Word-Level Language Identification and Predicting Codeswitching Points in Swahili-English Language Data

Mario Piergallini, Rouzbeh Shirvani, Gauri S. Gautam, and **Mohamed Chouikha**
Howard University Department of Electrical Engineering and Computer Science
2366 Sixth St NW, Washington, DC 20059
mario.piergallini@howard.edu,Rouzbeh.asgharishir@bison.howard.edu
gaurishankar.gaut@bison.howard.edu, mchouikha@howard.edu

Abstract

Codeswitching is a very common behavior among Swahili speakers, but of the little computational work done on Swahili, none has focused on codeswitching. This paper addresses two tasks relating to Swahili-English codeswitching: word-level language identification and prediction of codeswitch points. Our two-step model achieves high accuracy at labeling the language of words using a simple feature set combined with label probabilities on the adjacent words. This system is used to label a large Swahili-English internet corpus, which is in turn used to train a model for predicting codeswitch points.

1 Introduction

Language technology has progressed rapidly in many applications (speech recognition and synthesis, parsing, translation, sentiment analysis, etc.), but efforts have been focused mainly on large, high-resource languages and on monolingual data. Many tools have not been developed for low-resource languages nor can they be applied to mixed-language data containing codeswitching. In many cases, dealing with low-resource languages requires the ability to deal with codeswitching. For example, it is quite common to codeswitch between the lingua franca and English in many former English colonies in Africa, such as Kenya, Zimbabwe and South Africa (Myers-Scotton, 1993b). Thus, expanding the reach of language technologies to users of these languages may require the ability to handle mixed-language data, depending on which domains it is intended for.

Codeswitching produces additional challenges for NLP due to the simple fact that monolingual tools cannot be applied to mixed-language data. Beyond that, codeswitching also has its own peculiarities and can convey meaning in and of itself, and these aspects are worthy of study as well. Codeswitching can be used to increase or decrease social distance, indicate something about a speaker's social identity or their stance towards the subject of discussion, or to draw attention to particular phrases (Myers-Scotton, 1993b). Sometimes, of course, it may simply indicate that the speaker does not know the word in the other language, or is not able to recall it quickly in this instance. Computational approaches to discourse analysis will require tools specific to codeswitching in order to be able to make use of these social meanings.

Multiple theories propose grammatical constraints on codeswitching (Myers-Scotton, 1993a), and computational approaches may contribute to providing stronger evidence for or against these theories (Solorio and Liu, 2008). These grammatical constraints also can inform the social interpretation of codeswitching. If a codeswitch occurs in a position that is less expected, it may be more likely to have been used for effect. Similarly, when a codeswitch occurs in a less likely context based on features of the discourse, this also affects the interpretation. The longer a discussion is carried out in a single language, the more likely it would seem that a switch indicates a change in the discourse. For example, Carol Myers-Scotton (1993b) analyzes a conversation where a switch to Swahili and then to English after small talk in the local language adds

Proceedings of the Second Workshop on Computational Approaches to Code Switching, pages 21–29,
Austin, TX, November 1, 2016. ©2016 Association for Computational Linguistics

force to the speaker's rejection of a request. This type of switch could also be precipitated by a change in conversation topic, task (e.g. pre-class small talk transitioning into the beginning of lessons), location, etc. By contrast, in conversations where participants switch frequently between languages, each individual switch carries less social meaning. In those situations, it is the overall pattern of codeswitching that conveys meaning (Myers-Scotton, 1993b). A model should be able to see this pattern and adjust the likelihood of switches accordingly. Being able to predict how likely a switch is to occur in a particular position may thus provide information to aid in the social analysis of codeswitching behavior.

In this paper, we will be introducing two corpora of Swahili-English data. One is comprised of live interviews from Kenya, while the other was scraped from a large Tanzanian/Kenyan Swahili-language internet community. We will be analyzing codeswitching in both data sets. Human-annotated interviews and a small portion of human-annotated internet data are used to train a language identification model, which is then applied to the larger internet corpus. The interview data and this automatically-labeled data are then used in training a model for predicting codeswitch points.

There are few NLP tools for Swahili and we could find no prior computational work on Swahili that addressed codeswitching. Additionally, available corpora in Swahili are monolingual, so the creation of two sizable corpora of mixed Swahili-English data will be valuable to research in this area.

