HAL
open science

CESAR: A new metric to measure the level of code-switching in corpora -Application to Maghrebian dialects

Karima Abidi, Kamel Smaïli

To cite this version:

Karima Abidi, Kamel Smaïli. CESAR: A new metric to measure the level of code-switching in corpora -Application to Maghrebian dialects. IntelliSys2021, Sep 2021, Amsterdam, Netherlands. hal-03139685

HAL Id: hal-03139685

https://hal.science/hal-03139685v1

Submitted on 12 Feb 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

CESAR: A new metric to measure the level of code-switching in corpora - Application to Maghrebian dialects

Karima Abidi[1] and Kamel Smaïli[1]

SMarT Group, Loria, University of Lorraine, France, 54600, France
karima.abidi@loria.fr, kamel.smaili@loria.fr,

Abstract. In this paper, we are interested in a sociolinguistic phenomenon that occurs in daily conversations of Maghrebi people, commonly known as code-switching or also code-mixing. This problem consists of alternating languages during communication or writing. In this work, we measure the importance of this phenomenon in the Maghrebi languages. To this end, we harvested from YouTube, comments written in Algerian, Moroccan and Tunisian dialects. Each of which contains at least 17 million words. Although there are several metrics in the literature to measure the code-switching, to the best of our knowledge, there isn't yet a measure that takes into account the degree of mixture according to a reference language. In contrast to the existing measures, we propose a new metric named CESAR (CodE-Switching According to a Reference language) that estimates the degree of the language mixtures, in accordance with a reference language. Experiments are carried out on the three collected corpora by considering the local dialects as reference languages. Experimental results show that CESAR is well adapted to this purpose and allows to compare the three Maghrebi dialects according to their level of code-switching. ...

Keywords: Code-swtching, dialecte, reference language

1 Intrduction

Modern Standard Arabic (MSA), the official language shared by the entire Arab world, is a simplified form of the Classical Arabic, which is mostly used in the religious texts. In addition to MSA, there is another form of Arabic dedicated to the daily communication, named in the Maghreb countries *Darija* (Arabic dialect). Nowadays, with the advent of social networks, the Arabic dialect has become widely used. Indeed, the MSA is not the mother tongue of the Maghrebi people, they learn it language at school. That is why, in the daily conversation, they mostly prefer to use their dialect instead of MSA especially when they post messages on social networks. The Arabic dialect is henceforth written, which leads to several new NLP challenges. In fact, this form of Arabic has undergone great morpho-syntactic modifications by relaxing several grammatical constraints of MSA. Furthermore, in each country of North Africa, there are

different dialects with different linguistic variations. More importantly, people can mix several languages in addition to their own dialect, which makes the processing of these dialects more difficult.

The phenomenon of code-switching is not specific to North Africans, it is common to countries having an immigrant background or for those with long historical relationships with other cultures. For instance, the Turkish people living in Germany can switch from Turkish to German and the Indian people living in Great Britain may switch from Indian to English. Likewise, the Maghrebi people living in France switch from their Arabic dialect to French. This phenomenon, known as code-mixing or code-switching, has attracted significant attention by psychologists and sociologists for many years [1], [2], [3], [4], [5] and [6].
Recently, the code-switching has aroused keen interest among the NLP community [7], [8], [9], [10], [11],[12], etc. This community proposed to measure the level of the code-switching, with the aim of identifying the language break points in a document.
In this article, we aim to study the phenomenon of the code-switching in the three Maghrebi dialects : Algerian, Moroccan and Tunisian. This issue is not only common in formal communications, but also in informal ones, such as those used in social networks. Using a mixture of languages, in a conversation constitutes a real challenge for NLP. We are used to process one language with all its available resources. However, when we have more than one language in an utterance, should we use the resources of all the concerned languages? Should we translate all the phrases written or spoken in different languages to a unique target one, then processing the result as if we had a text in a unique language? These issues and others raise technical difficulties for automatic speech recognition systems [13], for machine translation and for other NLP applications.

The three countries of North Africa concerned by the study in this work (Algerian, Moroccan and Tunisian) have in common the Modern Standard Arabic (MSA),the French and sometimes the English. In addition each country has several dialects. A Maghrebian can mix in the same conversation his dialect with MSA, French and English or other foreign languages. Mixing may concern one isolated word or several contiguous words. In the following, we give an illustration of the phenomenon of code-switching for each of the three mentioned countries with their translation into English. These examples have been extracted from the comments posted under the videos of YouTube.

