# On Measuring the Complexity of Code-Mixing

**Björn Gambäck**
Norwegian University of Science and Technology
Trondheim, Norway
gamback@idi.ntnu.no

**Amitava Das**
University of North Texas
Denton, Texas, USA
amitava.santu@gmail.com

## Abstract

The paper discusses the practical applicability of a Code-Mixing Index as a measurement of the level of complexity and mixing in texts written in several different languages, and contrasts it to other ways of measuring the complexity of texts. In particular, we describe the application of the proposed Index to corpora of code-mixed Indian social media texts and compare their complexity to social media texts for other language pairs.

## 1 Introduction

Code-mixing and code-switching causes problems for many language processing systems that are based on particular language models. In order to select the correct model for a particular text, the language of the text is often assumed to be known *a priori*. However, if the text consists of sections written in several different languages, the model selection process gets substantially more complex.

The style of writing and the concise texts in social media further complicate the issue. Longer documents in other text genres tend to have fewer code-switching points, that often are caused by loan words or author shifts, and hence tend to occur at the sentence level or higher (i.e., inter-sentential switching). In contrast, code-switching in social media texts is more commonly caused by the writer (and reader) being bi- or multilingual and thus simply swapping as effortlessly between the languages in writing as the same person(s) would do when speaking. Notably, this type of code-switching can often appear at the word level, or even lower (i.e., word-internal).

Obviously, code-switching is more common in geographical regions with a high percentage of bilingual individuals, such as in Texas and California in the US, Hongkong and Macao in China, many European and African countries, and the countries in South-East Asia. Multi-linguality (and hence code-switching) is very common in India, that has close to 500 spoken languages (or over 1600, on some accounts), with about 30 languages having more than 1 million speakers. Language diversity and dialect changes trigger Indians to frequently change and mix languages, in particular in speech and in social media contexts.

In the present paper, 'code-mixing' will be the term mainly used for these type of phenomena and in particular taken to refer to intra-sentential switching, that is, to the cases where the language change occurs inside a sentence. The term 'code-switching' is equally common in the literature (Auer, 1999; Muysken, 2000; Gafaranga and Torras, 2002; Bullock et al., 2014), but here we will take it as specifically referring to inter-sentential switching.

Previous work and the nature of code-mixing in social media text will be the topic of the next section. Section 3 then introduces the Code-Mixing Index, while Section 4 shows how it can be applied in practice to give a measure of the level of multilinguality in a corpus. Section 5 discusses the implications of using this index and compares it to some other ways of measuring the complexity of texts. Finally, Section 6 sums up the paper and gives suggestions for how the Code-Mixing Index could be further extended and utilised.

## 2 Code-Switching in Social Media Text

Before turning to the topic of complexity of code-mixing, we first briefly discuss studies on the general characteristics of code-mixing and code-switching in social media text, in particular the types and the frequencies of code-mixing, the language levels it appears at, and the reasons for it to appear in the first place.

Regarding the language level, San (2009) reports on a predominance of inter-sentential code-switching in blog posts and Hidayat (2012) similarily claims intra-sentential switching to account for 33% of the code-switching in Facebook messages, with 59% of the switching being inter-sentential switching. In contrast, Das and Gambäck (2014) report on code-switching in Facebook messages by Indian students writing in mixed English-Bengali or English-Hindi, and state that intra-sentential switching accounts for 60 resp. 55% of the switching in those language pairs, with inter-sentential switching only accounting for 32 resp. 37% of the code-switching.

Other studies have looked at code-mixing in different types of short texts, such as information retrieval queries (Gottron and Lipka, 2010) or short text messages (Rosner and Farrugia, 2007). Lui and Baldwin (2014) note that users that mix languages in their writing still tend to avoid code-switching inside a specific Twitter message, a fact that both Carter (2012) and Lignos and Marcus (2013) utilize to investigate which language is the dominant one in a specific message, with the evaluation taking place at the post (tweet) level. However, this is not the case for chat messages, as indicated by the work of Nguyen and Doğruöz (2013) on mixed Turkish-Dutch chat posts, as well as by Das and Gambäck (2014) on Hindi/Bengali-English Facebook chat groups.

