Comparing the Level of Code-Switching in Corpora

Björn Gambäck*, **Amitava Das**†
*Department of Computer and Information Science
Norwegian University of Science and Technology
NO–7491 Trondheim, Norway
gamback@idi.ntnu.no

†Indian Institute of Information Technology
Sri City, Satyavedu Mandal, Chittoor District
Andhra Pradesh – 517588, India
amitava.das@iiits.in

Abstract

Social media texts are often fairly informal and conversational, and when produced by bilinguals tend to be written in several different languages simultaneously, in the same way as conversational speech. The recent availability of large social media corpora has thus also made large-scale code-switched resources available for research. The paper addresses the issues of evaluation and comparison these new corpora entail, by defining an objective measure of corpus level complexity of code-switched texts. It is also shown how this formal measure can be used in practice, by applying it to several code-switched corpora.

Keywords: Code-switching, Evaluation, Corpora, Social Media Text

1. Introduction

When two individuals who are bi- or multi-lingual in an overlapping set of languages communicate, they tend to switch seemlessly and effortlessly between the languages (codes) they share. When this code alternation occurs at or above the utterance level, the phenomenon is referred to as code-switching; when the alternation is utterance-internal, the term 'code-mixing' is common, even though 'code-switching' is frequently used in those cases as well. Code-mixing in itself is often an effect of what recently (particularly within language teaching) has started to be called 'translanguaging', that is, when truly bi-lingual individuals are creating new meanings based on their full and double language repertoire (Lewis et al., 2012).

Code-switching is most prominent in spoken language conversations and has thus traditionally mainly been studied by psycho- and sociolinguists (Auer, 1999; Muysken, 2000; Gafaranga and Torras, 2002; Bullock et al., 2014) and by speech researchers (Lyu et al., 2015), while the lack of large-scale textual corpora has made code-switching less attractive as a subject of study in computational or corpora linguistics. However, this started changing in 2003 with the advent of social media, where large amounts of texts are written that are more informal and more conversational in nature, and hence when produced by bilinguals tend to contain more code-switching (Paolillo, 1996).

This new availability of large-scale code-switched resources in turn raises questions of evaluation: how do we compare the results of applying language processing tools to one code-switched corpus to those on another? Or more specifically: how can we compare the level of code-switching in corpora? And for corpora containing a mix of a specific set of languages or across corpora from different sets of languages? These are the issues that the present paper aims to address.

That two texts come from social media does not in itself imply that they belong to *one*, delimited textual domain. Rather, there is a wide spectrum of different types of texts that are transmitted through social media, and the level of formality of the language in addition depends more on the style of the writer than on the actual media (Eisenstein, 2013; Androutsopoulos, 2011). They both argue that the common denominator of social media text is not that it is 'noisy' and informal *per se*, but that it describes language in (rapid) change. Furthermore, although social media often convey more ungrammatical text than more formal writings, Baldwin et al. (2013) have shown that the relative occurrence of non-standard syntax is fairly constant among many types of media, such as mails, tweets, forums, comments, and blogs.

Due to the ease of availability of Twitter, most research on social media text has so far focused on tweets (Twitter messages). Lui and Baldwin (2014) note that users that mix languages in their writing still tend to avoid code-switching inside a specific tweet, a fact that has been utilized to investigate which language is dominant in a tweet (Carter, 2012; Lignos and Marcus, 2013; Voss et al., 2014). However, tweets still tend to be somewhat formal by more often following grammatical norms and using standard lexical items (Hu et al., 2013), while chats are more conversational (Paolillo, 1999), and hence less formal, which tend to increase their level of code-switching (Cárdenas-Claros and Isharyanti, 2009; Paolillo, 2011; Nguyen and Doğruöz, 2013; Das and Gambäck, 2014).

The paper is organized as follows: Section 2. describes a formal measure that can be used to compare the complexity of code-switched corpora. Section 3. then uses this corpus level switching measure in practise, applying it to a set of recently produced code-switched corpora. Finally, Section 4. sums up and elaborates on the results.

2. Measuring Code-Switching in Corpora

When comparing different code-switched corpora to each other, it is desirable to have a measurement of the level of mixing between languages, in particular since error rates for various language processing application would be expected to increase as the level of code-switching increases. Both Kilgarriff (2001) and Pinto et al. (2011) discussed several statistical measures that can be used to compare corpora more objectively, but those measures presume that the corpora are essentially monolingual.

