An Arabic-Moroccan Darija Code-Switched Corpus

Younes Samih and Wolfgang Maier
Institute for Language and Information
University of Düsseldorf, Düsseldorf, Germany
{samih,maierwo}@phil.hhu.de

Abstract

In multilingual communities, speakers often switch between languages or dialects within the same context. This phenomenon is called code-switching. It can be observed, e.g., in the Arab world, where Modern Standard Arabic and Dialectal Arabic coexist. Recently, the computational treatment of code-switching has received attention. Just as other natural language processing tasks, this task requires annotated linguistic resources. In our work, we turn to a particular under-resourced Arabic Dialect, Moroccan Darija. While other dialects such as Egyptian Arabic have received their share of attention, very limited effort has been devoted to the development of basic linguistic resources that would support a computational treatment of Darija. Motivated by these considerations, we describe our effort in the development and annotation of a large scale corpus collected from Moroccan social media sources, namely blogs and internet discussion forums. It has been annotated on token-level by three Darija native speakers. Crowd-sourcing has not been used. The final corpus has a size of 223k tokens. It is, to our knowledge, currently the largest resource of its kind.
Keywords: code-switching, language identification, Moroccan Arabic

1. Introduction

Modern Standard Arabic (MSA) is the official language of most Arabic countries. It is spoken by more than 360 million people around the world and exists in state of diglossia (Ferguson, 1959). Arabic speakers tend to use Dialectal Arabic (DA) and MSA, two substantially different but historically related language varieties, for different purposes as the situation demands in their day-to-day lives.

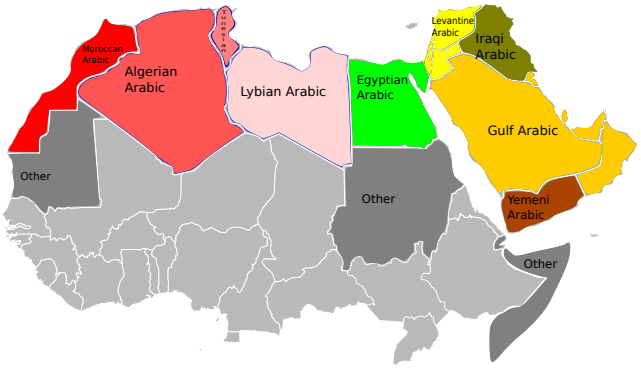

Figure 1: The Arab world and Arabic Dialects

Fig. 1 shows a schematic map of dialects. Note that often Moroccan, Algerian, Tunisian and Lybian Arabic are grouped together as Maghrebi Arabic, even though they are not necessarily mutually intelligible. While MSA is an established standard among educated Arabic speakers, DA is only used in everyday informal communication. Until recently, DA was considered as a partially under-resourced language, as the written production remains relatively very low in comparison to MSA. Increasingly, however, DA is emerging as the language of informal communication on the web, in emails, micro-blogs, blogs, forums, chat rooms, etc. This new situation amplifies the need for consistent language resources and language identification systems for Arabic and its dialects. While certain dialects, particularly Egyptian, have already received attention in NLP research, Moroccan Arabic (*Darija*) (Ennaji et al., 2004; Benmamoun, 2001), a dialect with over 21 million native speakers (Lewis et al., 2014), remains a particularly under-resourced variant of Arabic. It is strongly embedded in a multilingual context that entails frequent code-switching, i.e., switching between languages within the same context (Bullock and Toribio, 2009). Building linguistic resources and creating the necessary tools for Darija is a priority, not least because its vocabulary is particularly distant to MSA (Diab et al., 2010).

In this paper, we therefore contribute a corpus of Moroccan Darija with code-switching annotation on token level. The corpus has been collected from internet discussion forums and blogs, and is currently the largest manually annotated Arabic-Moroccan Darija code-switched corpus known to the authors. It will be of use for supporting research in the linguistic and sociolinguistic aspects of code-switching of Arabic and it will constitute an ideal data source for multilingual processing in general and for research in code-switching detection in particular, an area which recently has attracted attention (see Sec. 5.).

The remainder of the paper is organized as follows. In the following section, we outline the properties of code-switching. In Sec. 3., we describe the corpus creation. Sec. 4. presents the annotation. Sec. 5. reviews related work and Sec. 6. concludes the article.