2 Prior Research

2.1 Language Identification

Until recent years, most work on automatic language identification focused on identifying the language of documents. Work on language identification of very short documents can be found, for example, in Vatanen et al. (2010). But language identification at the word level in codeswitching data has begun to receive more attention in recent years, particularly with the First Workshop on Computational Approaches to Codeswitching (FWCAC). The workshop had a shared task in language identification, with eight different teams submitting systems on the four language pairs included (Spanish-English, Nepali-English, Mandarin-English and Modern Standard Arabic-Egyptian Arabic) (Solorio et al., 2014). Additionally, prior to this workshop, some work had been done on word-level language identification in Turkish-Dutch data (Nguyen and Doğruöz, 2013) and on language identification on isolated tokens in South African languages (Giwa and Davel, 2013), both with an eye towards analyzing codeswitching.

Most, if not all, of the previous approaches to word-level language identification utilized character *n*-grams as one of the primary features (Nguyen and Doğruöz, 2013; Giwa and Davel, 2013; Lin et al., 2014; Chittaranjan et al., 2014; Solorio et al., 2014). Those focused on intrasentential codeswitching also utilized varying amounts of context. Nguyen and Doğruöz (2013) and all but one of the systems submitted to the shared task at FWCAC used contextual features. A number of other types of features have been utilized as well, including capitalization, text encoding, word embedding, dictionaries, named entity gazetteer, among others (Solorio et al., 2014; Volk and Clematide, 2014). Significant variation in the difficulty of the task has been found between language pairs. More closely related languages can be more difficult if they also share similar orthographic conventions, as was found with the MSA-Egyptian Arabic language pair (Solorio et al., 2014). In the FWCAC shared task, notable declines in system performance were found when introduced to out-of-domain data.

2.2 Codeswitch Point Prediction

There has been significantly less work done on the task of predicting codeswitch points. We could only find two articles that deal precisely with this task, Solorio and Liu (2008) and Papalexakis, Nguyen and Doğruöz (2014). The two groups take fairly different approaches to feature design and performance evaluation, while both groups use naïve Bayes classifiers. Solorio and Liu also explore Voting Feature Intervals.

Solorio and Liu look at English-Spanish codeswitching in a relatively small conversational data set created for the study. They use primarily phrase constituent position and part-of-speech tagger outputs as features. The word, its language and its human-annotated POS were also used. These

	Interviews	JamiiForums
# Utterances/Posts	10,105	220,434
# Words (tokens)	188,188	16,176,057
Avg. words/item	18.6	73.4
% English words	84.5%	45.8%
% Swahili words	15.4%	54.1%
% Mixed words	<0.1%	<0.1%
% Other words	<0.1%	<0.1%

Table 1: Data set Stats

were tested both with and without the features for the previous word. Initial evaluation was done using F1-scores, but as noted in the paper, codeswitching is never a forced choice. As such, the upper-bound on this task should be relatively low. To get around this issue, Solorio and Liu came up with a novel approach to test performance by artificially generating codeswitched sentences. These sentences were scored for naturalness by bilingual speakers and compared to naturally-occurring codeswitched sentences. Their model achieved scores not far from the natural examples. This approach seems well-justified but requires significant human input.

Papalexakis et al. use simpler features focused on the context of the word. These include the language of the word and the two previous words, whether there was codeswitching previously in the document, the presence of emoticons in the previous two words and the following words, and whether the word is part of a common multi-word expression. These features are applied on a large data set from a Turkish-Dutch internet forum. The language of tokens in this data was labeled automatically using the system in Nguyen and Doğruöz (2013). They find that these features are useful, particularly the language sequence features. The exception is that the emoticon-based features actually reduced performance when combined with other features.

3 Data Sets

The two data sets we use in this paper come from very different domains. The first is comprised of live interviews, and as such is spoken conversation. The second is from a large internet forum, and so is casual, written data with use of emoticons and other behaviors specific to computer-mediated communication. The use of data from two linguistic domains also provides a test of the robustness of our model.

Some descriptive statistics about the two data sets can be seen in Table 2.2. The forum data set is a couple of orders of magnitude larger than the interview data set. Our utterances are similar in length to the posts analyzed in Nguyen and Doğruöz (2013). JamiiForums posts are considerably longer than both.