Debole and Sebastiani (2005) analysed the complexity of the different subsets of the Reuters-21578 corpus in terms of the relative hardness of learning classifiers on the subcorpora, a strategy which does not assume monolinguality in the corpora. However, they were only interested in the relative difficulty and give no measure of the complexity as such. In Gambäck and Das (2014) we instead suggested an initial Code-Mixing Index to assess the level of code-switching in an utterance. This measure will be taken as the starting point, and elaborated on here.

2.1. Utterance Level Switching

If an utterance x only contains language independent tokens, its code-mixing is zero; for other utterances, the level of mixing depends on the fraction of language dependent tokens that belong to *the matrix language* (the most frequent language in the utterance) and on N, the number of tokens in x except the language independent ones (i.e., all tokens that belong to any language L_i):[1]

$$C_u(x) = \begin{cases} \frac{N(x) - \max\limits_{L_i \in \mathbb{L}} \{t_{L_i}\}(x)}{N(x)} & : N(x) > 0 \\ 0 & : N(x) = 0 \end{cases} \qquad (1)$$

($L_i \in \mathbb{L}$, the set of all languages in the corpus; $1 \leq \max\{t_{L_i}\} \leq N$). Notably, for mono-lingual utterances $C_u = 0$ (since then $\max\{t_{L_i}\} = N$).[2]

This initial measure has several short-comings. In particular, it does not reflect what fraction of a corpus' utterances contain code-switching, nor take into account the number of code alternation points: arguably, a higher number of language switches in an utterance increases its complexity, while a corpus with a larger fraction of mixed utterances is (on average) more complex.[3]

Two main sources of information will be utilized to fully account for the code alternation at utterance level: the ratio of tokens belonging to the matrix language ($f_m = [N - max\{t_{L_i}\}]/N$ as in Equation 1) and the number of code alternation points per token ($f_p = P/N$, where P is the number of code alternation points; $0 \leq P < N$).

There are many ways to combine two (or several) information sources, in particular if they are independent; see, e.g., Genest and McConway (1990) for an overview. However, P partially depends on $\max\{t_{L_i}\}$,[4] which, for example, rules out the common *logarithmic opinion poll*:

$$p(x) = \prod_{k=1}^{n} p_k(x)^{w_k} \qquad : \sum_k w_k = 1 \qquad (2)$$

Instead we will use the *linear opinion poll:*

$$p(x) = \sum_{k=1}^{n} w_k \times p_k(x) \qquad : \sum_k w_k = 1 \qquad (3)$$

Combining $f_m(x)$ and $f_p(x)$ gives a revised utterance level measure for $N(x) > 0$:

$$\begin{aligned} C_u(x) &= w_m f_m(x) + w_p f_p(x) \qquad (4) \\ &= w_m \frac{N(x) - \max\limits_{L_i \in \mathbb{L}} \{t_{L_i}\}(x)}{N(x)} \cdot 100 + w_p \frac{P(x)}{N(x)} \cdot 100 \\ &= 100 \cdot \frac{w_m \Big(N(x) - \max\limits_{L_i \in \mathbb{L}} \{t_{L_i}\}(x)\Big) + w_p P(x)}{N(x)} \end{aligned}$$

where w_m and w_p are weights ($w_m + w_p = 1$). Again, $C_u = 0$ for mono-lingual utterances (since then $\max\{t_{L_i}\} = N$ and $P = 0$).

2.2. Corpus Level Switching

Moving to corpus level, the measure could be defined simply as average utterance level switching, as in Equation 5

[1] Note that the formula in Equation 1 differs from, but is equivalent to, the one given in Gambäck and Das (2014).

[2] Consider, e.g., an utterance U_1 with 10 words. If 5 of the words come from language L_1 and the other from language L_2, its C_u will be $(10 - 5)/10 = 0.50$. However, another 10-word utterance U_2 with all words coming from different languages gets $C_u(U_2) = (10 - 1)/10 = 0.90$, correctly reflecting the intuition that U_2 presents a more complex mix.

[3] Compare two 4-word utterances U_3 and U_4 with 2 words each from the languages L_1 and L_2. Thus $C_u(U_3) = C_u(U_4) = (4 - 2)/4 = 0.50$. But if U_3 only contains 1 code alternation point (e.g., if the words are $w_{L_1} w_{L_1} w_{L_2} w_{L_2}$), while U_4 contains 3 switches (e.g., $w_{L_1} w_{L_2} w_{L_1} w_{L_2}$), then U_4 will most likely be more difficult to analyse.

