Predicting Code-Switching in Multilingual Communication for Immigrant Communities

Evangelos E. Papalexakis
Carnegie Mellon University
Pittsburgh, USA
epapalex@cs.cmu.edu

Dong Nguyen
University of Twente
Enschede, The Netherlands
d.nguyen@utwente.nl

A. Seza Doğruöz
Netherlands Institute
for Advanced Study
Wassenaar, The Netherlands
a.s.dogruoz@gmail.com

Abstract

Immigrant communities host multilingual speakers who switch across languages and cultures in their daily communication practices. Although there are in-depth linguistic descriptions of code-switching across different multilingual communication settings, there is a need for automatic prediction of code-switching in large datasets. We use emoticons and multi-word expressions as novel features to predict code-switching in a large online discussion forum for the Turkish-Dutch immigrant community in the Netherlands. Our results indicate that multi-word expressions are powerful features to predict code-switching.

1 Introduction

Multilingualism is the norm rather than an exception in face-to-face and online communication for millions of speakers around the world (Auer and Wei, 2007). 50% of the EU population is bilingual or multilingual (European Comission, 2012). Multilingual speakers in immigrant communities switch across different languages and cultures depending on the social and contextual factors present in the communication environment (Auer, 1988; Myers-Scotton, 2002; Romaine, 1995; Toribio, 2002; Bullock and Toribio, 2009). Example (1) illustrates Turkish-Dutch code-switching in a post about video games in an online discussion forum for the Turkish immigrant community in the Netherlands.

Example (1)
user1: <dutch>vette spelllllllll </dutch>.. <turkish>bir girdimmi cikamiyomm .. yendikce yenesi geliyo insanin</turkish>
Translation: <dutch> awesome gameeeee </dutch>.. <turkish>once you are in it, it is hard to leave .. the more you win, the more you want to win</turkish>

Mixing two or more languages is not a random process. There are in-depth linguistic descriptions of code-switching across different multilingual contexts (Poplack, 1980; Silva-Corvalán, 1994; Owens and Hassan, 2013). Although these studies provide invaluable insights about code-switching from a variety of aspects, there is a growing need for computational analysis of code-switching in large datasets (e.g. social media) where manual analysis is not feasible. In immigrant settings, multilingual/bilingual speakers switch between minority (e.g. Turkish) and majority (e.g. Dutch) languages. Code-switching marks multilingual, multi-cultural (Luna et al., 2008; Grosjean, 2014) and ethnic identities (De Fina, 2007) of the speakers. By predicting code-switching patterns in Turkish-Dutch social media data, we aim to raise consciousness about mixed language communication patterns in immigrant communities. Our study is innovative in the following ways:

- We performed experiments on the longest and largest bilingual dataset analyzed so far.
- We are the first to predict code-switching in social media data which allow us to investigate features such as emoticons.
- We are the first to exploit multi-word expressions to predict code-switching.
- We use automatic language identification at the word level to create our dataset and features that capture previous language choices.

The rest of this paper is structured as follows: we discuss related work on code-switching and multilingualism in Section 2, our dataset in Section 3, a qualitative analysis in Section 4, our experimental setup and features in Section 5, our results in Section 6 and our conclusion in Section 7.

Proceedings of The First Workshop on Computational Approaches to Code Switching, pages 42–50,
October 25, 2014, Doha, Qatar.

2 Related Work

Code-switching in sociolinguistics There is rarely any consensus on the terminology about mixed language use. Wei (1998) considers alternations between languages at or above clause levels as *code-mixing*. Romaine (1995) refers to both inter-sentential and intra-sentential switches as code-switching. Bilingual speakers may shift from one language to another entirely (Poplack et al., 1988) or they mix languages partially within the single speech (Gumperz, 1982). In this study, we focus on code-switching within the same post in an online discussion forum used by Turkish-Dutch bilinguals.