There has also been work on analysing code-switching in spoken language; however, this has mainly been on artificially generated speech data (Chan et al., 2009; Solorio et al., 2011; Weiner et al., 2012), an exception being the small (129 intra-sentential language switches) spoken Spanish-English corpus introduced by Solorio and Liu (2008).

Clearly, there are (almost) as many reasons for why people code-switch as there are people code-switching. However, several studies of code-mixing in different type of social media texts indicate that it primarily is triggered by a need in the author to mark some in-group membership. This conclusion was reached by Sotillo (2012) and by Xochitiotzi Zarate (2010) who both analysed English-Spanish short text messages; by Bock (2013) looking at chat messages in English, Afrikaans and isiXhosa; by Shafie and Nayan (2013) in Facebook comments in Bahasa Malaysia and English; as well as by Negrón Goldbarg (2009) in a small study of code-switching in Spanish-English emails. In contrast, studies on Chinese-English code-mixing in Hong Kong by Li (2000) and in Macao by San (2009) indicate that the code-switching there mainly is triggered by linguistic motivations, with social motivations being less salient.

A common short-coming in several of the above-mentioned studies is that they have been performed on fairly small corpora and further that the level of mixing in those corpora often has been quite low. For example, Lignos and Marcus (2013) report that their English-Spanish bilingual corpora in fact were almost monolingual, with a baseline when just guessing on the majority language being as high as 92.3%. Nguyen and Doğruöz (2013) make the assumption that words from languages other than Turkish and Dutch (mainly English) appearing in the messages could be assumed to belong to the dominating language (i.e., Dutch in their case). They give no actual word-level baseline, but state that 83% of the posts are monolingual.

In contrast, Voss et al. (2014) worked on quite code-mixed tweets in Romanized Moroccan Arabic (Darija), English and French, with 20.2% of their data sets consisting of tweets in more than one language. The corpora introduced by Das and Gambäck (2014) are even more mixed, with the English-Hindi corpus having 28.5% of the messages written in at least two languages, and with the English-Bengali corpus even being twice as code-mixed as that (56.5%). This is partially explained by the English-Bengali texts also having many Hindi words mixed in.

## 3 A Code-Mixing Index

When comparing different code-mixed corpora to each other, it is desirable to have a measurement of the level of mixing between languages. To this end we introduced the *Code-Mixing Index*, CMI, in Das and Gambäck (2014). At the utterance level, this amounts to finding the most frequent language in the utterance and then counting the frequency of the words belonging to all other languages present, that is,

$$\text{CMI} = \frac{\sum_{i=1}^{N}(w_i) - max\{w_i\}}{n - u} \qquad (1)$$

where $\sum_1^N(w_i)$ is the sum over all $N$ languages present in the utterance of their respective number of words, $max\{w_i\}$ is the highest number of

```
Ek$hi Kisan$hi 1$univ murga$hi leke$hi aaya$hi .$univ Us$en
murge$hi ne$en aate$en hi$univ 150$univ murgiyo$hi ko$hi
ch0d$hi diya$hi .$univ Ye$en dekh$acro kar$acro kishan$acro
bahut$acro khush$acro hua$acro .$univ Sham$hi tak$en us$en
murge$hi ne$acro sari$acro batakho$acro ($univ duck$en )$univ
or$en Baki$hi janwaro$hi ko$hi b$hi chod$hi diya$hi ,$univ ye$en
dekhkar$en kishan$hi kuch$hi pareshan$hi hua$hi .$univ Agle$hi
din$en jab$hi subah$acro hui$acro to$en murga$en khet$en
me$en mara$hi pda$hi tha$en or$en upar$hi giddh$hi mandra$en
rhe$en the$en .$univ Use$en dekhkar$hi Kishan$hi bola$hi ,$univ
\$univ "$univ mar$en gaya$hi bhosdi$hi k$en ,$univ harkate$en
b$en to$en teri$hi aisi$hi thi\$acro "$univ tabhi$hi murge$hi
ne$acro ek$acro aankh$hi kholi$hi or$en bola$hi \$univ "$univ
chup$hi madarchod$hi inme$hi se$acro  ek$acro ko$hi niche$en
to$en aane$acro de$acro \$hi "$univ
```

Figure 1: Sentence $S_1$, 111 tokens (tagged as Hindi, English, acronymns or universal symbols).

```
Sharab$hi sachai$hi niklwa$hi hi$univ deti$hi h$hi :D$univ
```

Figure 2: Sentence $S_2$, 7 tokens (tagged as Hindi or universal symbols).

words present from any language (regardless of if more than one language has the same highest word count), $n$ is the total number of tokens, and $u$ is the number of tokens given other (language independent) tags.