[4] In utterance U_2 with 10 words, all from different languages ($\max\{t_{L_i}\} = 1$), there must be a code alternation between each word, so $P = 9$. If instead $\max\{t_{L_i}\} = 2$, then $4 \leq P \leq 9$, since the utterance, e.g., can contain 2-token sequences from five languages ($P = 4$) or ten 1-token sequences from up to eight different languages.

$$C_{avg} = \frac{1}{U}\sum_{x=1}^{U} C_u(x) \tag{5}$$

where U is the number of utterances in the corpus.

However, that would ignore two important points: that C_u does not account for code-alternation *between* two utterances, and that the frequency of code-switched utterances in a corpus increases its complexity.

Hence, when combining several utterances, an utterance's matrix language and the matrix language of the previous utterance need to be represented (if they differ, that implies adding a code-alternation point between the two utterances).[5] For each pair of utterances, a factor must be included to account for this, as shown in Equation 6:

$$C_u(x{-}1,x) = C_u(x{-}1) + C_u(x) + w_p\delta(x) : \left\{ \begin{array}{c} L_{x\text{-}1} = \max\limits_{L_i \in \mathbb{L}}\{t_{L_i}(x{-}1)\} \\ L_x = \max\limits_{L_i \in \mathbb{L}}\{t_{L_i}(x)\} \\ \delta(x) = \begin{cases} 0 : x{=}1 \lor L_{x\text{-}1} = L_x \\ 1 : x{\neq}1 \land L_{x\text{-}1} \neq L_x \end{cases} \end{array} \right\} \tag{6}$$

For combining a corpus' all utterances, we take inspiration from readability indices that are purely word frequency-based and (as C_u), e.g., make no distinction between different word classes. Those are calculated using the average sentence length and another factor, e.g., the average number of syllables per word as in the 'Reading Ease' score (Flesch, 1948), the frequency of multi-syllabic words in 'Fog' (Gunning, 1952), or the frequency of long words in 'LIX' (Björnsson, 1968).

Flesch' Reading Ease score is based on the average number of words per sentence and average number of syllables per word:

$$\mathrm{RE} = 206.835 - [1.015 \cdot (\frac{\mathrm{W}}{\mathrm{S}}) + 84.6 \cdot (\frac{\mathrm{L}}{\mathrm{W}})] \tag{7}$$

where W is the number of words in the text, S the total number of sentences, and L the total number of syllables (hence words/sentence are weighted as $1.2 \cdot$syllables/word).

The Fog Index is the number of words per sentence plus the percentage of multi-syllabic words:

$$\mathrm{Fog} = 0.4 \cdot [\frac{\mathrm{W}}{\mathrm{S}} + 100 \cdot (\frac{\mathrm{F}}{\mathrm{W}})] \tag{8}$$

where W is the number of words in the text, S the number of sentences, and F the number of "foggy" words, that is, mainly words with more than three syllables.

The LIX measurement is the number of words per sentence plus the percentage of long words:

$$\mathrm{LIX} = \frac{\mathrm{W}}{\mathrm{S}} + 100 \cdot (\frac{\mathrm{L}}{\mathrm{W}}) \tag{9}$$

where W is the number of words in the text, S the number of sentences, and L the number of long words (defined as words with more than five characters).

In the case of code-switching, the first factor is the average switching level per utterance, as calculated by inserting the C_u given by Equation 6 into the average of Equation 5, while the second factor is the frequency of utterances that contain any code-switching (i.e., utterances with $C_u > 0$). Thus arriving at Equation 10:

$$\begin{aligned} C_c &= \frac{\sum\limits_{x=1}^{U} C_u(x) + w_p\delta(x)}{U} + w_s\frac{S}{U} \cdot 100 \\ &= \frac{100}{U}\left[\sum_{x=1}^{U}\Big(w_m f_m(x) + w_p\big[f_p(x) + \delta(x)\big]\Big) + w_s S\right] \end{aligned} \tag{10}$$

where S is the number of utterances that contain code-switching ($0 \leq S \leq U$), and w_s the relative weight attached to the switching frequency.