There are different theoretical models which support (Myers-Scotton, 2002; Poplack, 1980) or reject (MacSwan, 2005; Thomason and Kaufman, 2001) linguistic constraints on code-switching. According to (Thomason and Kaufman, 2001; Gardner-Chloros and Edwards, 2004) linguistic factors are mostly unpredictable since social factors govern the multilingual environments in most cases. Bhatt and Bolonyai (2011) have an extensive study on socio-cognitive factors that lead to code-switching across different multilingual communities.

Although multilingual communication has been widely studied through spoken data analyses, research on online communication is relatively recent. In terms of linguistic factors Cárdenas-Claros and Isharyanti (2009) report differences between Indonesian-English and Spanish-English speakers in their amount of code-switching on MSN (an instant messaging client). Durham (2003) finds a tendency to switch to English over time in an online multilingual (German, French, Italian) discussion forum in Switzerland.

The media (e.g. IRC, Usenet, email, online discussions) used for multilingual conversations influence the amount of code-switching as well (Paolillo, 2001; Hinrichs, 2006). Androutsopoulos and Hinnenkamp (2001), Tsaliki (2003) and Hinnenkamp (2008) have done qualitative analyses of switch patterns across German-Greek-Turkish, Greek-English and Turkish-German in online environments respectively.

In terms of social factors, a number of studies have investigated the link between topic and language choices qualitatively (Ho, 2007; Androutsopoulos, 2007; Tang et al., 2011). These studies share the similar conclusion that multilingual speakers use minority languages to discuss topics related to their ethnic identity and reinforcing intimacy and self-disclosure (e.g. homeland, cultural traditions, joke telling) whereas they use the majority language for sports, education, world politics, science and technology.

Computational approaches to code-switching Recently, an increasing number of research within NLP has focused on dealing with multilingual documents. For example, corpora with multilingual documents have been created to support studies on code-switching (e.g. Cotterell et al. (2014)) To enable the automatic processing and analysis of documents with mixed languages, there is a shift in focus toward language identification at the word level (King and Abney, 2013; Nguyen and Doğruöz, 2013; Lui et al., 2014). Most closely related to our work is the study by Solorio and Liu (2008) who predict code-switching in recorded English-Spanish conversations. Compared to their work, we use a large-scale social media dataset that enables us to explore novel features.

The task most closely related to automatic prediction of code-switching is automatic language identification (King and Abney, 2013; Nguyen and Doğruöz, 2013; Lui et al., 2014). While automatic language detection uses the words to identify the language, automatic prediction of code-switching involves predicting whether the language of the next word is the same *without* having access to the next word itself.

Language practices of the Turkish community in the Netherlands Turkish has been in contact with Dutch due to labor immigration since the 1960s and the Turkish community is the largest minority group (2% of the whole population) in the Netherlands (Centraal Bureau voor de Statistiek, 2013). In addition to their Dutch fluency, second and third generations are also fluent in Turkish through speaking it within the family and community, regular family visits to Turkey and watching Turkish TV through satellite dishes. These speakers grow up speaking both languages simultaneously rather than learning one language after the other (De Houwer, 2009). In addition to constant switches between Turkish and Dutch, there are also literally translated Dutch multi-word expressions (Doğruöz and Backus, 2007; Doğruöz and Backus, 2009). Due to the religious backgrounds of the Turkish-Dutch community, Arabic

words and phrases (e.g. greetings) are part of daily communication. In addition, English words and phrases are used both in Dutch and Turkish due to the exposure to American and British media.

Although the necessity of studying immigrant languages in Dutch online environments has been voiced earlier (Dorleijn and Nortier, 2012), the current study is the first to investigate mixed language communication patterns of Turkish-Dutch bilinguals in online environments.

3 Dataset

Our data comes from a large online forum (Hababam) used by Turkish-Dutch speakers. The forum is active since 2000 and contains 28 subforums on a variety of topics (e.g. sports, politics, education). Each subforum consists of multiple threads which start with a thread title (e.g. a statement or question) posted by a moderator or user. The users are Turkish-Dutch bilinguals who reside in the Netherlands. Although Dutch and Turkish are used dominantly in the forum, English (e.g. fixed expressions) and Arabic (e.g. prayers) are occasionally used (less than 1%) as well. We collected the data between June 2005 and October 2012 by crawling the forum. Statistics of our data are shown in Table 1.