2. Code-Switching

2.1. Linguistic Analysis of Code-Switching

Code-switching[1] is common phenomenon in multilingual communities wherein speakers switch from one language or dialect to another within the same context (Bullock and Toribio, 2009). Communities where commonly, more than one language, resp. dialect is spoken can be found around the world. Examples include India, where speakers switch between English and Hindi (among other local languages)

[1]Note that for the purpose of this paper, we do not distinguish between *code-switching* and similar concepts such as *code-mixing*.

(Dey and Fung, 2014); the United States, where migrants from Spanish-speaking countries continue to use their native language alongside English (Poplack, 1980); Spain, where people switch between regional languages such as Basque and Spanish (Muñoa Barredo, 2003); Paraguay, where Spanish co-exists with Guarani (Estigarribia, 2015); and finally the Arab world, where speakers alternate between MSA and Dialectal Arabic.

In the literature, three types of codes-switching are distinguished. In *inter-sentential* code-switching languages are switched between sentences. An instance of this type of switching is (1) (from Muñoa Barredo (2003)), where the speaker switches from Basque to Spanish.

(1) egia ez dala erreala? *eso es otra cosa!*
you say that the truth is not real? *that's a different thing!*

Intra-sentential code-switching consist of a language switch *within* a sentence. An example is (2) (borrowed from Dey and Fung (2014)). Here, the speaker switches from Hindi to English within the same sentence.

(2) Tume nahi pata, *she is the daughter of the CEO,* yaha do char din ke liye ayi hai.
Dont you know, *she is the daughter of the CEO*, shes here for a couple of days.

A third type of code-switching is *intra-word* switching, where a language switch occurs in a single word. For instance, the morphology of one language involved can be applied on a stem of the other language. An corresponding example can be found in the Sec. 2.2..

Since the mid 1960s, there has been a large body of linguistic studies on code-switching, the bulk of them concentrating on social and linguistic factors that constrain its occurrences (Berk-Seligson, 1986). Various models of constraints have been proposed. Poplack (1980) formulates a model in terms of the *equivalence constraint* of the languages involved at the switch point. Namely, code-switching tends to occur at points in the sentence where the surface structure of the respective languages is the same. Myers-Scotton (1993) focuses on structural constrains in code-switching. She proposes the *matrix language-frame* (MLF). It is based on the assumption that one language is the matrix language (ML) and the other language is the embedded language (EL). While ML provides the grammatical and functional elements as well as the structural frame of the sentence, the EL can only provide content elements (Myers-Scotton, 1997). For a further, detailed linguistic overview, consult Muysken (2000).

2.2. Code-Switching in Morocco

The linguistic situation in Morocco complex due to its diverse ethnic and linguistic make-up and the colonial history. Following Benmamoun (2001), one can distinguish different languages and dialects that occupy the linguistic space:

Darija is the native language for the majority of the population and is the language of popular culture.

Berber is the language of the original people of Morocco. It is the native language of about 40% of the Moroccan population.

Modern Standard Arabic is a written language used mainly in formal education, media, administration, and religion.

French is not an official language, but dominant in higher education, in the media, and some industries.

In recent years, the Moroccan linguistic landscape has changed dramatically due to social, political, and technological factors. Darija, the colloquial, traditionally *unwritten* variety of Arabic, is increasingly dominating the linguistic scene. It is being written in a variety of ways in print media, advertising, music, fictional writing, translation, the scripts for dubbed foreign TV series, and a weekly news magazine (Elinson, 2013). It is also increasingly appearing on the web in blogs, emails, and social media platforms and is often code-switched with other languages and dialects, including MSA, English, and French, Spanish and Berber (Tratz et al., 2014).

As an example for intra-sentential and intra-word code-switching in Morocco, consider (3). It is taken from our own data.

(3) فرنسا لن يبقى فيها سوى الزواق و الشيكي و الجيب خاوي و نزيدكم و الاكل بالموس و الفورشيطا.
In France, the only things that remain are pretension, empty pockets, and, to add more, also eating with knife and fork.

The speaker switches between MSA, Darija, and uses a word where French is mixed with Arabic morphology. MSA words include, e.g., فرنسا (*France*), الاكل بالموس (*eating with knife*); Darija words include الزواق و الشيكي و الجيب خاوي و نزيدكم (*pretension, empty pockets, and, to add more*); finally, الفورشيطا is the French word for *fork* ("*fourchette*"), written in Arabic script. It is prefixed by the Arabic definite article and suffixed with an Arabic case marker.