3.1 Kenyan Interviews

The interviews in this data set were conducted in Kenya. The participants were students at a Kenyan university and the interviewers were a combination of other students and professors at the same university. Most of the participants were interviewed twice, once by a student and once by a professor. This provides two social contexts, one in which the participant and interviewer have the same social status, and one in which the interviewer has a higher social status. In our examination of the data, some differences can be seen in codeswitching behavior in these two situations, but this is beyond the scope of this paper.

The interviews were transcribed, translated, and the words were tagged by language by native speakers of Swahili who are fluent in English. Words were labeled as either English, Swahili, mixed, or other. Some words were originally labeled as Sheng, which is a term used in Kenya for a register of heavy codeswitching with urban street slang (Mazrui, 1995). The annotators were not instructed to use this label. Since most Sheng words clearly originate in either Swahili or English, they were relabeled accordingly. In contrast to the FWCAC shared task (Solorio et al., 2014), we did not label named entities or ambiguous words. Words that might have been labeled that way were instead labeled according to the context – a proper name was labeled as English if it was surrounded by English, or Swahili if it was surrounded by Swahili. Such words that occurred at language boundaries were labeled with the following words or the words within the same sentence (if it occurred at the end of a sentence). There were relatively few instances of this.

3.2 JamiiForums Internet Data

The internet data comes from a large Tanzania-based internet forum named JamiiForums[1]. It was scraped

[1] https://www.JamiiForums.com

by a Python script, but due to changes in the forum software and increased security, scraping was not completed. Thus the data only comprises a small fraction of the entire forum. As we already had a large amount of data, we did not feel it necessary to continue immediately[2]. Since our interview data came from Kenya, we prioritized scraping the entire Kenyan subforum first.

During scraping, full URLs, embedded images and email addresses were replaced with placeholder terms. Bare hostnames[3] were left alone since they can double as the name of an organization or website. Emoticons were replaced with the name of the emoticon as defined by the hover text or image file name. Text within quotation boxes was separated from text in the main body of the post. However, given that users do not always format their posts correctly, some improperly formatted forum code will inevitably have been included in our data.

Language labels for 22,592 tokens of the JamiiForums data were annotated by a native English speaker. These were annotated according to the same rules as the interview data. Annotation was done after applying the initial language identification model to the forum data, with only disagreements being labeled by the annotator. This significantly increased the speed that annotation could be done.

4 Language Identification Task

4.1 Methodology

For the language identification task, we applied some additional preprocessing to the data. First, the data was tokenized to split words from punctuation marks other than word-internal periods, apostrophes and hyphens. Then all punctuation except for periods, question and exclamation marks were removed. Prior work has explored whether emoticons have any influence on codeswitching behavior (Papalexakis et al., 2014), but did not find them to be significantly useful. Therefore other symbols with no lexical content such as emoticons and the placeholders for embedded images, etc. were also removed.

After this, we experimented with a few different features before settling on the final set. The first type of feature we used was character *n*-grams (unigrams, bigrams and trigrams), filtered to exclude *n*-grams that occurred less than 25 times. The symbol # was appended to the beginning and end of the word to enable the *n*-gram features to capture prefixes and suffixes. Additionally, we used a capitalization feature, English and Swahili dictionary features and a regular expression feature. The capitalization feature categorized words by whether the first letter only was capitalized and if so, whether it occurred at the beginning of a sentence. Otherwise, words were categorized as either all lower case, all upper case, or all numbers and symbols. Words which did not match any of those patterns were labeled as “other”. The dictionary-type features were generated using the English and Swahili models using the TreeTagger tool (Schmid, 1994). They were binary features based on whether the word was recognized by the English tagger or the Swahili tagger. The final feature we explored was a regular expression designed to match Swahili phonology. Since Swahili orthography is highly regular and native Bantu vocabulary conforms strictly to certain phonological constraints, it was possible to write a regular expression that matches >95% of Swahili words, with the primary exceptions being words borrowed from Arabic.