	Frequency
Number of posts	4,519,869
Number of users	14,923
Number of threads	113,517
Number of subforums	29

Table 1: Dataset Statistics

The subforums *Chit-Chat* (1,671,436), *Turkish youth & love* (447,436), and *Turkish news & updates* (418,135) have the highest post frequency whereas *Columns* (4727), *Science & Philosophy* (5083) and *Other Beliefs* (6914) have the lowest post frequency.

An automatic language identification tagger is used to label the language of the words in posts and titles of the threads. The tagger distinguishes between Turkish and Dutch using logistic regression (Nguyen and Doğruöz, 2013) and achieves a word accuracy of approximately 97%. We use the language labels to train our classifier (since given the labels we can determine whether there is a switch or not), and to evaluate our model.

4 Types of Code-Switching

In this section, we provide a qualitative analysis of code-switching in the online forum. We differentiate between two types of code-switching: code-switching across posts and code-switching within the same post.

4.1 Code-switching across posts

Within the same discussion thread, users react to posts of other users in different languages. In example (2), user 1 posts in Dutch to tease User 2. User 2 reacts to this message with a humorous idiomatic expression in Turkish (i.e. [*adim cikmis*] "I made a name") to indirectly emphasize that there is no reason for her to defend herself since she has already become famous as the *perfect* person in the online community. This type of humorous switch has also been observed for Greek-English code-switching in face-to-face communication (Gardner-Chloros and Finnis, 2003). The text is written with Dutch orthography instead of conventional Turkish orthography (i.e. [*adım çıkmış*]). It is probably the case that the user has a Dutch keyboard without Turkish characters. However, writing with non-Turkish characters in online environments is also becoming popular among monolingual Turkish users from Turkey.

Example (2)
User1: <dutch> je hoefde niet gelijk in de verdediging te schieten hoor </dutch> :P
Tra: "you do not need to be immediately defensive dear"

User2: <turkish> zaten adim cikmis mukemmel sahane kusursuz insana, bi de yine cikmasin </turkish> :(
Tra: "I already have established a name as a great amazing perfect person, I do not need it to spread around once more"

Example (3) is taken from a thread about breakfast traditions. The users have posted what they had for breakfast that day. The first user talks about his breakfast in Turkish and describes the culture specific food items (e.g. *borek* "Turkish pastry") prepared by his mother. The second user describes a typical Dutch breakfast and therefore switches to Dutch.

Example (3)
User1: <turkish>annemin peynirli borekleri ve cay</turkish>
Tra: "the cheese pastries of my mom and tea"

User2: <dutch>Twee sneetjes geroosterd bruin brood met kipfilet en een glas thee.</dutch>
Tra: "Two pieces of roasted brown bread with chicken filet and a cup of tea"

4.2 Code-switching within the same post

In addition to code-switching across posts, we encountered code-switching within the same post of a user as well. Manual annotation of a subset of the posts in Nguyen and Doğruöz (2013), suggests that less than 20% of the posts contain a switch. Example (4) is taken from a thread about Mother's Day and illustrates an intra-sentential switch. The user starts the post in Dutch (*vakantie boeken* "to book a vacation") and switches to Turkish since booking a vacation through internet sites or a travel agency is a typical activity associated with the Dutch culture.

Example (4)
<dutch>vakantie boeken</dutch>
<turkish> yaptim annecigimee </turkish>
Tra[1]:"(I) <dutch>booked a holiday</dutch>
<turkish>for my mother.</turkish>"

Example (5) is taken from a thread about Turkish marriages and illustrates an inter-sentential switch. The user is advising the other users in Turkish to be very careful about choosing their partners. Since most Turkish community members prefer Turkish partners and follow Turkish traditions for marriage, she talks about these topics in Turkish. However, she switches to Dutch when she talks about getting a diploma in the Dutch school system. Similar examples of code-switching for emphasizing different identities based on topic have been observed for other online and face-to-face communication as well (Androutsopoulos, 2007; Gardner-Chloros, 2009).