ALG: Verry simple for you *on veut toujours les ingrédients* **Okhti** *merci.*
(*Very simple for you, anyway my sister we would like to have the ingredients, thanks.*)

MAR: : تبارك الّله عليك *merci pour la vidéo* عجبني *les boucles d'oreilles* منين خديتهم **thanks.**
(*God bless you thank you for this video I liked the earrings where did you buy them. Thanks.*)

TUN: اما فكرة باهية تحسنت *audio quality !* **sinon** *keep it up* عندك مستقبل.
(*Very good idea the quality of audio has been improved ! otherwise keep it up you are talented.*)

In the given examples four languages have been combined, namely: English, French, MSA and the local dialects. In order to make easy the reading of the examples, each time a new language was used, we wrote it in a new format (bold or italic). In the Tunisian (TUN) example, the comment is written in three languages: a local dialect and/or MSA by using Arabic Script, French and English. In the Algerian post (ALG), in addition to English and French, an Arabic word (**Okhti**) has been written in Arabizi (using Latin Script). In the Moroccan (MAR) example, the author switched several times from Arabic (written in Arabic Script) to French and ends the comment with an English word. Despite the shortness of these comments extracted from YouTube, at several positions in the sentence, the speaker decided to change the language of writing. In the following, we will call these locations the language break points.

In this article we measure the complexity of code-switching of each of the three studied languages by using the metric proposed in [14]. This measure is not adapted to what we would like to quantify. In fact, we would like to estimate the noise brought by the other languages relative to a reference language (the local dialect of each country). That is why, we propose a new measure CESAR. This metric is bounded; a null value means that the corpus is entirely written in the reference dialect while a value of one corresponds to a corpus entirely written in languages different from the reference. CESAR allows to estimate how much a corpus is mixed with other languages.

The rest of the paper is organized as follows: Section 2 concerns the related works, while Section 3 describes the collected corpora for each dialect. In Section 4, we discuss the proposed CESAR metric that measures the complexity of code-switching in multilingual texts. We evaluate this measure and compare it against the one proposed by [14] in Section 5 we conclude.

2 Related works

Several works tried to tackle the problem of code-switching, by building annotated speech or textual code-switched corpora such as in [15]. The author built a switched corpus extracted from Twitter which contains 1029 Turkish-German code-switched tweets. This corpus has been manually tokenized and normalized. The same author in [16] built and annotated a code-switched speech corpus by recording conversations of bilingual speakers for the same pair of languages. In [17], a new approach for building a corpus of code-switched human–machine dialog was presented, the authors used for that purpose the HALEF [1] platform. They collected a total of over 700 dialogues. Samih and Maier in [18] harvested code-switched data from Moroccan social media sources (blogs and forums). The

[1] https://sourceforge.net/projects/halef/

constructed corpus was annotated by three Darija native speakers. The final obtained corpus had a size of 223k tokens. Authors in [11] built a code-switched corpus named FACST (French-Algerian Code-switching Triggered audio corpus) by recording spontaneous conversations of Algerian speakers living in France.

Regarding the identification of code-switching break points, Al-Badrashiny et al. in [8] proposed a hybrid system combining different classifiers and components such as a language model, a named entity identifier and a morphological analyzer to identify dialectal phrases in a stream of data in MSA. This method achieved a F-score of 90.6%. Another method allowing to identify the code switched break points was proposed in [7]. It was based on multi-structural word information such as graphemes, syllables and words. This method achieved a score of 96.36% in terms of accuracy. Jaech et al. in [19] proposed a new model called C2V2L ("character to vector to language") based on hierarchical neural model for language identification. This model was evaluated in code-switched data extracted from Twitter for the pair of languages English-Spanish. This method achieved F-score values of 93.1 and 97.7 for Spanish and English, respectively. In [**?**], the authors tackled the issue of translating a code-switched corpus into two different target langues, those concerned by the mixture of languages

The phenomenon of code-switching has been also handled in speech synthesis [20]. The authors proposed a mixed-lingual speech synthesis system by using a mixture of 4 monolingual corpora: Hindi, Telugu, Marathi and Tamil. The tests were performed by using a subjective evaluation. To do so, the authors conducted two types of listening tests by using Web Audio Evaluation Tool (WAET). They first evaluated the naturalness of the produced speech synthesis, on a scale of 1 (least natural) to 5 (highly natural) and then asked the listeners to mention the system of synthesis they prefer.