3. Comparing Corpora Level Switching

The main issue when applying an information source combination method (e.g., Equation 2 or 3) is how to choose the weights, and several strategies have been proposed. We tried a number of them experimentally at the utterance level, but the only combination giving reliable and intuitive values was the average (equal weights: $w_k = \frac{1}{n}$) reflecting an observation also made by Clemen (2008, p.765): "*Having spent much of my career studying various combination methods, it has been somewhat frustrating to consistently find that the simple average performs so well empirically.*"

With two information sources, the weights are $w_m = w_p = \frac{1}{2}$ and Equation 4 (for $N > 0$) reads:

$$C_u(x) = 100 \cdot \frac{N(x) - \max\limits_{L_i \in \mathbb{L}}\{t_{L_i}\}(x) + P(x)}{2N(x)} \tag{11}$$

Similarly, for determing the relative weight attached to the switching frequency at the corpus level (w_f in Equation 10), we again compare to readability indices. Fog and LIX combine two information sources without weighting (i.e., indirectly use the average); however, Reading Ease applies a weighting which treats the {words/sentence} factor as 1.2 × {syllables/word}. Flesch (1948) derived this through regression based on correlations between the average grade of children and those who could answer 50% and 75% of some test questions.

[5] This is different from, and more important than, checking whether the language of an utterance's first token differs from that of the previous utterance's last token.

Language Pair	words	utterances (U)	switched (S)	 (%)	C_{avg} (U)	 (S)	P_{avg} (U)	 (S)	δ (U)	C_c
DU – TR	70,874	3,065	382	12.46	4.11	**33.02**	0.23	1.88	**48.87**	14.50
EN – HI	27,167	2,583	570	22.07	1.87	8.46	0.57	2.56	17.81	20.26
EN – ES	140,746	11,400	2,335	20.48	4.91	23.95	0.38	1.84	13.83	21.97
EN – ZH	17,430	999	322	32.23	4.19	13.01	0.70	2.18	22.32	31.06
EN – NE	146,056	9,993	4,926	**49.29**	**7.98**	16.19	**1.52**	**3.08**	35.18	**49.06**
ARB – ARZ	119,317	5,839	931	15.94	3.77	23.67	0.19	1.20	13.29	17.06

Table 1: Code-switching levels in some corpora

Using these weights, Equation 10 becomes

$$C_c = \frac{100}{U}\left[\tfrac{1}{2}\sum_{x=1}^{U}\Big(f_m(x)+f_p(x)+\delta(x)\Big)+\tfrac{5}{6}S\right] \qquad (12)$$

$$= \frac{100}{U}\left[\tfrac{1}{2}\sum_{x=1}^{U}\Big(1-\frac{\max\limits_{L_i\in\mathbb{L}}\{t_{L_i}\}(x)+P(x)}{N(x)}+\delta(x)\Big)+\tfrac{5}{6}S\right]$$

To show how the measure can be used in practice to objectively compare the complexity of code-switching, C_c values as in Equation 12 were calculated for some recently produced code-switched corpora: the Dutch-Turkish chat corpus of Nguyen and Doğruöz (2013), the English-Hindi Twitter and Facebook chat corpus of Jamatia et al. (2015), and the four corpora[6] used in the shared task on word-level language detection in code-switched text (Solorio et al., 2014) organized by the workshop on Computational Approaches to Code Switching at the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP).

When comparing C_c values for different corpora, it is necessary to consider their respective tagsets and annotation guidelines. So does the annotation strategy chosen for the EMNLP corpora prescribe that elements such as abbreviations should be tagged with the language they belong to, while other annotations schemes treat them as language independent. Another potential problem can be to decide whether a tag is directly language related or not. The EMNLP tagset includes the tags 'mixed' and 'ambiguous', that here are treated as language items in the calculations, but without being assigned to any specific language (when selecting an utterance's matrix language), which follows the EMNLP annotation guidelines.

Table 1 shows statistics and C_c values for the corpora. The first columns give the number of words and total number of utterances (U) in each corpus, followed by the number and percentage of the utterances that really contain any code-switching (S). The second set of columns provide the C_{avg} values over both all the utterances and over only the utterances that actually contain code-switching, followed by the average number of intra-utterance code-alternation points (P) for the same two sets of utterances (the total and those containing switching), and finally the frequency of inter-utterance switching (δ), i.e., switching of matrix language between two utterances. The last column gives the actual C_c value for each corpus.

It is noticable that the EN-NE corpus from EMNLP exhibits the highest level of code-switching, both at corpus level ($C_c = 49.06$) and on average at utterance level ($C_{avg} = 7.98$ for all utterances, U), as well as highest average number of code-alternation points per utterance ($P_{avg} = 1.52$; for those utterances that contain switching: $P_{avg} = 3.08$), while DU-TR has the lowest C_c value, but the highest frequency of matrix language switching between utterances (there are $1,498$ switches for $3,065$ utterances), and also the highest utterance level switching ($C_{avg} = 33.02$) if counting the average over those utterances that contain switching (S).