We found that the Swahili regular expression was redundant with the use of character *n*-grams. Additionally, the English TreeTagger was highly overinclusive, marking many Swahili words as recognized, while the Swahili TreeTagger was underinclusive, making those features relatively weak. So we settled on using only the *n*-gram features along with the capitalization feature.

We then used the LIBLINEAR algorithm (Fan et al., 2008) with L2-regularization to generate context-free predictions over the words from the interview data. Punctuation tokens were excluded from this model since classifying them would be trivially easy. This context-free model was then used to expand the feature vector for each word. In addition to the original features, the generated probabilities for each class (English, Swahili, mixed, other) on the previous and following word were added to the feature vector. Punctuation was included as part of the context for generating these features. This achieved a high performance within our training set

[2] We discuss other aspects of the site in Section 6

[3] For example, Pets.com

Train / Test Set		Interview 10-fold CV		Interview JF Small		Intvw & JF Small JF Large	
Context Features		None	Word±1	None	Word±1	None	Word±1
English	Precision	94.2%	99.4%	41.6%	87.6%	90.1%	99.2%
	Recall	99.0%	99.7%	95.9%	96.6%	96.5%	98.8%
	F1 Score	96.5%	99.5%	58.0%	91.9%	93.2%	99.0%
Swahili	Precision	92.1%	97.9%	98.1%	99.0%	83.7%	95.3%
	Recall	67.0%	97.2%	62.4%	96.2%	64.1%	97.7%
	F1 Score	77.6%	97.5%	76.3%	97.6%	72.6%	96.5%
Accuracy		94.0%	99.3%	69.7%	96.5%	89.0%	98.4%
Cohen's Kappa		0.74	0.98	0.40	0.92	0.66	0.96

Table 2: Performance of Word-Level Language Identification Models

over a 10-fold cross-validation.

Next, we applied this model to a subset of the JamiiForums data. These labels were used to aid in annotating a portion of the forum data (JF Small). The 6,118 words annotated were then added to the training set and the resulting model was applied to an additional 16,475 words which were then hand-labeled (JF Large). The final model used all of the annotated data and was applied to the full 16+ million word JamiiForums data set.

4.2 Results

The results of the various iterations of the model are summarized in Table 3.2. We used Cohen's Kappa in addition to the measures of accuracy, precision, recall and F1 scores. Cohen's Kappa is used to measure inter-annotator agreement and is suitable for measuring performance across multiple classes and unbalanced label distributions. Effectively, we consider the model as one annotator and our annotators as the other. This measure is more robust across test sets with different label distributions, as is the case with the interview data, which is mostly English, and the JamiiForums data, which is balanced between English and Swahili.

As can be seen, the language probability scores of the word context improve performance significantly. Error analysis suggests that it primarily reduces the errors on named entities and numbers. Since we consider named entities and numbers as belonging to the language they're embedded in, it makes sense that these can sometimes only be correctly labeled using information about the context. But it also reduces errors on other words. For example, "wake" can be a word in both English and Swahili and context is necessary to disambiguate which language it is.

Overall, performance within the training set was highly accurate. The greater test was applying it to the out-of-domain forum data. As expected, performance decreased noticeably, with the context-dependent model going from 99.3% to 96.5% accuracy, and 0.98 Cohen's Kappa to 0.92. Nevertheless, this performance compares favorably to the performance of the systems in the FWCAC shared task on the out-of-domain "surprise" data (Solorio et al., 2014). There are several potential explanations for this. One obvious hypothesis is that the Swahili-English language pair is simply easier to distinguish than the language pairs in the shared task. English and Swahili are quite distinct phonologically; for example, Swahili words of Bantu stock universally end in vowels, so a final consonant is a strong indicator that a word is not Swahili. Another potential explanation is that our language label set was different and so the fact that we did not attempt to label named entities or ambiguous words explains the difference in performance. A final hypothesis is that using fewer features made our model more robust across domains. These explanations are difficult to disambiguate without direct comparisons of systems on similar data.

