"
- Marked usage deviates from the norm

Social motivations

Markedness theory (Myers-Scotton 1998)

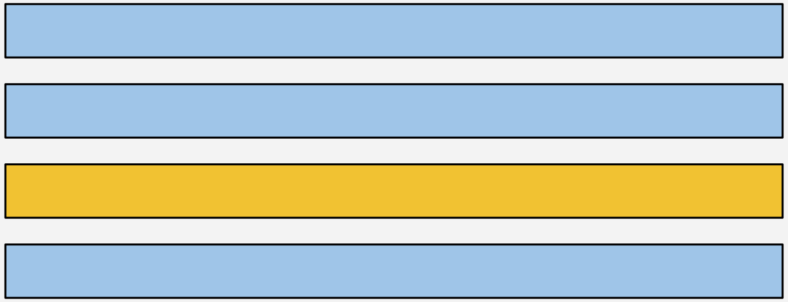
- Assumes a conversational "norm"
- Marked usage deviates from the norm



Social motivations

Markedness theory (Myers-Scotton 1998)

- Assumes a conversational "norm"
- Marked usage deviates from the norm
- Code-switching can be the norm!



Hypotheses

Hypotheses

- Arabic is unmarked
- English marked as *outsider*

Hypotheses

- Arabic is unmarked
- English marked as *outsider*
- Code-switching negative effect

Hypotheses

- Arabic is unmarked
- English marked as *outsider*
- Code-switching negative effect

> *Sorry for my English...*

Hypotheses

- Arabic is unmarked
- English marked as *outsider*
- Code-switching negative effect

> *Sorry for my English...*

Hypotheses

- Arabic is unmarked
- English marked as *outsider*
- Code-switching negative effect

> “*Sorry for my English...*”

- English as a prestige language (still could be marked)

Hypotheses

- Arabic is unmarked
- English marked as *outsider*
- Code-switching negative effect

> *Sorry for my English...*

- English as a prestige language (still could be marked)
- Code-switching positive effect

Hypotheses

- Arabic is unmarked
- English marked as *outsider*
- Code-switching negative effect