If an utterance only contains language independent tokens (i.e., if $n = u$), we define its index to be zero. For other utterances, we normalise the value (multiply by 100) to get digits in the range $[0 : 100)$. Further, since $\sum_1^N (w_i)$ in fact is equivalent to $n - u$, Equation 1 can be rewritten as

$$\boxed{\mathrm{CMI} = \begin{cases} 100 \times [1 - \frac{max\{w_i\}}{n-u}] & : n > u \\ 0 & : n = u \end{cases}} \tag{2}$$

where $w_i$ are the words tagged with each language tag and $max\{w_i\}$ thus is the number of words of the most prominent language (so for monolingual utterances, we will get CMI $= 0$, since then $max\{w_i\} = n - u$).

As an example, compare the two sentences in Figure 1 and Figure 2. For the first sentence, we would calculate that it contains 111 tokens in total, with the following number of tokens per category (with OTHERS being the sum of tokens tagged as not belonging to either of the languages, in this case acronyms and universal symbols):

$$\mathtt{HI} : 45 \quad \mathtt{EN} : 29 \quad \mathtt{OTHERS} : 37$$

On the other hand, the second sentence only contains seven tokens, distributed as follows:

$$\mathtt{HI} : \ 5 \quad \mathtt{EN} : \ 0 \quad \mathtt{OTHERS} : \ 2$$

Now if we simply would calculate the mixing as the fraction of Hindi words, that is, as:

$$\frac{\mathtt{HI}}{n} \tag{3}$$

for both the sentences, we would get a value of 0.41 for sentence $S_1$ and 0.71 for sentence $S_2$.

However, in reality $S_1$ is of course much more code-mixed than $S_2$, which basically is monolingual. Hence we need a measure that will capture the kinds of tokens (HI, EN and OTHERS) that actually are present in a sentence. The Code-Mixing Index reflects this, by giving the value 39.19 for sentence $S_1$, while sentence $S_2$ as expected gets a mixing score of 0.00.

As another example, consider a sentence $S_3$ with ten words. If five of the words come from language $L_1$ and the other five from language $L_2$, the CMI will be $100 \times (1 - \frac{5}{10}) = 50$. However, another 10-word sentence $S_4$ with all words coming from different languages will get CMI = $100 \times (1 - \frac{1}{10}) = 90$, correctly reflecting the intuition that $S_4$ presents a more complex mixing.

## 4 Using CMI in Practice

The idea behind the measure is that it will help researchers compare how difficult their work is in relation to that of others, depending on the level of

| Corpus | CMI | | Number of | | Mixed |
|---|---|---|---|---|---|
| | all | mixed | words | utterances | (%) |
| English–Bengali (Das and Gambäck, 2014) | 5.15 | 24.48 | 2309 | 71207 | 21.05 |
| Dutch–Turkish (Nguyen and Doğruöz, 2013) | 4.43 | 26.50 | 3065 | 70768 | 16.70 |

Table 1: Level of code-mixing in two chat corpora

| Language | CMI | | Num. | Mixed |
|---|---|---|---|---|
| (English+) | all | mixed | utt. | (%) |
| Bengali | 5.78 | 24.67 | 700 | 23.43 |
| Hindi | 19.35 | 24.18 | 700 | 80.00 |
| Gujarati | 5.43 | 25.47 | 150 | 21.33 |

Table 2: Code-mixing in the FIRE

| Language | CMI | | Num. | Mixed |
|---|---|---|---|---|
| | all | mixed | utt. | (%) |
| Nepalese | 18.28 | 25.11 | 9993 | 72.79 |
| Spanish | 6.93 | 24.13 | 11400 | 28.70 |
| Mandarin | 10.25 | 19.43 | 999 | 52.75 |
| Arabic | 4.41 | 25.60 | 5839 | 17.21 |

Table 3: Code-mixing in the EMNLP corpora

code-mixing in their corpora. Importantly, whatever language processing tool is being built, it can be argued that for more code-mixed text, the error rates would be expected to be higher.