Example (5)
<turkish>Allah korusun yani. Kocani iyi sec diyim=) evlilik evcilik degildir.</turkish>
<dutch>Al zou ik wanneer ik getrouwd ben een HBO diploma op zak hebben, zou ik hem dan denk ik niet verlaten.</dutch>
Tra:"<turkish> May God protect you. Choose your husband carefully. Marriage is not a game </turkish> <dutch> Even if I am married and have a university diploma, I don't think I will leave him </dutch>"

Code-switching through greetings, wishes and formulaic expressions are commonly observed in bilingual face-to-face communication and online immigrant forums as well (Androutsopoulos, 2007; Gardner-Chloros, 2009).

5 Experimental Setup

The focus of this paper is on code-switching within the same post. We discuss the setup and features of our experiment in this section.

5.1 Goal

We cast the prediction of the code-switch point within the post as a binary classification problem. We define the i-th token of the post as an instance. If the $i+1$th token is in a different language, the label is 1. Otherwise, the label is 0.

Obtaining language labels In order to label each token of a post, we rely on the labels obtained using automatic language identification at the word level (see Section 3). This process may not be the most accurate way of labeling each token of a post at a large scale. One particular artifact of this procedure is that an automatic tagger may falsely tag the language of a token in longer posts. As a result, some lengthy posts might appear to have one or more code-switches by accident. However, since the accuracy of our tagger is high (approx. 97% accuracy), we expect the amount of such spurious code-switches to be low. For future work, we plan to experiment on a dataset based on automatic language identification as well as a smaller dataset using manual annotation.

5.2 Creating train and test sets

Before we attempt to train a classifier on our data, we eliminate the biases and imbalances. The majority of posts do not contain any switches. As a consequence, the number of instances that belong to the '0' class (i.e. no code-switching occurring after the current word) grossly outnumber the instances of class '1', where code-switching takes place. In order to alleviate this class imbalance, for all our experiments, we sample an *equal* amount of instances from '0' and '1' classes randomly [2], both for our training and testing data. This way the result will not favor the '0' class even if we randomly decide on the class label for each instance. The average number of training and testing

[1] It is possible to drop the subject pronoun in Turkish. As typical in bilingual speech, an additional Turkish verb *yapmak* follows the Dutch verb *boeken* "to book".

[2] We do 100 iterations and average the results of all these independent samples.

instances per iteration was 4000 and 80000 respectively. By drawing 100 independent samples from the entire dataset, we cover a reasonable portion of the full data and do not sacrifice the balance of the two classes, which is crucially important for the validity of our results.

5.3 Feature selection

We use the following features (see Table 2) to investigate code-switching within a post.

5.3.1 Non-linguistic features

Emoticons Emoticons are iconic symbols that convey emotional information along with language use in online environments (Dresner and Herring, 2014). Emoticons have mostly been used in the context of sentiment analysis (e.g. Volkova et al. (2013), Chmiel et al. (2011)). Park et al. (2014) studied how the use of emoticons differ across cultures in Twitter data. Panayiotou (2004) studied how bilinguals express emotions in face-to-face environments in different languages. We are the first to investigate the role of emoticons as a non-linguistic factor in predicting code-switching on social media.

Emoticons in our data are either signified by a special tag `[smiley:smiley_type]` or can appear in any of the common ASCII emoticon forms (e.g. `:)`, `:-)` etc.). In order to detect the emoticons, we used a hand picked list of ASCII emoticons as our dictionary, as well as a filter that searched for the special emoticon tag. Since we rely on an automatic language tagger, the language label of a particular emoticon depends on its surrounding tokens. If an emoticon is within a block of text that is tagged as Turkish, then the emoticon will automatically obtain a Turkish label (and accordingly for Dutch). For future work, we will experiment with labeling emoticons differently (e.g. introducing a third, neutral label).

To assess the strength of emoticons as predictors of code-switching, we generate 4 different features (see Table 2). These features capture whether or not there is an emoticon *at* or *before* the token that we want to classify as the switch boundary between Dutch and Turkish. We record whether there was an emoticon at token i (i.e. the token we want to classify), token $i-1$ and token $i-2$.