In the recent few years, researchers started to measure the complexity of a text, including several languages, at different levels of granularity: sentence or a whole document [9]. These authors proposed a metric and tested it on several code-switched corpora (English-Nepalese, English-Mandarin, English-Hindi, etc.) collected from social networks. In [14], the authors proposed another measure to determine the level of code-switching. This metric combines three parameters: a language factor that indicates the number of languages present simultaneously in a text, a mixed factor which is related to the number of words not written in the dominant language and a switched factor that corresponds to the number of language break points. The proposed metric was evaluated on a corpus of 3701 code-switched sentences. They calculated the level of code-switching for each sentence and the whole corpus.

3 The extracted corpora

In order to study the phenomenon of the code-switching, we build corpora by harvesting data from the YouTube comments. To do so, we used the Google's API [2] to extract comments of videos posted by Algerian, Moroccan and Tunisian

[2] Available at: https://developers.google.com/YouTube

people. The issue is how to extract comments of the appropriate dialect. As there is no standard method to do it, we opted for the approach we proposed in a previous work [12], where we harvested data from YouTube, by selecting several hashtags or keywords specific to each country. Thereafter, a filtering and cleaning process were applied. Table 1 gives some statistics about the collected corpus, where $|C|$ denotes the number of comments, $|W|$ the number of words and $|V|$ the size of the vocabulary.

Table 1. Statistics of the harvested corpora.

	Algerian (M)	**Moroccan (M)**	**Tunisian(M)**
$\|C\|$	1.61	1.60	1.26
$\|W\|$	23	22	17
$\|V\|$	1.2	1.3	1

4 CESAR: A new measure for CodE-Switching According to a Reference language

Our main purpose is to quantify the code-switching phenomenon according to a reference language within a corpus. In other words, we aim to determine how much a document is clean relatively to a reference language. The scores proposed in [9] and [14] are not suitable for our purpose since they calculate the complexity factor (CF) of a corpus independently from any language. They are interested in the phenomenon of code-switching without paying attention to the tongue mother of the Internet user. The measure of code-switching complexity proposed in [14], named in the following Gosh measure, is given below:

$$CF = \frac{50 * \frac{W' - max\{w\}}{W'} + 50 * \frac{S}{W-1}}{\frac{W}{N}} \tag{1}$$

Where W is the number of words, N is the number of distinct languages in the utterance, S is the number of break points and $\frac{W' - max\{w\}}{W'}$ is the ratio of the number of words which are not written in the dominant language of the sentence over the total number of the language-dependent words present in the sentence.

Existing code-switching measures are not able to measure the cleanliness of a corpus relatively to a specific language, especially for those extracted from social networks. For instance, if the objective is to harvest an Algerian dialect corpus, then how to ensure that we will get a corpus with a minimum of noise? A suitable measure may accept or refuse a harvested sentence relatively to its degree of code-switching. In the following, we propose a new measure that takes into account the noise existing in a document in accordance to a reference language. This means if a text is composed of only words of this language, the score is equal

to zero and the more the text is code-switched the closer the score is to 1. For that, we define several parameters:

$$P_r(C) = \frac{1}{n}\sum_{i=1}^{n}\delta(d_i)LF(d_i) \tag{2}$$

Where $\delta(d_i)$ is defined as follows:

$$\delta(d_i) = \begin{cases} 1 \ if \ \exists w \in d_i \ \text{ where } \ L(w) \neq L_r \\ 0 \ else \end{cases} \tag{3}$$

Where $L(w)$ is the language of the word w, d_i is a document of the corpus C, n is the number of documents of C and $\delta(d_i)$ delivers 1 if at least one code-switching break point does exist in d_i. LF is the language factor that takes into account the number of languages in the document d_i. LF is equal to 0 if the document is composed only by the reference language, it is equal to one if no word in the document is written in the reference language and it is proportional to the number of languages different from the reference in the other case. $P_r(C)$ gives 0 if the corpus C is entirely written in the reference language. It delivers 1 if no document of C is code-switched and 0 if no document is written entirely in the reference language. In the other cases, $P_r(C)$ is equal to the proportion of the number of languages for each d_i of C.

$$B_r(C) = \frac{1}{n}\sum_{i=1}^{n}\frac{N(w_{d_i})}{N_{d_i}}LF(d_i) \ \text{ where } L(w_{d_i}) \neq L_r \tag{4}$$

Where $N(w_{d_i})$ is the number of words of the document d_i, for which the language of these words is different from the reference language. N_{d_i} is the total number of words of the document d_i, while $B_r(C)$ is the rate of the noise introduced by the words that are different from the reference language. Therefore, the CESAR measure allowing to estimate the code-switching in a corpus C according to a reference language r, is given below:

$$CESAR(C) = \alpha P_r(C) + \beta B_r(C) \tag{5}$$

Where α and β are the weights assigned to P_r and B_r, they are determined empirically by respecting the constraint $\alpha + \beta = 1$.