### 4.1 CMI for Two Chat Corpora

As an example, we utilise the Code-Mixing Index to compare the level of language mixing in the English–Bengali corpus of Das and Gambäck (2014) to that of the Dutch–Turkish corpus of Nguyen and Doğruöz (2013). Table 1 shows the CMI values for these corpora, both on average over all utterances and on average over the utterances having a non-zero CMI, that is, over the utterances that contain some code-mixing. The last column of the table gives the fraction of such mixed utterances in the respective corpora.

It is interesting to compare these two corpora, that are of almost exactly the same size and from similar sources (chat messages), but from different languages. For the Dutch–Turkish corpus we can see a clearly lower average CMI overall than in the English–Bengali corpus, and that the fraction of mixed sentences is smaller. However, the utterances in the Dutch–Turkish corpus that actually are mixed receive a higher CMI, on average.

### 4.2 The FIRE Shared Task Corpora

As a comparison, we have also calculated the CMI values for the various Indian language corpora used in the FIRE 2014 (Forum for IR Evaluation)[1] shared task on transliterated search, which has released data for English (mainly) mixed with six different Indian languages. However, the Kannada text is definitely too short for any statistical purposes (55 words), as is the Tamil one (29 words). Both the Tamil and Malayalam corpora are also only partially and inconsistently annotated (which is a pity, since the Malayalam corpus is of a decent size: 2112 words). Hence, apart from the Hindi and Bengali data, only the Gujarati text could potentially be used for comparison, although it is also fairly short (938 words). On the other hand, the Hindi and Bengali corpora from the FIRE shared task are both of reasonable size: both consist of 700 utterances with in total 23,967 and 20,660 words, respectively. The CMI values for these three languages (when mixed primarily with English) are shown in Table 2.

The very high code-mixing percentage for English–Hindi can partially be explained with tagging problems: neither of the tagsets used in the FIRE shared task, by Das and Gambäck (2014) or by Nguyen and Doğruöz (2013) explicitly account for words that are ambiguous in the context (i.e., words that even given the contextual information could potentially belong to two or more of the languages in the corpus). This is particularly problematic for the Hindi corpus used in the FIRE shared task, which is taken from a Facebook group for Indian university students writing confessionals in a very much short-hand language, where the actual words sometimes are difficult to interpret, which often cause annotation errors.

### 4.3 The EMNLP Shared Task Corpora

Another recent shared task has addressed the problem of code-switching in social media text: The First Workshop on Computational Approaches to Code Switching held in connection to the 2014

[1] http://www.isical.ac.in/~fire/

Conference on Empirical Methods in Natural Language Processing (EMNLP).[2] For the shared task in that workshop, four different code-switched corpora were collected from Twitter (Solorio et al., 2014). Three of these corpora contain English-mixed data from Nepalese, Spanish and Mandarin Chinese, while the fourth corpus consists of tweets code-switched between Modern Standard Arabic and Egyptian Arabic. The CMI values for these four language pairs can be seen in Table 3.

It is interesting to note the extremely high level of mixing in the Nepalese corpus (and fairly high level in the Mandarin Chinese). This could of course possibly have been caused by errors and problems in tagging, but in contrast to the corpora in Tables 1 and 2, the tagset used in the EMNLP shared task included a tag for ambiguous words, which potentially could ease the annotation task. Hence, the high mixing level is more likely a result of the way the data was collected: the data collection was specifically targeted at finding code-switched tweets (rather than finding a representative sample of tweets). This approach to the data collection clearly makes sense in the context of a shared task challenge, although it might not reflect the actual level of difficulty facing a system trying to separate “live” data for the same language pair.

### 4.4 Comparing CMI Values Across Corpora

When comparing CMI values for different corpora, it is important to keep in mind the respective tagsets and guidelines used annotation for the corpora. One such thing to notice is that abbreviations in the EMNLP data have been tagged with the language they belong to, while in the other corpora they rather were tagged as being language independent. This of course affects the CMI values.

