INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Addressing Code-Switching in French/Algerian Arabic Speech

Djegdjiga Amazouz [1], *Martine Adda-Decker*[1,2], *Lori Lamel*[2]

[1]LPP-CNRS Université Sorbonne Nouvelle Paris-III, Paris, France
[2]LIMSI, CNRS, Paris Saclay University, Orsay, France
djegdjiga.amazouz@univ-paris3.fr, madda@limsi.fr, lamel@limsi.fr

Abstract

This study focuses on code-switching (CS) in French/Algerian Arabic bilingual communities and investigates how speech technologies, such as automatic data partitioning, language identification and automatic speech recognition (ASR) can serve to analyze and classify this type of bilingual speech. A preliminary study carried out using a corpus of Maghrebian broadcast data revealed a relatively high presence of CS Algerian Arabic as compared to the neighboring countries Morocco and Tunisia. Therefore this study focuses on code switching produced by bilingual Algerian speakers who can be considered native speakers of both Algerian Arabic and French. A specific corpus of four hours of speech from 8 bilingual French Algerian speakers was collected. This corpus contains read speech and conversational speech in both languages and includes stretches of code-switching. We provide a linguistic description of the code-switching stretches in terms of intra-sentential and inter-sentential switches, the speech duration in each language. We report on some initial studies to locate French, Arabic and the code-switched stretches, using ASR system word posteriors for this pair of languages.

Index Terms: Code-switching, Language Identification, Algerian Arabic, French.

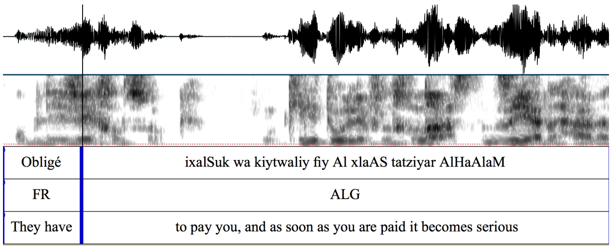
1. Introduction

Code-switching (CS) is the process of switching from one language to another in the same conversational sequence [1, 2, 3]. CS is a spoken phenomenon of bilingual and multilingual communities [4, 5]. Given the high number of bilingual speakers worldwide, the phenomenon of CS is extremely frequent although it has not yet been frequently studied by the automatic speech processing community.

Bilingual speakers may spontaneously introduce words and phrases from one language (the *embedded* language) placed within the basic language (the *matrix* language). CS may take different forms within the speech, linguists distinguish between *inter-sentential* and *intra-sentential* CS [6]. We notice that there are different terms used to refer to bilingual speech phenomena that we consider CS such as *mixed structure* [7], *mixed code* [5], *transcodic marks*, *loan words*. All these phenomena, that are included in this present CS study, are typically studied in interactional sociolinguistics [8, 5], language learning/acquisition [9, 10] as well as general linguistics. More recently, CS speech has also raised interest in computational linguistics and automatic speech processing research [11, 12, 13]. A lot of the studied CS data come from media and web podcast corpora. These types of corpora are relatively easy to obtain and may contain large amounts of data. However, for acoustic-phonetic studies, media data are often inadequate as the acoustic signal quality may be low (noise, weak sound quality and audibility) and they may present some difficulties to process (low quantity of CS, high speech rate, overlapping speech).
One of the aims of this study is to collect and annotate a French/Arabic CS speech corpus of high acoustic quality and with a relatively high density of CS. To this purpose, we selected speakers who are used to language switch in their daily lives. The data were not collected in a field study, the speakers were invited to come to our lab. However, we designed a special protocol to collect CS data in different production rates and styles. The potential usages of the data are multiple: classification of foreign accent, work on L2 pronunciation learning, phonetic and prosodic comparisons of monolingual and bilingual speech.

In this paper we will first present related work, before introducing our corpus collection and the resulting data. Too, we address the challenge of how to obtain a CS corpus in laboratory environment by creating a sociolinguistic context favoring speech in two languages. As the corpus annotation is still underway, we provide some partial statistics on the content and some first experiments at detecting CS segments.

Figure 1 illustrates an example of intra-sentential French/Arabic code switching. The sentence starts with a word in French and is then completed in Arabic. This figure provides annotations at the word level, the language segmentation and tags, and the English translation (bottom line).