5 Experimentation

In the following, we will present several experiments in order to show the capacity of CESAR to measure the level of code-switching in a corpus.

5.1 Comparing Gosh and CESAR measures on elementary examples

In order to show that the measure of Gosh [14] is not adapted to our purpose, let us calculate the code-switching scores with this metric and compare it to CESAR on the examples given in Table 2.

Num	Sentence	Gosh	CESAR
S_1	$x_1^a x_2^a x_3^a x_4^a x_5^a$	0	1
S_2	$x_1^r x_2^r x_3^r x_4^r x_5^r$	0	0
S_3	$x_1^r x_2^e x_3^a x_4^r x_5^r$	34.5	0.29

Table 2. Code-switching scores with Gosh and CESAR

In the second column we give examples on which we calculated the code-switching scores, each word x_i^l is associated to its language l where $l \in \{$a: Arabic, r: reference and e: English$\}$. In the sentence S_1, no word from the reference language does exist. CESAR gives to this sentence the maximum value, while Gosh assigns it a 0. For S_2, the sentence is written entirely in the reference language. CESAR assigns it the minimum value and that is what we would like to get, while Gosh gives it a value of 0 which is the same value such as for the sentence S_1. For S_3, the sentence is code-switched, CESAR assigns it a small value depending on the number of used languages in the text. This sentence is not highly code-switched (only two words from foreign languages), whereas Gosh gives it a value of 34.5, a score which is difficult to interpret because Gosh's measure is not bounded, which makes it difficult to make a correlation between the score and the complexity of the sentence in terms of code-switching. If we duplicate the sentence S_3 such as: $x_1^r x_2^e x_3^a x_4^r x_5^r x_1^r x_2^e x_3^a x_4^r x_5^r$, the Gosh metric assigns it a score of 15.9 while CESAR gives it a value of 0.43. This means that Gosh assumes that the sentence is less complex, contrary to CESAR which considers it as more code-switched, which is really the case.

5.2 Measuring the code-switching degree for the three Maghrebian corpora

After showing the main difference between CESAR and Gosh measures by means of the examples above, in Table 3, we estimate the code-switching values of the corpora collected from YouTube. For each corpus, the reference language is the dialect of the corresponding country. CESAR leads to almost similar values for the three dialects. We notice that the Algerian corpus is the most code-switched one, even if the difference of scores is slight. Concerning the results obtained by the Gosh measure, the conclusion is opposite to the one given by CESAR. But, as mentioned before, the Gosh measure is not related to a reference language, and furthermore, it is not normalized. The longer the document, the lower the Gosh measure. Which does not reflect the actual situation within the meaning of the code-switching level of the document.

	Algerian	Moroccan	Tunisian
CESAR(C)	0.28	0.27	0.25
Gosh(C)	15.49	16.96	18.23

Table 3. Measuring the code-switching of the three Maghrebian dialects

5.3 The sensitivity of Gosh and CESAR measures on a balanced corpus

In order to compare the two measures, we achieved several experiments on the Algerian dialect, one of the languages of our corpus, named C in the following experiments. This corpus contains 2000 comments which are not code-switched, they are entirely written in the reference language (The Algerian dialect). The particularity of C is that all the comments have the same size, more precisely each of them is composed of fifty words. We conducted three separate experiments E_1, E_2 and E_3 on the corpus C. For each experiment E_i, we regularly injected new words (one by one) in the comments of C in order to measure the sensitivity of each measure in accordance to the added words. If these words come from the reference language, that means that no code-switching does exist in the comment and consequently the initial measures should not change. In the case where the added words come from other languages, the values of the measures should be modified since the comments are henceforth code-switched.

In each experiment E_i, we added randomly fifty words, one at each step and then we reported the measures of Gosh and CESAR in Figure 1. In the Experiment E_1, the added words come from the same language. In the experiment E_2, the added words are French, while in the experiment E_3, the added words are French, Arabic and English.