The last emoticon feature records whether there is any emoticon *after* the current token. We note that this feature looks ahead (after the i-th token), and therefore cannot be implemented in a real time system which predicts code-switching on-the-fly. However, we included the feature for exploratory purposes.

5.3.2 Linguistic features

Language around the switch point We also investigate whether the knowledge of the language of a couple of tokens before the token of interest, as well as the language at the token of interest, hold some predictive strength. These features correspond to #1-3 in Table 2. Generally, the language label is binary. However, if there are no tokens in positions $i-2$ or $i-1$ for features #1 and #2, we assign a third value to represent this non-existence. Additionally, we explore whether a previous code-switching in a post triggers a second code-switching later in the same post. We test this hypothesis by recording feature #4 which represents the existence of code-switching before token i.

Single word versus multi-word switch There is an on-going discussion in multilingualism about the classification of switched tokens (Poplack, 2004; Poplack, 2013) and whether there are linguistic constraints on the switches (Myers-Scotton, 2002). In addition to switches across individual lexical tokens, multilingual speakers also switch across multi-word expressions.

Automatic identification of multi-word expressions in monolingual language use have been widely discussed (Baldwin et al., 2003; Baldwin and Kim, 2010) but we know little about how to predict switch points that include multi-word expressions. We are the first to include multi-word expressions as a feature to predict code-switching. We are mostly inspired by (Schwartz et al., 2013) in identifying MWEs.

More specifically, we built a corpus of 3-gram MWEs (2,241,484 in total) and selected the most frequent 100 MWEs. We differentiate between two types of MWEs: Let the i-th token of a post be the switch point. For *type 1*, we take 3 tokens (all in the same language) right before the switch token (i.e. terms $i-3, i-2, i-1$). [*Allah razi olsun*] "May the Lord be with you" and [*met je eens*] "agree with you" are the two of the most frequent MWEs (in Turkish and Dutch respectively).

For *type 2*, we take the tokens $i-2$, $i-1$, i and the last token is in a different language (e.g. [*Turkse premier Recep*] "Turkish prime-minister

Table 2: Features

Feature #	Feature Description
1	Language of token in position $i-2$
2	Language of token in position $i-1$
3	Language of token in position i (current token)
4	Was there code-switching before the current token?
5	Is there an emoticon in position $i-2$?
6	Is there an emoticon in position $i-1$?
7	Is there an emoticon in position i?
8	Are there any emoticons in positions after i?
9	Is the i-th token the first word of a 3-word multi-word expression?
10	Is the i-th token the second word of a 3-word multi-word expression?
11	Is the i-th token the third word of a 3-word multi-word expression?

Recep").

The first type of MWEs captures whether an MWE (all three words in the same language), signifies code-switching for token i or not.

The second type investigates whether there are MWEs that "spill over" the code-switching point (i.e. the first two tokens of an MWE are in the same language, but the third token is in another language). In order to get a good estimate of the MWEs in our corpus, we count the occurrences of all these 3-grams and keep the top scoring ones in terms of frequency, which end up as our dictionary of MWEs.

6 Results

To evaluate the predictive strength of our features, we conduct experiments using a Naive Bayes classifier.

In order to measure the performance, we train the classifiers for various combinations of the features shown in Table 2. As we described in the previous section, we train on randomly chosen, class-balanced parts of the data and we test on randomly selected balanced samples (disjoint from the training set), averaging over 100 runs. For each combination of features, we measure and report average precision, recall, and F1-score, with respect to positively predicting code-switching.

Table 3 illustrates the performance of individual features used in our classifier. Features that concern the language of the previous tokens (i.e. features #1 & #2) seem to perform better than chance in predicting code-switching. On the other hand, features #3 (*language of the token in position i*) and #4 (*previous code-switching*) have the worst performance. In fact, the obtained classifier always predicts *no code-switching* regardless of the value of the feature. Therefore, both precision and recall are 0. Features #1 & #2 behave differently from features #3 & #4 because #1 & #2 have ternary values (the token language, or *non-existing*). This probably forces the classifiers to produce a non-constant decision. For instance, the model for feature #1 decides positively for code-switching if the language label is either *Turkish* or *Dutch* and decides negatively if the label is *non-existing*.