Obligé	ixalSuk wa kiytwaliy fiy Al xlaAS tatziyar AlHaAlaM
FR	ALG
They have	to pay you, and as soon as you are paid it becomes serious

Figure 1: *An example of intra-sentential FR (French) / AA (Algerian Arabic) codeswitching.*

2. Related Work

From LID and ASR perspectives, CS has received attention in several studies using a variety of data from speech web data to carefully recorded speech. In Mandarin-Taiwanese CS, switches of language where a high CS was observed, Lyu and Lyu [14] used three sets of data for their LID system: a monolingual training data set for Mandarin, another monolingual training set for Taiwanese, and finally a code-switching test data set (Test-CS). The Test-CS data consist of 4.8 hours of speech and it corresponds to 4600 utterances. Their LID system proposes to evaluate the CS and contribute to error reduction for LID and ASR. In the same line, Lyu et al., [15] designed the SEAME (Mandarin–English code-switching speech corpus in South-East Asia) corpus with 93 hours recorded from media

http://dx.doi.org/10.21437/Interspeech.2017-1373

interviews of Singaporean and Malaysian speakers who mix Mandarin and English in their speech. In their corpus, intra-sentential CS represented 30 hours of spontaneous Mandarin-English with Mandarin being the dominant language. In CS radio broadcast data, Modipa et al., [16] tested the performance of ASR system in Sepedi spoken mixed with English as dominant language, detection of the language change in the utterance. They analysed the mechanism of CS in this pair of languages by quantifying the frequency of CS. An acoustic analysis was also done using this corpus in order to develop an ASR system. They produced and used a dictionary for the two languages in a single bilingual ASR system.

3. Research questions

In this paper, we pose and attempt address the following questions:

i. Is CS a common practice in Maghrebian broadcast speech and in which country is CS practice more frequent?

ii. How can we collect conversational speech including code-switching? Are there methods to incite CS?

iii. Can code-switching be detected using LID systems [17]? what is the minimum time span to be successful?

iv. Can we detect code-switching using ASR systems?

v. How can speech technologies help in quantifying different types of CS?

The first question is addressed using a large collection of broadcast and podcast speech data from the Maghreb countries referred to as the MCSM corpus. The results of this preliminary study on Arabic/French code-switching guided our choices to collect a CS corpus of read and spontaneous speech focused mainly on Algerian Arabic (question ii) . The last three questions deal with evaluating the potential help of automatic speech processing tools to detect code-switching and/or in providing support in annotating such corpora for further linguistic studies.

4. The French-Algerian FACST corpus

In this section we present the bilingual community and describe the corpus and the different steps of the data collection. We also provide some basic statistics of the collected CS data.

4.1. Code-switching in FR/AA bilingual community

Algerian Arabic (AA) is one of the Arabic dialects used in informal situations and in daily life contrary to Modern Standard Arabic (MSA) which is mainly used in formal communication as in eduction, press news, media and political speeches. AA is the mother tongue of about 40 millions of Algerians. It is an oral language and has few written resources. In Algeria, French is the major second language inherited from the bygone colonial era and still learned during education. It is also used in daily life in contact with AA. French is the language of several scientific faculties and it is used by part of Algerian press and media. Hence, AA has been in contact with FR for historical and educational reasons over many decades [18]. As a consequence, Algerian people tend to speak fluently both French and AA, and CS phenomena frequently appear in their daily communication. One can notice that, in the Algerian community, a lot of languages and dialects coexist: Arabic and its dialects, Berber and its variants, French in the larger cities of the country and Spanish in the North West of Algeria. In this study, we focus on spoken *Algérois* [19], the AA dialect of the Algiers area. This region shows a high degree of contact between AA and FR, where more CS practice can be expected than in other regions of Algeria. The situation of CS in France, has a tendency to be dominated by the French language. In their professional or educational contexts, Algerians tend to prefer French as basic communication language, while occasionally switching to AA.