Error analysis on the JF Small set suggested that many of the errors were simply due to out-of-vocabulary *n*-grams. Our interview data included very few numerals and no symbols such as '&', since transcribers were instructed to write only the words as spoken. However, these characters are common in written communication. Rather than adjusting our feature set, we decided to add this annotated data to the training set and see how this im-

	Interviews	JamiiForums
# Codeswitches	8,508	922,547
Codeswitch %	4.5%	5.7%

Table 3: Codeswitch Point Statistics

proved performance. Adding the JF Small set to the interview data and testing on the JF Large set cut the error rate by over half and brought the Cohen's Kappa up to 0.96, almost as high as the performance within the training set. The accuracy of over 98% made us feel confident in applying this model to the full JamiiForums set, which would be used for the codeswitch point prediction task, discussed below.

5 Predicting Codeswitch Points

The second task was to explore how well we could predict whether a speaker or writer would codeswitch based on the language behavior prior to the current word. In this task, if the current word is word$_i$, then word$_i$ is labeled as a codeswitch point if word$_{i+1}$ is of a different language. Otherwise it is a non-switch point.

(1) Okay, *na unafikiria ni* important *kujua* native language?
(**Translation**: Okay, *and do you think it is* important *to know* native language?)

Consider example (1) from our interview data. At each word in the sentence, we want to predict whether a codeswitch will occur in the next word. The words "okay", "*ni*", "important" and "*kujua*" would all be codeswitch points. If the current word is "*ni*", we want to be able to predict that the next word "important" would be in English. For this, we use only the evidence available in the utterance to that point: "Okay, *na unafikiria ni*". This task is obviously more difficult with less evidence, as would be the case for the word "okay".

5.1 Methodology

As mentioned, we labeled the 16,176,057 words in the full JamiiForums data set using the language identification system we described above. The data used to train the language identification model was excluded from this set. In generating these labels for codeswitch points, we ignored punctuation. We then experimented with predicting codeswitch points using the Kenyan interview data and this much larger internet forum data set. The distribution of codeswitch points in our data sets can be seen in Table 5. The amount of codeswitching appears to be fairly similar, despite the difference in language distribution.

Previous approaches to this problem have used naïve Bayes classifiers trained using contextual and POS features (Solorio and Liu, 2008; Papalexakis et al., 2014). We explored a similar set of features, but additionally tried to represent a few other intuitions. The set of features for a potential codeswitch point at word$_i$ are shown in Table 5. In total, we explored eleven features. For features 6-10, we used binning to make the values more appropriate for the naïve Bayes algorithm. Features 1-3 and 11 were previously used in either Solorio and Liu (2008), Papalexakis et al. (2014), or both and found to have predictive value. Features 4 and 5 are meant as alternative versions of 2 and 3. The idea was that this can reduce sparsity in the data since a three-word sequence of Swahili generates the same values as a three-word sequence of English. Features 6-9 represent the intuition that the longer a speaker continues in a single language, the less likely a switch is at any particular point. We explored using a logarithmic scale since it seemed that after a long stretch of words in the same language, the likelihood of a codeswitch would not decrease much after a few more. The documents in our data have a large variation in length, as can be seen in Table 2.2. There are a number of very long documents in the JamiiForums data, which increases the range of values for these features. Finally, feature 10 is similar to features 6-9, but is not influenced by the length of the document. We had also explored using the POS taggers, as POS had been a useful feature in Solorio and Liu (2008). It did not provide an increase in performance on the interview data, and since applying the TreeTagger algorithm to 16 million words would have been very time-consuming, we did not explore it further. Other free POS taggers are not available for Swahili, nor could we find any large, easily accessible and POS-annotated Swahili corpus available to train our own.

Using these features, a naïve Bayes model was trained on the two data sets. In the unbalanced condition, this was done with a 10-fold cross-validation

Feature #	Feature Name	Description
1	lang_i	Language of word_i
2	lang_{i-1}	Language of word_{i-1}
3	lang_{i-2}	Language of word_{i-2}
4	$\text{match}(\text{lang}_{i-1})$	Are lang_i and lang_{i-1} the same?
5	$\text{match}(\text{lang}_{i-2})$	Are lang_i and lang_{i-2} the same?
6	# same lang words	# of words of lang_i in $\text{words}_{[0..i]}$
7	# diff lang words	# of words *not* of lang_i in $\text{words}_{[0..i]}$
8	log # same lang words	log_2(1+value(feature 6))
9	log # diff lang words	log_2(1+value(feature 7))
10	% same lang words	% of words of lang_i in $\text{words}_{[0..i]}$
11	Previous codeswitch	Did a codeswitch occur before word_i?