Another potential cause of differences can be unclear status of some tags, whether they are directly language related or not. For the EMNLP shared task corpora, we have treated the tags ‘mixed’ and ‘ambiguous’ as language items in the CMI calculations, but without assigning them to any of the languages (when selecting the majority language of an utterance). That can be debated, since the ‘ambiguous’ words could potentially be included among the non-language ones, depending on the interpretation of ambiguous, i.e., if the word is ambiguous between the two languages or between all the tags. However, according to the annotation guidelines for the EMNLP shared task, it should be the former. In practice this does not make a major difference, though: words tagged ‘ambiguous’ are quite few for Spanish and Nepalese, and even 0 for Mandarin, so they could only affect the CMI figures for Arabic in any noticeable way.

A very interesting thing about the EMNLP data is that the corpus with the highest CMI (Nepalese) was the second easiest one to label for the systems participating in the shared task, while the one with the lowest CMI (Arabic) was the most difficult. Obviously, the CMI value only gives an indication of the mixing, not an absolute measure of how difficult it will be to separate the languages in the end. That will also depend on factors such as the closeness between the languages, on external factors like the scripts used to write them, etc. In the EMNLP shared task, the language pair which was the easiest to separate was Mandarin–English, but not for any linguistic reasons, but simply since the two languages were written in different scripts. The most difficult language pair was in contrast triggered by linguistic cues: Standard Arabic–Egyptian Arabic, with the latter being a dialect of the former, so that the two are very close and hence difficult to separate.

## 5 Discussion

Social media text code-mixing in Eurasian languages is a new problem, and needs more efforts to be fully understood and solved. It is also difficult to compare the results reported in different studies to those obtained in other media and for other types of data: While previous work on speech mainly has been on artificially generated data, previous work on text has mainly been on language identification at the document level, even when evidence is collected at word level. Longer documents tend to have fewer code-switching points.

We have here discussed two average code-mixing measurements, ‘CMI all’ which is the average for all sentences in a corpus (i.e., including also all sentences with CMI $= 0$) and ‘CMI mixed’ which is the average only for the sentences that actually contain code-mixing (i.e., only those with CMI $> 0$). Possibly, the ’CMI mixed’ value combined with the fraction of mixed sentences in the corpus give the most interesting information, together showing how multilingual the corpus is and how mixed the multilingual sentences are.

[2] http://emnlp2014.org/workshops/CodeSwitch/call.html

The Code-Mixing Index carries some similarity to the F-factor, or 'formality' (Heylighen and Dewaele, 1999). The F-factor is based on the frequency of the different word classes used in a text, namely the sum of the frequency of nouns, adjectives, prepositions and articles (these classes are called 'formal' by Heylighen and Dewaele, hence the name of the measurement) minus the frequency of pronouns, verbs, adverbs and interjections; all normalised to 100, as follows:

$$F = \frac{1}{2} * (F_N + F_{ADJ} + F_{PREP} + F_{DET} \\ - F_{PRO} - F_V - F_{ADV} - F_{INTER} + 100) \quad (4)$$

with Heylighen and Dewaele (1999) noting that, for example, spoken language recieves F-factors of $40 - 44$ (depending on the education level of the speaker), while novels have $F = 52$ on average and scientific texts $F = 66$. In a way, code-mixing can be seen as a new type of *informality* and it is thus reasonable to ask whether we can give a similar kind of measure for code-mixing. Then we might be able say, for example, that for formal social media text this measure is 25 while for informal social media text it is 80.

However, the CMI described here is actually closer related to simpler (i.e., purely based on word frequencies) readability indices such as the Reading Ease score (Flesch, 1948) or the LIX measure (Björnsson, 1968), since there is no distinction in the CMI between different word classes (which is the main point of the formality measure). In contrast, in LIX the main distinction is binary, between long ($>$ 5 characters) and short words, and the measurement's main part is the frequency of long words, although it also includes a factor based on text length, which supposedly is relevant for readability (but not for code-mixing, clearly). Similarly, Flesch' Reading Ease formula is just based on average sentence length and the number of syllables per 100 words.