Table 3: Performance of individual features

Feature #	Precision	Recall	F1 score
1	0.6305	1	0.7733
2	0.6362	1	0.7776
3	0	0	-
4	0	0	-
5	0.704	0.2116	0.3254
6	0.7637	0.2324	0.3564
7	0.8025	0.1339	0.0954
8	0.4879	0.3214	0.3875
9	0.5324	0.7819	0.6335
10	0.5257	0.8102	0.6376
11	0.5218	0.8396	0.6436

The rest of the individual features perform similarly but worse than #1 and #2. Therefore, it is necessary to use a combination of features instead of single ones.

After examining how features perform individually, we further investigate how features behave in groups. We first group the features into homogenous categories (e.g. #1-#3 focus on the language of tokens, #5-#8 record the presence of emoticons and #9-#11 refer to MWEs). Subsequently, we test the performance of these categories in different combinations, and finally measure the effect of

Table 4: Performance of groups of features

	Features	Precision	Recall	F1 score
1-3	Language of tokens	0.6362	1	0.7777
1-4	Language + previous code-switching	0.6663	0.1312	0.6663
5-8	Emoticons	0.6638	0.397	0.2766
9-11	MWEs	0.5384	0.7476	0.626
5-11	Emoticons + MWEs	0.52	0.8718	0.6466
1-8	Language + previous code-switching + emoticons	0.6932	0.5114	0.4634
1-4, 9-11	Language + previous code-switching + MWEs	0.712	0.7297	0.7113
1-11	All	0.6847	0.8034	0.7106

using all our features for the task. Table 4 shows the combinations of the features we used, as well as the average precision, recall, and F1-score.

According to Table 4, the combination of the language of the tokens (features #1-#3) and the previous code-switching earlier in the post (features #1-#4), and MWEs (features #9-#11) perform the highest in terms of precision/recall. Features #3 and #4 have rather low performances on their own but they yield a strong classifier in combination with other features.

When we use features that record emoticons (#5-#8) or MWEs (#9-#11) alone, the performance of our classifier decreases. In general, MWEs outperform emoticons. We observe this performance boost when we combine emoticon features with other features (e.g. #1-#8) and with MWEs together in the same subset (#1-#4, #9-#11).

7 Conclusion

We focused on predicting code-switching points for a mixed language online forum used by the Turkish-Dutch immigrant community in the Netherlands. For the first time, a long term data set was used to investigate code-switching in social media. We are also the first to test new features (e.g. emoticons and MWEs) to predict code-switching and to identify the features with significant predictive strength. For future work, we will continue our investigation with exploring the predictive value of these new features within the Turkish-Dutch immigrant community as well as others.

8 Acknowledgements

The first author was supported by the National Science Foundation (NSF), Grant No. IIS-1247489. The second author was supported by the Netherlands Organization for Scientific Research (NWO) grant 640.005.002 (FACT). The third author was supported by a Digital Humanities Research Grant from Tilburg University and a research fellowship from Netherlands Institute for Advanced Study.

References

Jannis Androutsopoulos and Volker Hinnenkamp. 2001. Code-switching in der bilingualen chatkommunikation: ein explorativer blick auf# hellas und# turks. *Beisswenger, Michael (ed.)*, pages 367–401.

Jannis Androutsopoulos, 2007. *The Multilingual Internet*, chapter Language choice and code-switching in German-based diasporic web forums, pages 340–361. Oxford University Press.

Peter Auer and Li Wei, 2007. *Handbook of multilingualism and multilingual communication.*, chapter Introduction: Multilingualism as a problem? Monolingualism as a problem, pages 1–14. Berlin: Mouton de Gruyter.

Peter Auer. 1988. A conversation analytic approach to code-switching and transfer. *Codeswitching: Anthropological and sociolinguistic perspectives*, 48:187–213.

Timothy Baldwin and Su Nam Kim. 2010. Multiword expressions. *Handbook of Natural Language Processing, second edition. Morgan and Claypool*.