4.2. Corpus

Prior to collecting our French/Arabic CS speech corpus in France, we made a preliminary study on the MCSM database (Maghrebian Code-switching in Media) [20], which consists of a collection of Maghrebian broadcasts from Algeria, Morocco and Tunisia. In this study, we compared the quantity of CS occurring in dialectal Arabic with a focus on code switches to French. The corpus consists of 53 hours of television media containing entertainment and TV talk shows in Arabic dialects: Algerian (14 hours), Moroccan (15 hours), Tunisian (24 hours). This study aimed at answering the following question: is Arabic-French CS frequent in Maghreb dialects? In which country is the CS effect the strongest? Table 1 displays the figures describing the MCSM corpus in terms of total duration and number of CS segments and duration. Results show that Maghrebian bilingual communities all use Arabic-French CS in media, but the quantity of CS varies with the country. It can be seen that Algerian/French CS is much more frequent (148 CS segments per hour) than Moroccan/French CS (62 CS segments per hour) and Tunisian/French (21 CS segments per hour).

Table 1: *CS segment quantities and number of CS speakers of each country in MCSM*

Country show	Algeria	Morocco	Tunisia
Nb. of CS segments	2081	938	509
Total show duration	14h	15h	24h
Nb. CS/hour	148.6	62.5	21.2

Although the number of switches is relatively high, the segments tend to be of very short duration and typically include either discourse markers and adverbs (alors 'then', donc 'so', justement 'precisely', d'accord 'OK',) or technical vocabulary such as fashion products. This latter switches can be considered as loan word switches. Given these results on the media data, we decided to focus our CS corpus collection on the French/Algerian pair of languages as the chance to find code-switching speakers seems to be higher.

4.2.1. FACST

The FACST corpus (French Arabic Code-switching Triggered) consists of records of CS conversations with bilingual adult speakers who tend to code-switch in there daily lives. At present, FACST includes 8 bilingual speakers aged from 20 to 35. They all have lived a part of their lives in Algeria and another in France. They all have studied at university and use both languages daily and often with CS. We recorded the speakers in a soundproof room at LPP (Labortoire de Phonétique et Phonologie) of Sorbonne-Nouvelle University in Paris. The corpus is intended to serve multiple research purposes in CS including language identification studies, language boundary detection, as well as phonetic and prosodic issues in CS speech and other linguistics studies. The aim of recording these conversations at LPP was on one hand to get speech with a high quantity of CS in the records, on the other hand, recordings in a soundproof booth ensure a high acoustic quality of the signal.

4.2.2. Data collection

We selected the participants using a sociolinguistic online questionnaire (ECSP 2016) *"Experience of Code-switching practice"* with questions about their linguistic autobiography, the environment in which languages are practised, language acquisition/learning, CS habits. . .

Recording sessions started with a preliminary unrecorded conversation with the speaker to get her/him in a relaxed setting practicing both languages in the same interaction. First, we asked the speakers to perform oral readings of two texts with three different speech rates (slow, normal, fast), one in AA and and other one in FR. We asked the speakers to read a text excerpt of "Le Petit Prince" for French, and an excerpt from the movie scenario "Bab El-Oued city" in Algerian Arabic. The stimuli of the controlled read speech are summarized in Table 2.

Table 2: *Controlled read speech in FR and AA. Number of words and average reading time in seconds (slow-medium-fast)*

Language	**Nb. of words**	**Average reading time for each rate (s)**
FR	185	92 - 60 - 55
AA	102	50 - 37 - 30

The goal of this experiment was, firstly to obtain a controlled monolingual speech corpus in AA and FR for bilingual speakers before proceeding to the bilingual speech. Secondly, we would like to highlight potential pronunciation differences of consonants and vowels in each language separately. Thirdly, using the three speech rate task, we intend to study realisation differences in consonants and vowels.