Table 4: Classification features for codeswitch point at word_i

on the full set. In the balanced condition, random samples of approximately 10,000 switch points and 10,000 non-switch points were taken from the data sets in a manner similar to Papalexakis et al. (2014). This allows a more direct comparison to both previous papers.

5.2 Results

The results of our prediction experiments are summarized in Table 5.2. The precision, recall and F1 score are for the codeswitch point class. Since our data is highly unbalanced, you could achieve an accuracy of 94% and 95% on our data sets by never predicting a codeswitch, so we also provide Cohen's Kappa which accounts for the label distribution.

The combination of features that worked best on our data was (1, 4, 5, 6, 9, 11). Reducing sparsity by making features 2 and 3 relative to the language of word_i appears slightly better. It is less clear why using the raw number of same language words worked better in combination with the logarithmic scale on different language words.

Performance on the two data sets is fairly similar despite the differing language distribution, the spoken vs. written domain, and the human-annotated vs. automatic language labels. This could indicate that English-Swahili codeswitching conventions are similar across these two domains. Relative to previous work on codeswitch prediction, our F1 score is similar but higher than in Solorio and Liu (2008) in the unbalanced condition. In the balanced condition, our F1 scores are similar to those reported in Papalexakis et al. (2014) on the interview data.

As mentioned earlier, codeswitching is never a forced choice (Solorio and Liu, 2008), so it would not be expected that these types of features could fully predict codeswitching behavior. In many sentences, there are multiple valid points at which one could codeswitch, and whether one does is informed by social considerations as well as grammatical constraints (Myers-Scotton, 1993a; Myers-Scotton, 1993b).

Given the important social component of codeswitching behavior, another avenue we would like to explore is the use of conversational features. Who someone is communicating with and why can also influence codeswitching behavior (Myers-Scotton, 1993b). Our interview data has a structure of questions and replies, making it possible to examine the influence of previous utterances. The JamiiForums data also has structure in the forum threads, although who is talking to whom is not always obvious since it does not use a nested reply structure. We discuss some of these future directions in the next section.

6 Discussion & Future Directions

As mentioned in our introduction, one of the motivations for predicting codeswitch points is that it could be used to aid in a social analysis of codeswitching behavior. Knowledge of when a decision to codeswitch – or not to – is more or less likely can mark such a decision as more or less meaningful. If the other participant in a conversation has been engaging in frequent codeswitching, this may generally lead the speaker to engage in more codeswitching. If the speaker does not accommodate to their interlocutor, it can give insight into the

Measure	Interviews		JamiiForums	
	unbal	bal	unbal	bal
Model accuracy	97.5%	74.4%	96.9%	67.4%
Precision	28.5%	78.3%	27.4%	81.4%
Recall	52.2%	72.6%	51.3%	58.1%
F1 Score	36.8%	75.3%	35.7%	67.8%
Cohen's Kappa	0.327	0.524	0.306	0.448

Table 5: Codeswitch point prediction performance

social relationship between the participants. For example, Myers-Scotton (1993b) noted that power relations can be reflected in codeswitching behavior. If one participant is more powerful, they may be able to control the language of the interaction and how much language mixing is allowed within it. Other computational studies of linguistic accommodation have found correlations between with power relations (Danescu-Niculescu-Mizil et al., 2012). It is likely that accommodation in codeswitching behavior would follow similar patterns.

The data sets we have collected have metadata that can be used for such social analyses. The Kenyan interview data set has, in addition to conversational structure, pairs of interviews of the same students conducted by another student and by a professor, creating a difference in power across those conditions. The JamiiForums data has less explicit power relations to exploit, with the only easily recognizable hierarchy being between regular users, moderators and administrators, and only a handful of members fall into the latter categories. However, social closeness can be represented by certain interactions between users, such as quotation replies, "liking" each others posts, and following other users. Exploiting these social relations to inform our analysis could yield improvements in our prediction of codeswitching behavior. It would also be possible to track changes in behavior over time, given the decade-long history of the site. An important note about this data set is that while we have approximately 220,000 posts in our collection, the full JamiiForums site has over 17 million posts and an estimate of over 1 billion words. Gaining access to the full data set would increase the scale of the corpus by a couple of orders of magnitude.