Timothy Baldwin, Colin Bannard, Takaaki Tanaka, and Dominic Widdows. 2003. An empirical model of multiword expression decomposability. In *Proceedings of the ACL 2003 workshop on Multiword expressions: analysis, acquisition and treatment-Volume 18*, pages 89–96. Association for Computational Linguistics.

Rakesh M Bhatt and Agnes Bolonyai. 2011. Code-switching and the optimal grammar of bilingual language use. *Bilingualism: Language and Cognition*, 14(04):522–546.

Barbara E Bullock and Almeida Jacqueline Toribio. 2009. *The Cambridge handbook of linguistic code-switching*, volume 1. Cambridge University Press Cambridge.

Monica S. Cárdenas-Claros and Neny Isharyanti. 2009. Code-switching and code-mixing in internet chatting: Between'yes,"ya,'and'si'-a case study. *The Jalt Call Journal*, 5(3):67–78.

Centraal Bureau voor de Statistiek. 2013. Bevolking, generatie, geslacht, leeftijd en herkomstgroepering. 2013.

Anna Chmiel, Julian Sienkiewicz, Mike Thelwall, Georgios Paltoglou, Kevan Buckley, Arvid Kappas, and Janusz A Hołyst. 2011. Collective emotions online and their influence on community life. *PloS one*, 6(7):e22207.

Ryan Cotterell, Adithya Renduchintala, Naomi Saphra, and Chris Callison-Burch. 2014. An algerian arabic-french code-switched corpus. In *LREC*.

Anna De Fina. 2007. Code-switching and the construction of ethnic identity in a community of practice. *Language in Society*, 36(03):371–392.

Annick De Houwer. 2009. *Bilingual first language acquisition*. Multilingual Matters.

A Seza Doğruöz and Ad Backus. 2007. Postverbal elements in immigrant Turkish: Evidence of change? *International Journal of Bilingualism*, 11(2):185–220.

A. Seza Doğruöz and Ad Backus. 2009. Innovative constructions in Dutch Turkish: An assessment of ongoing contact-induced change. *Bilingualism: Language and Cognition*, 12(01):41–63.

Margreet Dorleijn and Jacomine Nortier, 2012. *The Cambridge Handbook of Linguistic Code-switching*, chapter Code-switching and the internet, pages 114–127. Cambridge University Press.

Eli Dresner and Susan C Herring. 2014. Emoticons and illocutionary force. In *Perspectives on Theory of Controversies and the Ethics of Communication*, pages 81–90. Springer.

Mercedes Durham. 2003. Language choice on a Swiss mailing list. *Journal of Computer-Mediated Communication*, 9(1):0–0.

European Comission. 2012. Europeans and their languages: Special barometer 386. Technical report, European Comission.

Penelope Gardner-Chloros and Malcolm Edwards. 2004. Assumptions behind grammatical approaches to code-switching: When the blueprint is a red herring. *Transactions of the Philological Society*, 102(1):103–129.

Penelope Gardner-Chloros and Katerina Finnis. 2003. How code-switching mediates politeness: Gender-related speech among London Greek-Cypriots. *Sociolinguistic Studies*, 4(2):505–532.

Penelope Gardner-Chloros, 2009. *Handbook of Code-switching*, chapter Sociolinguistic Factors in Code-Switching, pages 97–114. Cambridge University Press.

Francois Grosjean. 2014. Bicultural bilinguals. *International Journal of Bilingualism*, xx(xx):1–15.

John J Gumperz. 1982. *Discourse strategies*, volume 1. Cambridge University Press.

Volker Hinnenkamp. 2008. Deutsch, Doyc or Doitsch? Chatters as languagers–The case of a German–Turkish chat room. *International Journal of Multilingualism*, 5(3):253–275.

Lars Hinrichs. 2006. *Codeswitching on the Web: English and Jamaican Creole in E-mail Communication (Pragmatics & Beyond, Issn 0922-842x)*. John Benjamins.

Judy Woon Yee Ho. 2007. Code-mixing: Linguistic form and socio-cultural meaning. *The International Journal of Language Society and Culture*, 21.