3. Corpus Creation

We acquire our data from internet discussion forums and blogs which are hosted in Morocco or extensively used by Moroccans. The crawled output is stripped from HTML tags and other meta-data. Since sentence splitting is not a trivial task in Arabic and no such tool is available for Darija, we leave the downloaded text units ("posts") intact. Then we tokenize the text with a simple heuristic, delete all diacritic marks as usual in Arabic NLP (Habash, 2010),

and Buckwalter transliterate the text. Finally, we store the data token-wise as pairs (original and transliteration) in a MySQL database. The size of the resulting corpus is 15 million tokens in total. It comprises a wide range of topics including politics, religion, sport and economics.

Existing code-switched data sets are often highly skewed towards one language (Solorio et al., 2014), with a high percentage of the sentences not exhibiting code-switching at all. Also in our corpus, MSA is more prevalent than Darija. In order to obtain a resource that concentrates on code-switching, we aim at minimizing the skewing by extracting only a subset of the data set that contains more instances of code-switching. The subset is extracted with the following iterative process. We first compile an initial seed list of 439 commonly used Darija words and phrases collected from the internet.[2] Then we repeat the following steps. Each word and phrase in the seed list is formulated in a MySQL query as a keyword to retrieve more code-switched examples from the original data set. The retrieved examples are put into the code-switched data set. Then, the seed list is updated with all words of the retrieved text units, and the procedure is repeated until the code-switched data set has reached a certain size. Note that the extraction procedure does not guarantee that only code-switched examples are included in the final data set. However, we do achieve a rate of 73.9% of text units with code-switching (see below), which contrasts with code-switching ratios of around 20% in data sets in previous literature (Solorio et al., 2014).

Sec. 4. shows more detailed statistics of the data.

4. Annotation

We adapt the annotation guidelines for the data used in the shared task on code-switching detection at EMNLP 2014 (Solorio et al., 2014). We use all of their labels:

- `lang1` is used for MSA words,
- `lang2` for Darija words,
- `mixed` is used to mark words that mix a Darija stem with MSA morphology or vice versa,
- `ne` is used to mark named entities, including dates,
- `other` is used to mark other numbers, punctuation, and other non-language material, and
- `ambiguous` is used for material which could be interpreted as either `lang1` or `lang2`.

We additionally introduce a new label `lang3`, which accounts for the special linguistic situation in Morocco. It marks words belonging to a language which is neither MSA nor Darija, notably either French, English, Spanish or Berber. Note that French words might be written with Arabic script. Berber is furthermore written with its own script, called Tifinagh. Since `lang3` material is scarce, we have decided against the introduction of further language labels.

[2] `darijadictionary.com` and `en.mo3jam.com/dialect/Moroccon`

	all	%	length av.	length med.
`lang1`	109,025	48.82	6.13	4
`lang2`	76,732	34.36	5.22	4
`lang3`	1,383	0.62	1.88	1
`ne`	17,087	7.65	1.29	1
`mixed`	86	0.04	1.00	1
`ambiguous`	141	0.06	1.26	1
`other`	18,830	8.43	1.02	1

Table 1: Data statistics

Several DA annotation tools have been reported in the literature, such as COLANN_GUI and COLABA (Benajiba and Diab, 2010; Diab et al., 2010), DATOOL (Tratz et al., 2013), DIWAN (Al-Shargi and Rambow, 2015), among others, and have been successful in serving different annotation tasks, yet they were either not available or did not suit our needs exactly. Therefore, we have built a custom web-based annotation tool. The annotation has been performed by three Moroccan Darija native speakers, two of them with no prior linguistic knowledge. They worked independently at different physical locations; crowd-sourcing services like Amazon Mechanical Turk have not been used. In our annotation tool, a single text unit is shown at a time. The words can be annotated in random order. When the annotation of a text unit is done, it is saved by a single click back to the database, thereby the label is stored for each token together with its Arabic script version and its Buckwalter transliterated version.

To ensure the agreement among the annotators, various training sessions were provided and regular inter-annotator agreement measures were performed to check the annotation quality. The final inter-annotator agreement (Cohen's κ) was computed on a selection of 50 text units between 0.82 and 0.86.

In total, 3,862 text units with around 223k tokens have been annotated. The annotation is still ongoing and the resource continues to grow. Tab. 1 the number of tokens for each label, as well as the respective ratio among all labels, and average as well as median length of tokens per label. We found the average number of switches between languages per text unit to be 2.6, and the median to be 2. Out of the 3,862 text units, 284 (7.4%) contain only MSA tokens, and 725 (18,8%) contain only Darija tokens. In other words, 2,853 (73,9%) text units are true code-switched instances. In 1,950 (50.5%) text units, more than 50% of all language tokens (i.e., not `mixed` tokens, etc.) are `lang1`, the same holds for `lang2` in 1,970 (51.0%) text units.