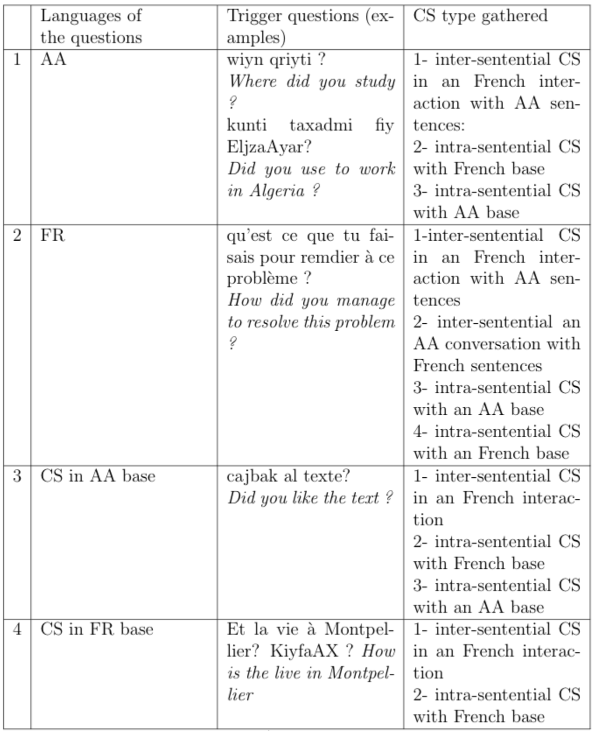
The second step consisted of recording dual conversations between the linguist (who is a bilingual speaker of both languages) and the speaker. The CS conversations are triggered by questions about life and study in both countries. So, the role of the linguist is to ask the questions and let the speaker answer freely and hopefully make use of CS. The recordings lasted from 15 to 25 minutes for each speaker. Figure 2 illustrates the methodology used to entice natural CS conversations. The linguist asked questions in one or the other of the targeted languages (FR or AA) also including CS to stimulate spontaneous speech with FR/AA code-switching.

The main role of the linguist here was to create a sociolinguistic context in which conversational CS occurs naturally in guided dialogs. However, we want to point out that this collection experiment, carried out in a soundproof room, is clearly not a sociolinguistic survey in the field – it is an attempt to trigger similar naturalistic CS phenomena in a lab setting.

4.2.3. Annotation

We used the Transcriber [21] program to manually segment the audio utterances into sentences, breath groups, speech turns, languages used. We also annotated the speakers and the language of the segments as follows:

```
Speakers    Time           Gender     Language
Speaker5    13.75-15.33    <male>     ALG
Speaker7    329.54-331.49  <male>     FRA
Speaker8    17.86-18.83    <female>   FRA
```



	Languages of the questions	Trigger questions (examples)	CS type gathered
1	AA	wiyn qriyti ? *Where did you study ?* kunti taxadmi fiy EljzaAyar? *Did you use to work in Algeria ?*	1- inter-sentential CS in an French interaction with AA sentences: 2- intra-sentential CS with French base 3- intra-sentential CS with AA base
2	FR	qu'est ce que tu faisais pour remdier à ce problème ? *How did you manage to resolve this problem ?*	1-inter-sentential CS in an French interaction with AA sentences 2- inter-sentential an AA conversation with French sentences 3- intra-sentential CS with an AA base 4- intra-sentential CS with an French base
3	CS in AA base	cajbak al texte? *Did you like the text ?*	1- inter-sentential CS in an French interaction 2- intra-sentential CS with French base 3- intra-sentential CS with an AA base
4	CS in FR base	Et la vie à Montpellier? KiyfaAX ? *How is the live in Montpellier*	1- inter-sentential CS in an French interaction 2- intra-sentential CS with French base

Figure 2: *Examples of questions asked by the linguist to trigger spontaneous FR/AA CS in the subject's responses. Type of CS produced that follows the questions in the speech.*

The transcription of the speech has been done after the segmentation and language annotation. The AA speech was transcribed using a transliteration orthography inspired by Buckwalter Arabic transliteration [22]. We chose this transcription convention in order to facilitate the use of the manual transcripts for phonetic analyses while keeping the possibility to convert the transliterated characters to Arabic charterers in future studies. For the time being, the speech data from 5 subjects has been manually transcribed. Their production in terms of words and duration by language are shown in Table 3. Figure 3 shows a histogram of length of CS speech segments for FR/AA.