Finally, going beyond analyzing patterns of codeswitching to interpreting individual instances of codeswitching will require a finer-grained analysis. For example, differentiating between quotative, parenthetical and emphatic uses of codeswitching requires not merely an estimate of how expected a codeswitch is in that position, but some understanding of the semantics of the language used. While some uses may be easier to distinguish (codeswitching to quote someone is likely to be preceded by a quotative verb, for example), interpreting the socio-pragmatic meaning of codeswitching will generally be far more difficult. Distinguishing between what Myers-Scotton (1993b) refers to as "codeswitching as the unmarked choice" and marked uses of codeswitching is an important first step in that direction.

7 Conclusion

In this paper, we built models for language identification and to predict codeswitching using Swahili-English data. This is, to our knowledge, the first computational paper addressing Swahili codeswitching. We achieved a high accuracy on the language identification task, and modest improvement on the codeswitch point prediction task.

Future directions of study will focus on social analyses of codeswitching behavior, such as the connection between power and linguistic accommodation, or codeswitching and social solidarity. Further work can be done on the Kenyan interview data, while the language identification model will enable analysis of other aspects of codeswitching within the large JamiiForums corpus.

References

Gokul Chittaranjan, Yogarshi Vyas, Kalika Bali, and Monojit Choudhury. 2014. Word-level language identification using CRF: Code-switching shared task report of MSR India system. In *Proceedings of The First Workshop on Computational Approaches to Code*

Switching, pages 73–79. Association for Computational Linguistics.

Cristian Danescu-Niculescu-Mizil, Lillian Lee, Bo Pang, and Jon Kleinberg. 2012. Echoes of power: Language effects and power differences in social interaction. In *Proceedings of the 21st International Conference on World Wide Web*, WWW '12, pages 699–708, New York, NY, USA. ACM.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874.

Oluwapelumi Giwa and Marelie H. Davel. 2013. N-gram based language identification of individual words. In *Proceedings of the Twenty-Fourth Annual Symposium of the Pattern Recognition Association of South Africa*, pages 15–22. Association for Computational Linguistics.

Chu-Cheng Lin, Waleed Ammar, Lori Levin, and Chris Dyer. 2014. The CMU submission for the shared task on language identification in code-switched data. In *Proceedings of The First Workshop on Computational Approaches to Code Switching*, pages 80–86. Association for Computational Linguistics.

Alamin Mazrui. 1995. Slang and codeswitching: The case of Sheng in Kenya. *Afrikanistische Arbeitspapiere*, 42:168–179.

Carol Myers-Scotton. 1993a. *Duelling Languages: Grammatical Structure in Codeswitching*. Oxford University Press, Oxford, UK.

Carol Myers-Scotton. 1993b. *Social Motivations for Codeswitching: Evidence from Africa*. Oxford University Press, Oxford, UK.

Dong Nguyen and A. Seza Doğruöz. 2013. Word level language identification in online multilingual communication. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 857–862. Association for Computational Linguistics.

Evangelos E. Papalexakis, Dong Nguyen, and A. Seza Doğruöz. 2014. Predicting code-switching in multilingual communication for immigrant communities. In *Proceedings of The First Workshop on Computational Approaches to Code Switching*, pages 42–50. Association for Computational Linguistics.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*.

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

Thamar Solorio, Elizabeth Blair, Suraj Maharjan, Steve Bethard, Mona Diab, Mahmoud Ghoneim, Abdelati Hawwari, Fahad AlGhamdi, Julia Hirshberg, Alison Chang, and Pascale Fung. 2014. Overview for the first shared task on language identification in codeswitched data. In *Proceedings of The First Workshop on Computational Approaches to Code Switching*, pages 62–72. Association for Computational Linguistics.

Tommi Vatanen, Jaakko J. Väyrynen, and Sami Virpioja. 2010. Language identification of short text segments with n-gram models. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, pages 3423–3430. European Language Resources Association (ELRA).

Martin Volk and Simon Clematide. 2014. Detecting code-switching in a multilingual Alpine heritage corpus. In *Proceedings of The First Workshop on Computational Approaches to Code Switching*, pages 24–33. Association for Computational Linguistics.