As mentioned above, we have not performed sentence splitting, i.e., our text units can consist of more than one sentence. Fig. 2 shows a histogram revealing the distribution of lengths of text units. We see that it is positively skewed; most text units have 50 or less tokens. Nevertheless, our data does contain very long text units with over 300 tokens.

As an example for the annotation, Fig. 3 shows a sample forum post with free translation, and the annotated version of the text. Aside from MSA words, this text unit contains Darija words, and words with mixed morphology. Darija

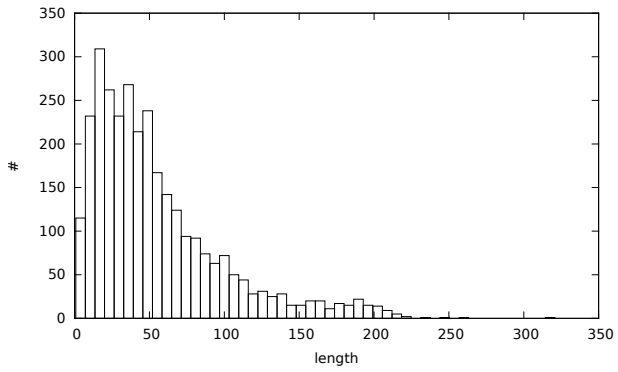

Figure 2: Text unit length histogram

رآه غاتمشي هاذ الحكومة و غادجي حكومة اخرى و لكن كرف من هذي لأن الحكومة الوحيدة اللي بغات تصلح لبلاد هي هذي منذ الاستقلال و ها حنا غانشوفو الشفارة اللي غاي يرجعو (الاتحاد الدستوري ، الاستقلال ، الأحرار و لقتالة) .

rAh/lang2 gAtm$y/lang2 hA*/lang2 AlHkwmt/lang2 w/lang2 gAdjy/lang2 Hkwmt/lang2 AxrY/lang2 w/lang2 lkn/lang2 krf/lang2 mn/lang2 h*y/lang2 l>n/lang1 AlHkwmt/lang1 AlwHydt/lang1 Ally/lang2 bgAt/lang2 tSlH/lang2 lblAd/lang2 hy/lang2 h*y/lang2 mn*/lang1 AlAstqlAl/lang1 w/lang2 hA/lang2 HnA/lang2 gAn$wfw/lang2 Al$fArt/lang2 Ally/lang2 gAy/lang2 yrjEw/lang2 (/other AlAtHAd/ne Aldstwry/ne ,/other AlAstqlAl/ne ,/other Al>HrAr/ne w/lang2 lqtAlt/lang2)/other ./other

This government will go, and another one will come which will be worse. The current government is the only government since the independence which has worked hard to develop the country; and we will see that the thieves coming back (Independence Party, Constitutional Party, Liberal Party, and murderers).

Figure 3: Sample forum post with annotated version and free translation

words are, e.g., رآه (demonstrative pronoun), كرف (meaning *worse*). Words with mixed morphology are, e.g., غاتمشي and غادجي, which both exhibit the Darija future prefix غا on words which otherwise are MSA. Last, note that the punctuation in the text unit is labeled as `other`. For the purpose of presentation, i.e., in order to not having to mix writing directions, the annotated tokens are shown only in the Buckwalter transliterated form; labels are separated from the tokens they annotated by a single forward slash.

5. Related Work

Research in processing of code-switching and of language varieties and dialects has recently attracted attention. This is reflected by recent workshops (Diab et al., 2014; Solorio et al., 2014; Nakov et al., 2014; Zampieri et al., 2014). A number of language resources has been created, such as the ones described by Tratz et al. (2013), Maharjan et al. (2015), Dey and Fung (2014), and the data set from the Shared Task at the First Workshop on Computational Approaches to Code-Switching at EMNLP 2014 (Solorio et al., 2014), with a particular focus on inter-operable annotation guidelines. A popular use for those resources can be found in approaches to automatic detection of code-switching points in text. This task has mostly been treated as a sequence labeling problem. Different techniques have been applied, ranging from Naive Bayes (Solorio and Liu, 2008) over Conditional Random Fields (King and Abney, 2013; Elfardy et al., 2014), Support Vector Machines (Bar and Dershowitz, 2014), Markov Models (King et al., 2014) and n-gram based approaches (Shrestha, 2014; Bacatan et al., 2014) to Recurrent Neural Networks (Chang and Lin, 2014). POS tagging of code-switched text has also been investigated (Solorio and Liu, 2008b; Rodrigues and Kübler, 2013).