Table 3: *Some statistics about the FACST spontaneous speech. Number of words and duration of CS speech in each language.*

Speakers and gender	**FR words**	**Duration**	**AA words**	**Duration**
S 5 M	1619	531.11s	348	118.20s
S 3 F	1283	480.88s	369	138.97s
S 7 M	1157	425.66s	243	90.93s
S 8 F	829	391.72s	277	145.56s
S 2 M	277	118.16s	285	116.67s

5. Speech Technologies for Code-Switching

In the following, we report on some first experiments to detect CS using automatic speech processing tools. First, the speech files were automatically segmented into acoustically homogeneous segments, which ideally correspond to speaker turns and/or to a given language or stable acoustic conditions (broad

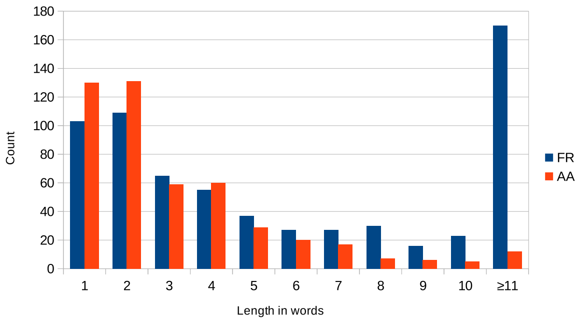


Figure 3: *Histogram of the CS segment lengths for FR/AA.*

band/telephone band...). These segments were then automatically transcribed using different ASR systems [23, 17] in parallel: a French system, a multi-dialect Arabic system (predominantly Lebanese) and an Algerian Arabic (dialect) system. The systems were trained on several hundreds of hours of speech from a large number of speakers. The Algerian models are the result of adapting the multi-dialect Arabic system with about 300 hours of data from Algerian speakers. Our expectation is that the French system will produce the highest scores on French speech and vice versa, that AA speech will be best decoded by one of the two Arabic systems. Figure 4 shows the French system's confidence scores on an excerpt of speech including CS. The x-axis corresponds to the word numbers. Words on the left (70-79) and right (85-89) are in French and the middle words are in Arabic. The purple curve shows the French ASR word posteriors, which as can be expected are higher for French than for Arabic. The green curve is a smoothed version of the posteriors, as the raw values are quite brittle. The bottom line specifies the true language (0.1 is French, 0.2 is Arabic).

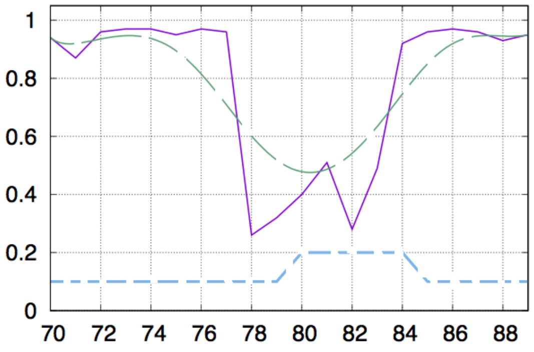


Figure 4: *ASR word posteriors of the French transcription systems (raw scores and smoothed). The X-axis corresponds to an excerpt of speech: words numbered from 70-79 and 85-89 are in FR, words numbered from 80-84 correspond to an AA code-switch. The lowest curve (blue) denotes FR (0.1) or AA (0.2).*

Table 4 gives the average word posteriors (confidence scores) for each speaker as a function of the manually annotated language and the ASR system used to transcribe the data. Overall, and for most speakers, a higher confidence is obtained when the data is processed with the matching system.

Table 4: *Average word level posterior scores by speaker with the French (left) and Algerian (right) ASR systems, for words of the CS segments manually annotated as French and Arabic.*

CS language →	**Fre**	**Ara**	**Fre**	**Ara**
Speaker	**French ASR**		**Arabic ASR**	
Speaker 5	**0.74**	0.72	0.54	**0.56**
Speaker 3	**0.79**	0.57	0.58	**0.65**
Speaker 7	0.74	0.78	0.51	**0.54**
Speaker 8	**0.78**	0.61	0.56	**0.58**
Speaker 2	**0.62**	0.59	0.59	0.56
Overall	**0.76**	0.64	0.56	**0.59**