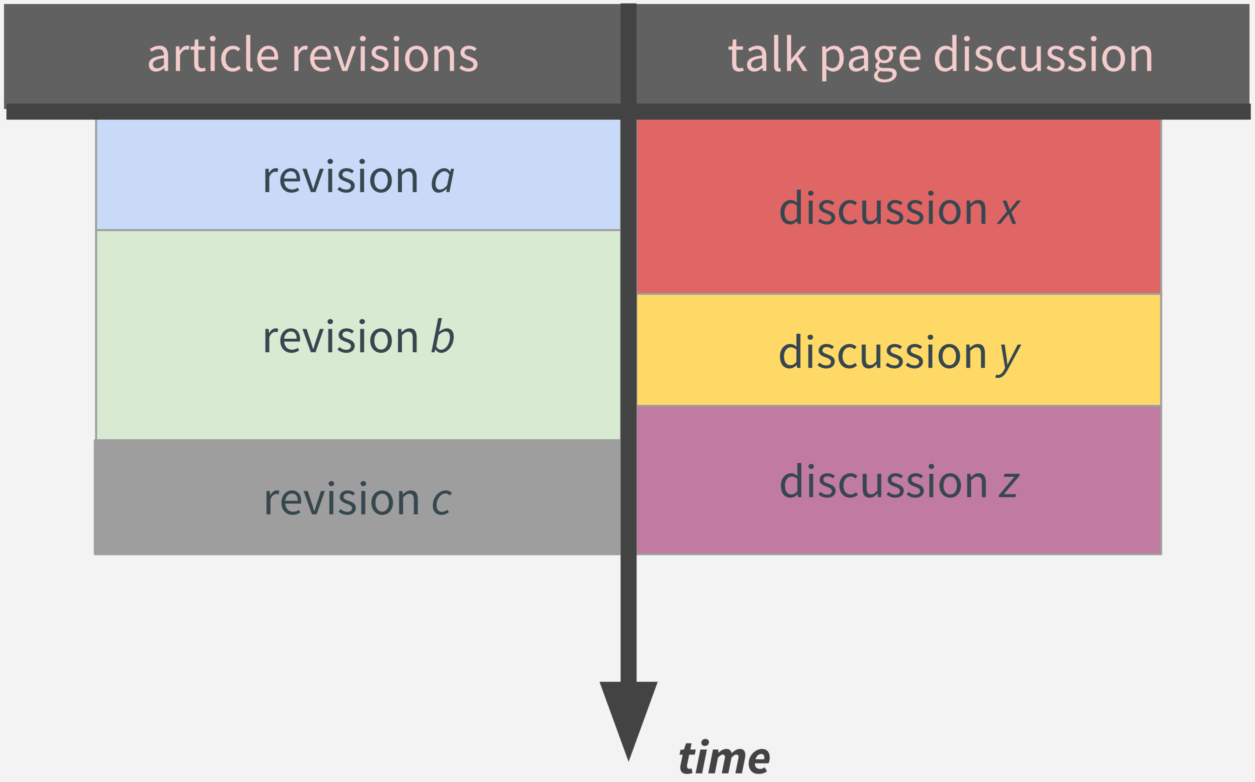
> *Sorry for my English...*

- English as a prestige language (still could be marked)
- Code-switching positive effect

> *Dear Mr. Webmaster...*

Data and task

Arabic Wikipedia corpus



Arabic Wikipedia corpus

- For a conversation to be considered, must have a post with at least 3 words with all Latin characters

# conversations	2103
# editors	917
# editor-conversation pairs (data points)	5259
# data points with CS	786

Social influence

Social influence

- Want an outcome measure that reflects how successful editors are in collaboration

Social influence

- Want an outcome measure that reflects how successful editors are in collaboration: *how successful their article edits are*

Social influence

- Editor influence as proportion of edits that "last" (Priedhorsky 2007)

Social influence

- Editor influence as proportion of edits that "last" (Priedhorsky 2007)
- Example
  - edit vector *e* = { 'old' : +1, 'young': -1, 'people': +2}

Social influence

- Editor influence as proportion of edits that "last" (Priedhorsky 2007)
- Example
  - edit vector *e* = { 'old' : +1, 'young': -1, 'people': +2}
  - Look at difference from "final revision", calculate change vector *c* = {'old': -1, 'people': -1}

Social influence

- Editor influence as proportion of edits that "last" (Priedhorsky 2007)
- Example
  - edit vector *e* = { 'old' : +1, 'young': -1, 'people': +2}
  - Look at difference from "final revision", calculate change vector *c* = {'old': -1, 'people': -1}

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

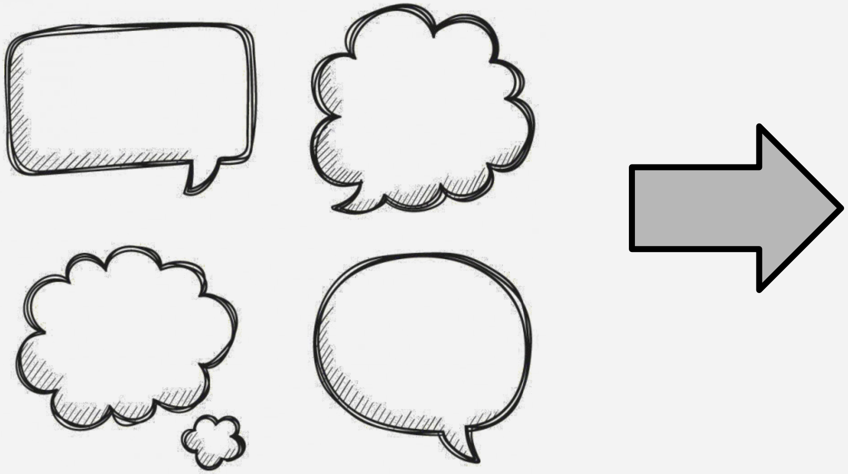
Prediction task

Prediction task

- How does code-switching influence editor success?

Prediction task

- How does code-switching influence editor success?
- Talk page features to predict outcome based on article interaction



$$score(u, t)$$

Code-switching features

Code-switching features

- Presence of code-switching
- Proportion of words in Latin characters

Code-switching features

- Presence of code-switching
- Proportion of words in Latin characters
- Proportion of switches (out of all possible switch-points)
- Presence of quotes and words in Latin characters
- Proportion of named entities in Latin characters
- Apologies
- Presence of specific languages (~94% English)

Preliminary experiments

Statistical analysis

feature	**positive feature mean**	**negative feature mean**	***p***
presence of CS			

Statistical analysis

- Significant positive effect

feature	**positive feature mean**	**negative feature mean**	***p***
presence of CS	.628	.593	.0001

Linear regression

- CS features improve over unigrams

feature set	RMSE
editor unigrams†	.350
editor CS	
editor unigrams + CS	

† with 1000 feature selection

Linear regression

- CS features improve over unigrams

feature set	**RMSE**
editor unigrams†	.350
editor CS	**.315***
editor unigrams + CS	.349