Concerning the processing of Arabic in general, there is an ample body of research. For an overview, see Habash (2010). With regard to the processing of Dialectal Arabic, most of the existing work concentrates on Levantine and Egyptian Arabic (see, e.g., Elfardy and Diab (2012) and Elfardy et al. (2014)). An exception is Cotterell et al. (2014), who works on Algerian Arabic. In linguistics, Moroccan Arabic has found attention very early (Harrell, 1962; Harrell and Sobelman, 1966). Work on the computational processing of Darija, however, remains very scarce. To our knowledge, there is only the work of Tratz et al. (2013), who present a data collection and annotation environment for romanized Darija, and the work of Voss et al. (2014) who present an approach for finding romanized Darija in code-mixed tweets.

6. Conclusion and Future Work

We have presented a corpus of Moroccan Arabic Darija, obtained from internet discussion forums and blogs, manually annotated for code-switching on token level without crowd-sourcing. With its 223k tokens, to our knowledge, it is currently the largest resource of its kind.

In future work, we will additionally annotate the data with part-of-speech information.

7. Acknowledgments

We would like to thank Ikram Zeraa and Miloud Samih for their invaluable annotation work. The authors were partially funded by Deutsche Forschungsgemeinschaft (DFG).

8. Bibliographical References

Al-Shargi, F. and Rambow, O. (2015). Diwan: A dialectal word annotation tool for Arabic. In *Proceedings of the Second Workshop on Arabic Natural Language Processing*, pages 49–58, Beijing, China. Association for Computational Linguistics.

Bacatan, A. C., Castillo, B. L., Majan, M. J., Palermo, V., and Sagum, R. (2014). Detection of intra-sentential code-switching points using word bigram and unigram

frequency count. *International Journal of Computer and Communication Engineering*, 3(3):184–188.

Bar, K. and Dershowitz, N. (2014). The Tel Aviv University system for the code-switching workshop shared task. In *Proceedings of the First Workshop on Computational Approaches to Code Switching*, pages 139–143, Doha, Qatar. Association for Computational Linguistics.

Benajiba, Y. and Diab, M. (2010). A web application for dialectal Arabic text annotation. In *Proceedings of the LREC Workshop for Language Resources (LRs) and Human Language Technologies (HLT) for Semitic Languages: Status, Updates, and Prospects*, Valletta, Malta. ELRA.

Benmamoun, E. (2001). Language identities in Morocco: A historical overview. *Studies in the Linguistic Sciences*, 31(1):95–106.

Berk-Seligson, S. (1986). *Linguistic Constraints on Intrasentential Code Switching: A Study of Spanish-Hebrew Bilingualism*. Cambridge University Press.

Bullock, B. E. and Toribio, A. J. (2009). *The Cambridge handbook of linguistic code-switching*. Cambridge University Press.

Chang, J. C. and Lin, C.-C. (2014). Recurrent-neural-network for language detection on twitter code-switching corpus. *arXiv preprint arXiv:1412.4314*.

Cotterell, R., Renduchintala, A., Saphra, N., and Callison-Burch, C. (2014). An Algerian Arabic-French code-switched corpus. In *LREC Workshop on Free/Open-Source Arabic Corpora and Corpora Processing Tools*, Reykjavik, Iceland.

Dey, A. and Fung, P. (2014). A Hindi-English code-switching corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

Diab, M., Habash, N., Rambow, O., Altantawy, M., and Benajiba, Y. (2010). COLABA: Arabic dialect annotation and processing. *Proceedings of the LREC Workshop for Language Resources (LRs) and Human Language Technologies (HLT) for Semitic Languages: Status, Updates, and Prospects*, pages 66–74.

Mona Diab, et al., editors. (2014). *Proceedings of the First Workshop on Computational Approaches to Code Switching*. Doha, Qatar.

Elfardy, H. and Diab, M. (2012). Token level identification of linguistic code switching. In *Proceedings of COLING 2012: Posters*, pages 287–296, Mumbai, India.

Elfardy, H., Al-Badrashiny, M., and Diab, M. (2014). AIDA: Identifying code switching in informal Arabic text. In *Proceedings of the First Workshop on Computational Approaches to Code Switching*, pages 94–101, Doha, Qatar.