The objective of the experiment E_1 (the green curve) is to analyse the behaviour of the two measures when the documents are not code-switched. The CESAR and Ghosh values are constant, which demonstrates that the two mesure evaluat correctly the level of code-switching in the corpus.

In the experiment E_2, represented by the blue curves, we added French words to the comments composed only by dialect words. This operation makes the comments code-switched, which theoretically should increase the values of Gosh and CESAR. The Gosh's measure increased when we added the first foreign word, that is what was expected, but unfortunately the measure started decreasing from the introduction of the second word and this, until the end. While, CESAR's measure increases in accordance to the number of the injected foreign words.

Similarly to E_2, we introduced foreign words in the experiment E_3, represented by the black curves. The only difference is that we introduced words in several languages in order to appreciate the possibility of the measures to capture the level of code-switching in the comments composed by several languages. As in the previous experiment, the Gosh's measure increased in the beginning and then dropped continuously, while the CESAR's values increased until the end of the process of the words injection.

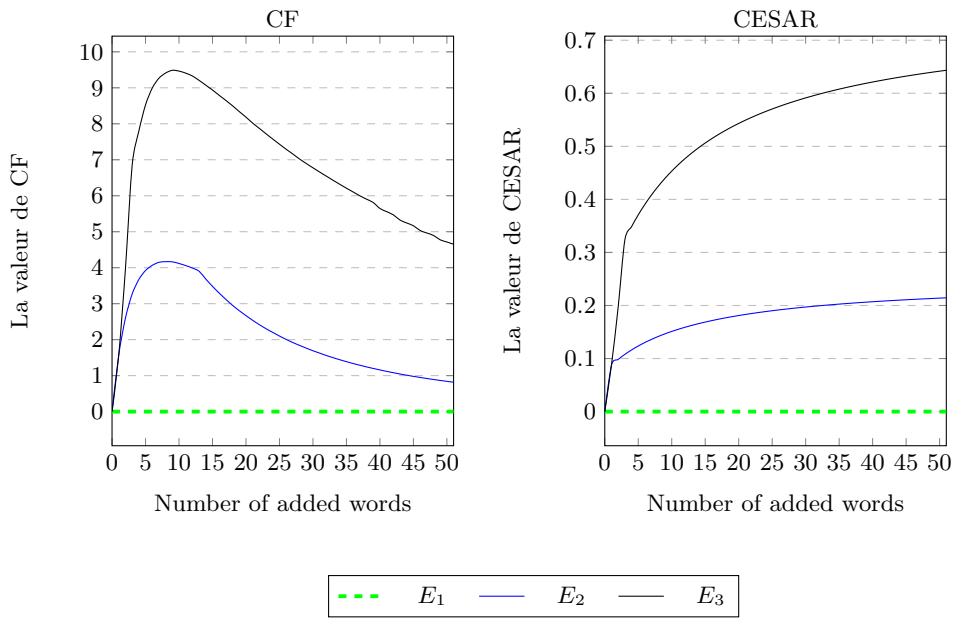

Fig. 1. The progression of Gosh and CESAR in accordance to document with and without code-switched segments

6 Conclusion

The purpose of this paper is to study the dialects of the three Maghreb countries, namely Algeria, Morocco and Tunisia. We were particularly interested in their degree of code-switching. This phenomenon usually occurs in daily conversations and through social networks in several countries especially in Arab world, India, Philippines, in few regions in switzerland, etc. To this end, we collected three corpora from Youtube concerning the comments made by Algerians, Moroccans and Tunisians. Then, we measured the level of code-switching for each dialect for which there are no dictionaries. There are measures to meet this objective, however, those we studied including the measure of Gosh are not suited to our needs. In fact, most of the measures merely estimate the degree of code-switching. Our goal is not only to do this, but also to locate the level of code-switching compared to a reference language. In other words, we want to know how much the text is polluted relatively to a given language. That is why, we proposed a new measure named CESAR which is a bounded metric that associates 0 to a document in which the text is entirely in the studied language (reference) without any break point. A maximum value equal to 1, is assigned to any document without any word in the reference language. When the text is a mixture of several languages, CESAR assigns it a value between 0 and 1 according to its level of code-switching. This measure was compared to Gosh's measure in several experiments on elementary samples and on the three collected corpora. The

experiments showed that CESAR is well adapted to measure the uncleanness of a text regarding other languages. Interestingly, this measure could be used to extract corpora containing a minimum number of foreign phrases by minimizing the value of CESAR.