In detail, the LIX measurement is calculated as the number of words per sentence plus the percentage of long words:

$$\mathrm{LIX} = \frac{\mathrm{W}}{\mathrm{S}} + \frac{\mathrm{L} \cdot 100}{\mathrm{W}} \quad (5)$$

where $W$ is the number of words in the text, $S$ the number of sentences in the text, and $L$ the number of long words as described above. Björnsson (1968) argued that the readability of a text could be evaluated by this measurement and that different text genres would have different measures on the LIX scale, e.g., with children books having values below 25, simple texts being in the range 25–30, etc., up to difficult scientific texts that would have values above 60.

## 6 Conclusion and Future Work

The paper has described the application of a Code-Mixing Index to measure the complexity of texts written in several different languages, a phenomenon which is particularly common in social media text in geographical regions with high percentages of bi- and multilingual inhabitants, such as on the Indian subcontinent.

We have discussed two average code-mixing measurements, 'CMI all' which is the average for all sentences in a corpus and 'CMI mixed' which is the average only for the sentences that actually contain code-mixing. In the future, maybe it would be an idea to give a combined measure which includes both the fraction of mixed sentences and the (current) 'CMI mixed' (and possibly adding a factor for the multi-linguality) — even though the 'CMI all' in a way gives this, but possibly not as clearly; doing it in a similar way to LIX (by addition) might be clearer and also allows for weighting the fraction of mixed sentences higher (since that is probably the most important distinguishing factor).

Another factor that could be included in the index is the number of code-switching points in a sentence. It is fair to argue that a higher number of switches in a sentence increases its complexity: compare two four-word sentences with two words each from the languages $L_1$ and $L_2$. They will both get CMI $= 100 \times (1 - \frac{2}{4}) = 50$. However, if the first sentence only contains one code-switching point (e.g., if the words are $w_{L_1} w_{L_1} w_{L_2} w_{L_2}$), while the second sentence contains three switches (e.g., with the words $w_{L_1} w_{L_2} w_{L_1} w_{L_2}$), the second sentence will most likely be more difficult to analyse, a fact which potentially could be reflected in the Code-Mixing Index.

The index works well when comparing corpora tagged using the same annotation strategies and similar tagging schemes, but on a general scale, it would be important to factor out differences caused by tagging schemes when comparing corpora from different sources. As discussed in Section 4.3, the EMNLP corpora differ from the other corpora tested in this paper in that the EMNLP tagset included a tag for words that are ambiguous between several languages (which potentially would make the annotation process simpler and more robust).

Further, the annotation strategy choosen for the EMNLP corpora prescribed that elements such as abbreviations should be tagged as language specific, while other annotations schemes treated them as language independent tokens (see Section 4.4). It is likely that these differences in tagsets and annotation approaches has had an effect on the CMI values in the way the measurements were carried out in the present paper. However, it ought to be possible to address this by closer studying the different corpora and their annotations in order to find some neutralising mappings before calculating the Code-Mixing Index.

Certainly, though, an index will never be able to capture all types of differences between corpora. In particular, the ways corpora are collected in the first place and their intended usage can also affect the CMI values (see the discussion of the EMNLP Nepalese corpus in Section 4.3). However, levelling out such differences should arguably not be the aim of the Code-Mixing Index 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 take this into account when deciding on whether a straightforward comparison really makes sense.

## Acknowledgements

Thanks to the different researchers who have made their data sets available: the organisers of the shared tasks in code-switching at EMNLP 2014 and in translateration at FIRE 2014, as well as Dong Nguyen and Seza Doğruöz (respectively University of Twente and Tilburg University, The Netherlands).

Many thanks also to the annonymous reviewers for pointing out unclear issues and for giving us ideas for further work.

## References

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

Carl-Hugo Björnsson. 1968. *Läsbarhet*. Liber, Stockholm, Sweden. (in Swedish).

Zannie Bock. 2013. Cyber socialising: Emerging genres and registers of intimacy among young South African students. *Language Matters: Studies in the Languages of Africa*, 44(2):68–91.

Barbara E. Bullock, Lars Hinrichs, and Almeida Jacqueline Toribio. 2014. World Englishes, code-switching, and convergence. In Markku Filppula, Juhani Klemola, and Devyani Sharma, editors, *The Oxford Handbook of World Englishes*. Oxford University Press, Oxford, England. Forthcoming. Online publication: March 2014.