Ben King and Steven P Abney. 2013. Labeling the languages of words in mixed-language documents using weakly supervised methods. In *HLT-NAACL*, pages 1110–1119.

Marco Lui, Jey Han Lau, and Timothy Baldwin. 2014. Automatic detection and language identification of multilingual documents. *Transactions of the Association for Computational Linguistics*, 2:27–40.

David Luna, Torsten Ringberg, and Laura A Peracchio. 2008. One individual, two identities: Frame switching among biculturals. *Journal of Consumer Research*, 35(2):279–293.

Jeff MacSwan. 2005. Codeswitching and generative grammar: A critique of the mlf model and some remarks on "modified minimalism". *Bilingualism: language and cognition*, 8(01):1–22.

Carol Myers-Scotton. 2002. *Contact linguistics: Bilingual encounters and grammatical outcomes*. Oxford University Press Oxford.

Dong Nguyen and A. Seza Doğruöz. 2013. Word level language identification in online multilingual communication. In *Proceedings of EMNLP 2013*.

Jonathan Owens and Jidda Hassan, 2013. *Information Structure in Spoken Arabic*, chapter Conversation markers in Arabic-Hausa code-switching, pages 207–243. Routledge Arabic Linguistics. Routledge.

Alexia Panayiotou. 2004. Switching codes, switching code: Bilinguals' emotional responses in english and greek. *Journal of multilingual and multicultural development*, 25(2-3):124–139.

John C Paolillo. 2001. Language variation on internet relay chat: A social network approach. *Journal of sociolinguistics*, 5(2):180–213.

Jaram Park, Young Min Baek, and Meeyoung Cha. 2014. Cross-cultural comparison of nonverbal cues in emoticons on twitter: Evidence from big data analysis. *Journal of Communication*, 64(2):333–354.

Shana Poplack, David Sankoff, and Christopher Miller. 1988. The social correlates and linguistic processes of lexical borrowing and assimilation. *Linguistics*, 26(1):47–104.

Shana Poplack. 1980. Sometimes i'll start a sentence in spanish y termino en espanol: toward a typology of code-switching1. *Linguistics*, 18(7-8):581–618.

Shana Poplack, 2004. *Soziolinguistik. An international handbook of the science of language*, chapter Codeswitching, pages 589–597. Walter de Gruyter, 2nd edition.

Shana Poplack. 2013. "sometimes i'll start a sentence in spanish y termino en español": Toward a typology of code-switching. *Linguistics*, 51(Jubilee):11–14.

Suzanne Romaine. 1995. Bilingualism (2nd edn). *Malden, MA: Blackwell Publishers*.

H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Stephanie M Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin EP Seligman, et al. 2013. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, 8(9):e73791.

Carmen Silva-Corvalán. 1994. *Language Contact and Change: Spanish in Los Angeles.* ERIC.

Thamar Solorio and Yang Liu. 2008. Learning to predict code-switching points. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 973–981. Association for Computational Linguistics.

Dai Tang, Tina Chou, Naomi Drucker, Adi Robertson, William C Smith, and Jeffery T Hancock. 2011. A tale of two languages: strategic self-disclosure via language selection on facebook. In *Proceedings of the ACM 2011 conference on Computer supported cooperative work*, pages 387–390. ACM.

Sarah Grey Thomason and Terrence Kaufman. 2001. *Language contact*. Edinburgh University Press Edinburgh.

Almeida Jacqueline Toribio. 2002. Spanish-english code-switching among us latinos. *International journal of the sociology of language*, pages 89–120.

Liza Tsaliki. 2003. Globalization and hybridity: the construction of greekness on the internet. *The Media of Diaspora, Routledge, London*.

Svitlana Volkova, Theresa Wilson, and David Yarowsky. 2013. Exploring demographic language variations to improve multilingual sentiment analysis in social media. In *EMNLP*, pages 1815–1827.

Li Wei, 1998. *Codeswitching in conversation: Language, interaction and identity*, chapter The 'why' and 'how' questions in the analysis of conversational codeswitching, pages 156–176. Routledge.