4. Discussion and Conclusion

The paper has defined an objective measure of the complexity of code-switched texts, i.e., texts written in several different languages, something which is particularly common in social media. Certainly, though, no such measure will ever be able to capture all types of differences between corpora. In particular, the ways corpora were collected and annotated, and their intended usage also need to be taken into account. However, levelling out such differences should arguably not be the aim of the code-switching measure itself, but rather be left to the users: when comparing corpora with widely different scopes, the users themselves need to be aware of the potential variation and consider this when deciding on whether a straight-forward comparison really makes sense.

The English-Nepalese EMNLP corpus showed an extremely high level of switching and the Mandarin Chinese a fairly high level. This could of course possibly have been caused by errors and problems in tagging, but in contrast to other corpora, the tagset used in the EMNLP shared task included a tag for words that are ambiguous in a context (i.e., words that even given the contextual information could potentially belong to two or more of the languages in the corpus), which potentially should ease the annotation task.

It is important to keep in mind that the code-switching corpus complexity measurement is intended to be independent of the languages contained in the corpus, while the per-

[6] The EMNLP corpora mix English with Spanish, Mandarin Chinese and Nepalese. The forth EMNLP corpus is dialectal: Standard Arabic mixed with Egyptian Arabic (ARB-ARZ).

formance of a language processing system of course also will depend on the actual languages, their relationship, and their annotation schemes. Thus the reported level of mixing in the ARB-ARZ corpus is quite low (only 15.94%, with $C_c = 17.06$), but the "languages" involved are both Arabic dialects, so very closely-related, and hence the potential overlap between them is high, even if that has not been reflected in the annotation. (Notably, almost all words of the "standard" language will also belong to a dialect, while the opposite relation does not hold. So only utterances containing strictly Egyptian Arabic words and expressions would be expected to be annotated as ARZ in this case, and all others as ARB, Modern Standard Arabic.)

For this reason, the dialectal Arabic corpus actually was the one causing most problems for the processors in the EMNLP 2014 shared task on code-switched language identification. On the other hand, the corpus with the highest C_c (EN-NE) was the second easiest one to label for the systems participating in the shared task (the language pair which was the easiest to separate was Mandarin–English, although not for linguistic reasons, but simply since the two languages were written in different scripts).

5. Acknowledgements

The EMNLP shared task corpora are public domain, as tweet IDs. The Dutch-Turkish corpus has gratefully been made available by the researchers involved. Interested readers should contact: Dong Nguyen (U Twente) or Seza Doğruöz (Tilburg U). Many thanks to them and to the EMNLP shared task organisers (Solorio et al., 2014), as well as to Anupam Jamatia, Kunal Chakma and Dwijen Rudrapal (all at NIT Agartala).

6. Bibliographical References

Androutsopoulos, J. (2011). Language change and digital media: a review of conceptions and evidence. In Tore Kristiansen et al., editors, *Standard Languages and Language Standards in a Changing Europe*, pages 145–159. Novus, Oslo, Norway, February.

Auer, P. (1999). From codeswitching via language mixing to fused lects: Toward a dynamic typology of bilingual speech. *International Journal of Bilingualism*, 3(4):309–332.

Baldwin, T., Cook, P., Lui, M., MacKinlay, A., and Wang, L. (2013). How noisy social media text, how diffrnt social media sources? In *Proceedings of the 6th International Joint Conference on Natural Language Processing*, pages 356–364, Nagoya, Japan, October. AFNLP.

Björnsson, C.-H. (1968). *Läsbarhet*. Liber, Stockholm, Sweden. (in Swedish).

Bullock, B. E., Hinrichs, L., and Toribio, A. J. (2014). World Englishes, code-switching, and convergence. In Markku Filppula, et al., editors, *The Oxford Handbook of World Englishes*. Oxford University Press, Oxford, England. Forthcoming. Online publication: March 2014.

Cárdenas-Claros, M. S. and Isharyanti, N. (2009). Code switching and code mixing in internet chatting: between 'yes', 'ya', and 'si' a case study. *Journal of Computer-Mediated Communication*, 5(3):67–78.

Carter, S. (2012). *Exploration and Exploitation of Multilingual Data for Statistical Machine Translation*. PhD Thesis, University of Amsterdam, Informatics Institute, Amsterdam, The Netherlands, December.