Elinson, A. E. (2013). Dārija and changing writing practices in Morocco. *International Journal of Middle East Studies*, 45(04):715–730.

Ennaji, M., Makhoukh, A., Es-saiydy, H., Moubtassime, M., and Slaoui, S. (2004). *A Grammar of Moroccan Arabic*. Number 25 in Pars Lettres. Faculty of Letters Dhar El Mehraz, Fès, Fès, Morocco.

Estigarribia, B. (2015). Guaran-Spanish Jopara mixing in a Paraguayan novel – does it reflect a third language, a language variety, or true codeswitching? *Journal of Language Contact*, 8(2):182–222.

Ferguson, C. (1959). Diglossia. *Word*, 15:325–340.

Habash, N. Y. (2010). Introduction to Arabic natural language processing. *Synthesis Lectures on Human Language Technologies*, 3(1):1–187.

Richard Harrell et al., editors. (1966). *A dictionary of Moroccan Arabic*. Georgetown University Press.

Harrell, R. S. (1962). *A Short Reference Grammar of Moroccan Arabic*. Georgetown University Press.

King, B. and Abney, S. (2013). Labeling the languages of words in mixed-language documents using weakly supervised methods. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1110–1119, Atlanta, Georgia. Association for Computational Linguistics.

King, L., Baucom, E., Gilmanov, T., Kübler, S., Whyatt, D., Maier, W., and Rodrigues, P. (2014). The IUCL+ system: Word-level language identification via extended Markov models. In *Proceedings of the First Workshop on Computational Approaches to Code Switching*, pages 102–106, Doha, Qatar. Association for Computational Linguistics.

M. Paul Lewis, et al., editors. (2014). *Ethnologue: Languages of the World*. SIL International, Dallas, Texas, seventeenth edition. Online version: `http://www.ethnologue.com`.

Maharjan, S., Blair, E., Bethard, S., and Solorio, T. (2015). Developing language-tagged corpora for code-switching tweets. In *Proceedings of The 9th Linguistic Annotation Workshop*, pages 72–84, Denver, Colorado, USA.

Muñoa Barredo, I. (2003). Pragmatic functions of code-switching among Basque-Spanish bilinguals. In *Comunidades e individuos bilinges: Actas do I Simposio Internacional sobre o Bilingismo*. Universidade de Vigo: Servizo de Publicacins.

Muysken, P. (2000). *Bilingual speech: A typology of code-mixing*, volume 11. Cambridge University Press.

Myers-Scotton, C. (1993). Common and uncommon ground: Social and structural factors in codeswitching. *Language in Society*, 22:475–475.

Myers-Scotton, C. (1997). *Duelling Languages: Grammatical Structure in Codeswitching*. Clarendon Press.

Preslav Nakov, et al., editors. (2014). *Proceedings of the EMNLP'2014 Workshop on Language Technology for Closely Related Languages and Language Variants*. Doha, Qatar.

Poplack, S. (1980). Sometimes I'll start a sentence in Spanish y termino en español: Toward a typology of code-switching. *Linguistics*, 18(7-8):581–618.

Rodrigues, P. and Kübler, S. (2013). Part of speech tagging bilingual speech transcripts with intrasentential model switching. In *AAAI Spring Symposium: Analyzing Microtext*.

Shrestha, P. (2014). Incremental n-gram approach for language identification in code-switched text. In *Proceed-*

ings of the First Workshop on Computational Approaches to Code Switching, pages 133–138, Doha, Qatar. Association for Computational Linguistics.

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

Solorio, T. and Liu, Y. (2008b). 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. Association for Computational Linguistics.

Solorio, T., Blair, E., Maharjan, S., Bethard, S., Diab, M., Ghoneim, 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 First Workshop on Computational Approaches to Code Switching*, pages 62–72, Doha, Qatar.

Tratz, S., Briesch, D., Laoudi, J., and Voss, C. (2013). Tweet conversation annotation tool with a focus on an Arabic dialect, Moroccan Darija. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 135–139, Sofia, Bulgaria.

Tratz, S., Briesch, D., Laoudi, J., Voss, C., and Holland, V. M. (2014). Language and dialect identification in social media analysis. *Proceedings of SPIE*, 9122:91220K–91220K–11.

Voss, C., Tratz, S., Laoudi, J., and Briesch, D. (2014). Finding romanized Arabic dialect in code-mixed tweets. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland.

Marcos Zampieri, et al., editors. (2014). *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*. Dublin, Ireland.