† with 1000 feature selection
* $p < 0.001$

Discussion

Implications

Implications

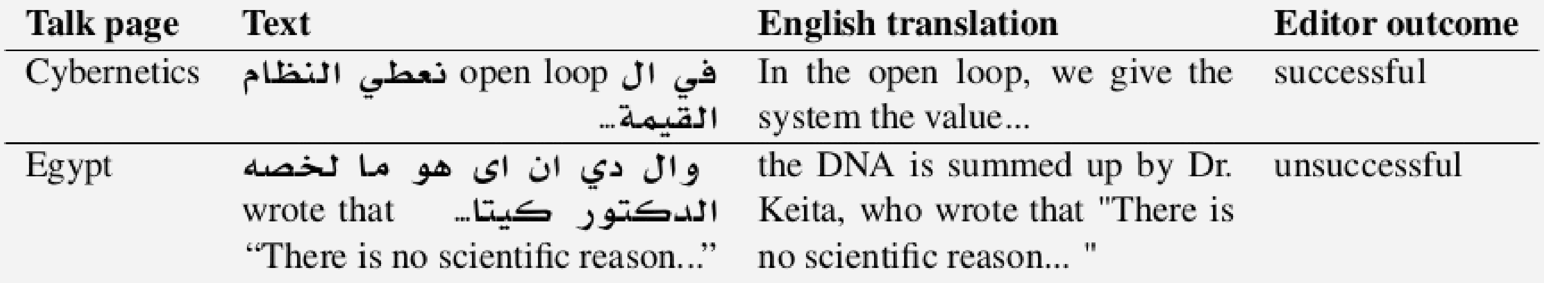
- Positive influence supports hypothesis about possible value of European languages in this context

Implications

- Positive influence supports hypothesis about possible value of European languages in this context, but is that really what's going on here?

Implications

- Positive influence supports hypothesis about possible value of European languages in this context, but is that really what's going on here?
- Qualitative evidence that effectiveness depends on page topic



Talk page	Text	English translation	Editor outcome
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

Annotation

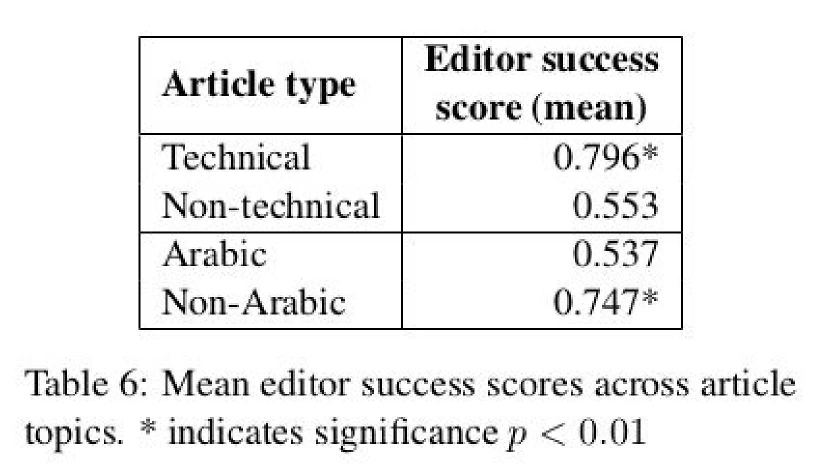
- Drilled down by CS type (Begum et al. 2016) and article topic (from DBPedia)

Annotation

- Drilled down by CS type (Begum et al. 2016) and article topic (from DBPedia)
- Technical CS high success (0.818 over 0.631 avg for CS)

Annotation

- Drilled down by CS type (Begum et al. 2016) and article topic (from DBPedia)
- Technical CS high success (0.818 over 0.631 avg for CS)
- Non-Arabic, technical topics high CS success



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 $p < 0.01$

Michael Miller Yoder, Shruti Rijhwani, Carolyn Penstein Rosé, and Lori Levin. 2017 (in press). Code-Switching as a Social Act: The Case of Arabic Wikipedia Talk Pages. *Proceedings of the 2017 ACL Workshop on Natural Language Processing and Computational Social Science.*

a

Michael Miller Yoder, Shruti Rijhwani, Carolyn Penstein Rosé, and Lori Levin. 2017 (in press). Code-Switching as a Social Act: The Case of Arabic Wikipedia Talk Pages. *Proceedings of the 2017 ACL Workshop on Natural Language Processing and Computational Social Science.*

Thank you!

Google Drive: **shoutkey.com/be**

wikipedia_talk_page_examples.xlsx

visualization: **shoutkey.com/lease**