References

1. A. K. Joshi, "Processing of Sentences with Intra-sentential Code-switching," in *Proceedings of the* 9th *Conference on Computational Linguistics - Volume 1*, COLING '82, pp. 145–150, 1982.
2. P. Auer, "From codes-witching via language mixing to fused lects: Toward a dynamic typology of bilingual speech," *International Journal of Bilingualism*, vol. 3, no. 4, pp. 309–332, 1999.
3. J. Gafaranga and M.-C. Torras, "Interactional otherness: Towards a redefinition of code-switching," *International Journal of Bilingualism*, vol. 6, no. 1, pp. 1–22, 2002.
4. A. A. Tawwab and S. Eldin, "Socio Linguistic Study of Code Switching of the Arabic Language Speakers on Social Networking," *International Journal of English Linguistics*, vol. 4, no. 6, 2014.
5. A. Alhazmi, "Linguistic Aspects of Arabic-English Code Switching on Facebook and Radio in Australia," *International Journal of Applied Linguistics and English Literature*, vol. 5, no. 3, 2015.
6. R. Redouan, " Linguistic Constraints on Code-switching and Code-mixing of Bilingual Moroccan Arabic-French Speakers in Canada ," in *Proceedings of the* 4th *International Symposium on Bilingualism, MA: Cascadilla Press*, 2005.
7. Y. Yeong and T. Tan, "Language identification of code switching sentences and multilingual sentences of under-resourced languages by using multi structural word information," in 15th *Annual Conference of the International Speech Communication Association, Singapore,Interspeech*, pp. 3052–3055, 2014.
8. M. Al-Badrashiny, H. Elfardy, and M. Diab, "AIDA2: A Hybrid Approach for Token and Sentence Level Dialect Identification in Arabic," in *Proceedings of the* 19th *Conference on Computational Natural Language Learning, CoNLL, Beijing, China*, pp. 42–51, 2015.
9. B. Gambäck and A. Das, "Comparing the level of code-switching in corpora," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC, Portorož, Slovenia*, 2016.
10. K. Abidi and K. Smaïli in *International Conference on Natural Language, Signal and Speech Processing (ICNLSSP)*, (Casablanca, Morocco), 2017.
11. D. Amazouz, M. Adda-Decker, and L. Lamel, "The French-Algerian Code-Switching Triggered audio corpus (FACST)," in 11th *edition of the Language Resources and Evaluation Conference, LREC*, 2018.
12. K. Abidi, M. A. Menacer, and K. Smaïli, "CALYOU: a Comparable spoken ALgerian corpus extracted from YOUTube," in 18th *Annual Conference of the International Speech Communication Association, Stockholm Sweden, Interspeech*, 2017.
13. M. Menacer, O. Mella, D. Fohr, D. Jouvet, D. Langlois, and K. Smaïli, "Development of the Arabic Loria Automatic Speech Recognition system (ALASR) and its evaluation for Algerian dialect," in *Third International Conference On Arabic Computational Linguistics, Dubai*, 2017.

14. S. Ghosh, S. Ghosh, and D. Das, "Complexity metric for code-mixed social media text," *CoRR*, vol. abs/1707.01183, 2017.
15. Ö. Çetinoglu, "A turkish-german code-switching corpus," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC, Portorož*, 2016.
16. Ö. Çetinoglu, "A code-switching corpus of turkish-german conversations," in *Proceedings of the* 11th *Linguistic Annotation Workshop, Valencia, Spain*, pp. 34–40, 2017.
17. V. Ramanarayanan and D. Suendermann-Oeft, "Jee haan, I'd like both, por favor: Elicitation of a Code-Switched Corpus of Hindi-English and Spanish-English Human-Machine Dialog," in 18th *Annual Conference of the International Speech Communication Association, Stockholm, Sweden, Interspeech*, pp. 47–51, 2017.
18. Y. Samih and W. Maier, "An arabic-moroccan darija code-switched corpus," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC, Portorož, Slovenia*, 2016.
19. A. Jaech, G. Mulcaire, M. Ostendorf, and N. A. Smith, "A neural model for language identification in code-switched tweets," in *Proceedings of the Second Workshop on Computational Approaches to Code Switching, Austin, Texas, USA,EMNLP*, pp. 60–64, 2016.
20. S. K. Rallabandi and A. W. Black, "On building mixed lingual speech synthesis systems," in 18th *Annual Conference of the International Speech Communication Association, Stockholm, Sweden, Interspeech*, pp. 52–56, 2017.