Clemen, R. T. (2008). Comment on Cooke's classical method. *Reliability Engineering & System Safety*, 93(5):760–765, May.

Das, A. and Gambäck, B. (2014). Identifying languages at the word level in code-mixed Indian social media text. In *Proceedings of the 11th International Conference on Natural Language Processing*, pages 169–178, Goa, India, December.

Debole, F. and Sebastiani, F. (2005). An analysis of the relative hardness of Reuters-21578 subsets. *Journal of the American Society for Information Science and Technology*, 58(6):584–596, April.

Eisenstein, J. (2013). What to do about bad language on the internet. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 359–369, Atlanta, Georgia, June. ACL.

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology*, 32(3):221–233, June.

Gafaranga, J. and Torras, M.-C. (2002). Interactional otherness: Towards a redefinition of codeswitching. *International Journal of Bilingualism*, 6(1):1–22.

Gambäck, B. and Das, A. (2014). On measuring the complexity of code-mixing. In *Proceedings of the 11th International Conference on Natural Language Processing*, pages 1–7, Goa, India, December. 1st Workshop on Language Technologies for Indian Social Media.

Genest, C. and McConway, K. J. (1990). Allocating the weights in the linear opinion pool. *Journal of Forecasting*, 9(1):53–73, Jan/Feb.

Gunning, R. (1952). *The Technique of Clear Writing*. McGraw-Hill, New York, New York.

Hu, Y., Talamadupula, K., and Kambhampati, S. (2013). *Dude, srsly?*: The surprisingly formal nature of Twitter's language. In *Proceedings of the 7th International Conference on Weblogs and Social Media*, Boston, Massachusetts, July. AAAI.

Jamatia, A., Gambäck, B., and Das, A. (2015). Part-of-speech tagging for code-mixed English-Hindi Twitter and Facebook chat messages. In *Proceedings of the 10th International Conference on Recent Advances in Natural Language Processing*, pages 239–248, Hissar, Bulgaria, September.

Kilgarriff, A. (2001). Comparing corpora. *International Journal of Corpus Linguistics*, 6(1):97–133.

Lewis, G., Jones, B., and Baker, C. (2012). Translanguaging: origins and development from school to street and beyond. *Educational Research and Evaluation*, 18(7):641–654, October.

Lignos, C. and Marcus, M. (2013). Toward web-scale analysis of codeswitching. In *87th Annual Meeting of the Linguistic Society of America*, Boston, Massachusetts, January. Poster.

Lui, M. and Baldwin, T. (2014). Accurate language identification of twitter messages. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 17–25, Göteborg, Sweden, April. ACL. 5th Workshop on Language Analysis for Social Media.

Lyu, D.-C., Tan, T.-P., Chng, E.-S., and Li, H. (2015). Mandarin–English code-switching speech corpus in South-East Asia: SEAME. *Language Resources and Evaluation*, 49(3):581–600, September.

Muysken, P. (2000). *Bilingual speech: A typology of code-mixing*. Cambridge University Press, Cambridge, England.

Nguyen, D. and Doğruöz, A. S. (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, Seattle, Washington, October. ACL.

Paolillo, J. (1996). Language choice on soc.culture.punjab. *Electronic Journal of Communication*, 6(3), June.

Paolillo, J. (1999). The virtual speech community: Social network and language variation on IRC. *Journal of Computer-Mediated Communication*, 4(4), June.

Paolillo, J. (2011). "conversational" codeswitching on usenet and internet relay chat. *Language@Internet*, 8(article 3), June.

Pinto, D., Rosso, P., and Jiménez-Salazar, H. (2011). A self-enriching methodology for clustering narrow domain short texts. *The Computer Journal*, 54(7):1148–1165, July.

Solorio, T., Blair, E., Maharjan, S., Bethard, S., Diab, M., Gohneim, M., Hawwari, A., AlGhamdi, F., Hirschberg, J., Chang, A., and Fung, P. (2014). Overview for the first shared task on language identification in code-switched data. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 62–72, Doha, Qatar, October. ACL. 1st Workshop on Computational Approaches to Code Switching.

Voss, C., Tratz, S., Laoudi, J., and Briesch, D. (2014). Finding romanized Arabic dialect in code-mixed tweets. In *Proceedings of the 9th International Conference on Language Resources and Evaluation*, pages 188–199, Reykjavík, Iceland, May. ELRA.