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

Joyce YC Chan, Houwei Cao, PC Ching, and Tan Lee. 2009. Automatic recognition of Cantonese-English code-mixing speech. *International Journal of Computational Linguistics and Chinese Language Processing*, 14(3):281–304.

Amitava Das and Björn Gambäck. 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*, Goa, India, December. ACL.

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

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

Thomas Gottron and Nedim Lipka. 2010. A comparison of language identification approaches on short, query-style texts. In *Advances in Information Retrieval: 32nd European Conference on IR Research, Proceedings*, pages 611–614, Milton Keynes, UK, March. Springer.

Francis Heylighen and Jean-Marc Dewaele. 1999. Formality of language: definition, measurement and behavioral determinants. Internal report, Center 'Leo Apostel', Vrije Universiteit Brüssel, Brussels, Belgium, November.

Taofik Hidayat. 2012. An analysis of code switching used by facebookers (a case study in a social network site). BA Thesis, English Education Study Program, College of Teaching and Education (STKIP), Bandung, Indonesia, October.

David C. S. Li. 2000. Cantonese-English code-switching research in Hong Kong: a Y2K review. *World Englishes*, 19(3):305–322, November.

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

Marco Lui and Timothy Baldwin. 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.

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

Rosalyn Negrón Goldbarg. 2009. Spanish-English codeswitching in email communication. *Language@Internet*, 6:article 3, February.

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, Seattle, Washington, October. ACL.

Mike Rosner and Paulseph-John Farrugia. 2007. A tagging algorithm for mixed language identification in a noisy domain. In *Proceedings of the 8th Annual INTERSPEECH Conference*, volume 3, pages 1941–1944, Antwerp, Belgium, August. ISCA.

Hong Ka San. 2009. Chinese-English code-switching in blogs by Macao young people. MSc Thesis, Applied Linguistics, University of Edinburgh, Edinburgh, Scotland, August.

Latisha Asmaak Shafie and Surina Nayan. 2013. Languages, code-switching practice and primary functions of Facebook among university students. *Study in English Language Teaching*, 1(1):187–199, February.

Thamar Solorio and Yang Liu. 2008. Part-of-speech tagging for English-Spanish code-switched text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1051–1060, Honolulu, Hawaii, October. ACL.

Thamar Solorio, Melissa Sherman, Yang Liu, Lisa M. Bedore, Elisabeth D. Peña, and Aquiles Iglesias. 2011. Analyzing language samples of Spanish-English bilingual children for the automated prediction of language dominance. *Natural Language Engineering*, 17(3):367–395, July.

Thamar Solorio, Elizabeth Blair, Suraj Maharjan, Steven Bethard, Mona Diab, Mahmoud Gohneim, Abdelati Hawwari, Fahad AlGhamdi, Julia Hirschberg, Alison Chang, and Pascale Fung. 2014. Overview for the first shared task on language identification in code-switched data. 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.

Susanna Sotillo. 2012. *Ehhhh utede hacen plane sin mi???:@ im feeling left out:(* form, function and type of code switching in SMS texting. In *ICAME 33 Corpora at the centre and crossroads of English linguistics*, pages 309–310, Leuven, Belgium, June. Katholieke Universiteit Leuven.

Clare Voss, Stephen Tratz, Jamal Laoudi, and Douglas Briesch. 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.

Jochen Weiner, Ngoc Thang Vu, Dominic Telaar, Florian Metze, Tanja Schultz, Dau-Cheng Lyu, Eng-Siong Chng, and Haizhou Li. 2012. Integration of language identification into a recognition system for spoken conversations containing code-switches. In *Proceedings of the 3rd Workshop on Spoken Language Technologies for Under-resourced Languages*, pages 76–79, Cape Town, South Africa, May.

Alma Lilia Xochitiotzi Zarate. 2010. Code-mixing in text messages: Communication among university students. In *Memorias del XI Encuentro Nacional de Estudios en Lenguas*, pages 500–506, Tlaxcala de Xicohtencatl, Mexico. Universidad Autónoma de Tlaxcala.