6. Conclusions

In this paper, we presented our ongoing studies on French/Algerian Arabic code switching in spoken language. We tried to answer several different research questions. First, we used a large set of media data from the Maghrebian countries to measure the Arabic/French code switching reality in entertainment shows. This study revealed that the Algerian shows featured a much higher CS rate as compared to Tunisian and Moroccan broadcast shows. We described a speech collection protocol to record both a monolingual speech and a CS speech corpus testing different ways to elicit or trigger natural code-switching in spontaneous speech. The purpose of the corpus is to facilitate acoustic-phonetic studies of segments in bilingual and CS situations. At present, the corpus contains 8 hours of speech which are still in process of being manually segmented, annotated and transcribed. The four hours of transcribed data which have been used in this work show a high density of code-switching with mainly short intra-sentential CS, although some segments are of longer duration and cover stretches of CS beyond sentential boundaries. These first results confirm that our CS triggering protocol was successful and that using it we were able to obtain spontaneous code switched speech with both inter- and intra-sentential CS. Concerning the question as to whether CS can be automatically detected, we presented first results using three different ASR systems in parallel (French, multi-dialectal Arabic, Algerian Arabic). These results show that short duration CS segments poses serious challenges to automatic language identification in CS speech, although parallel ASR systems may produce word posteriors which are good indicators of language change. We will pursue this line of investigation in our future work. Concerning our last question on the use of speech technologies to study CS, our present assessment is that automatic speech transcription is of great help to achieve a high quality transcription and temporal alignments into words and phones which opens new perspectives for large scale CS studies on acoustic-phonetic and prosodic levels.

7. Acknowledgements

This research is supported in one hand by the French National Agency for Research as part of the SALSA project (Speech and Language technologies for Security Applications) project under grant ANR-14-CE28-0021 and in the other by the French Investissements d'Avenir - Labex EFL program (ANR-10-LABX-0083). The study was supported by LPP-CNRS Paris-III Sorbonne Nouvelle University and LIMSI-CNRS Paris-Orsay University. We thank the participants who contributed to the FACST corpus.

8. References

[1] C. Myers-Scotton, *Duelling languages: Grammatical structure in codeswitching*. Oxford University Press, 1993.

[2] C. M. Scotton, *Multiple voices: An introduction to bilingualism*. Blackwell Pub., 2006.

[3] L. Isurin, D. Winford, and K. De Bot, *Multidisciplinary approaches to code switching*. John Benjamins Publishing, 2009, vol. 41.

[4] P. Auer, *Language and Space: An International Handbook of Linguistic Variation. Theories and Methods*. Walter de Gruyter, 2010, vol. 1.

[5] ——, *Code-switching in conversation: Language, interaction and identity*. Routledge, 2013.

[6] H. Kebeya, "Inter-and intra-sentential switching: are they really comparable?" Ph.D. dissertation, Kenyatta University, 2013.

[7] C. Canut and D. Caubet, *Comment les langues se mélangent: codeswitching en francophonie*. Editions L'Harmattan, 2001.

[8] J. J. Gumperz, *Discourse strategies*. Cambridge University Press, 1982, vol. 1.

[9] M.-J. Ezeizabarrena and S. Aeby, "Les phénomènes de code-switching dans les conversations adulte-enfant (s) en basque-espagnol: une approche syntaxique," *Corpus*, no. 9, 2010.

[10] D. Moore, "Code-switching and learning in the classroom," *International journal of bilingual education and bilingualism*, vol. 5, no. 5, pp. 279–293, 2002.

[11] N. T. Vu, D.-C. Lyu, J. Weiner, D. Telaar, T. Schlippe, F. Blaicher, E.-S. Chng, T. Schultz, and H. Li, "A first speech recognition system for mandarin-english code-switch conversational speech," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE, 2012, pp. 4889–4892.

[12] E. Yılmaz, H. van den Heuvel, and D. van Leeuwen, "Investigating bilingual deep neural networks for automatic recognition of code-switching frisian speech," *Procedia Computer Science*, vol. 81, pp. 159–166, 2016.

[13] T. Solorio and Y. Liu, "Learning to predict code-switching points," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2008, pp. 973–981.

[14] D.-C. Lyu and R.-Y. Lyu, "Language identification on code-switching utterances using multiple cues." in *Interspeech*, 2008, pp. 711–714.

[15] D.-C. Lyu, T.-P. Tan, E.-S. Chng, and H. Li, "Mandarin–english code-switching speech corpus in south-east asia: Seame," *Language Resources and Evaluation*, vol. 49, no. 3, pp. 581–600, 2015.

[16] T. I. Modipa, M. H. Davel, and F. De Wet, "Implications of sepedi/english code switching for asr systems," 2013.

[17] A. Laurent, T. Fraga-Silva, L. Lamel, and J.-L. Gauvain, "Investigating techniques for low resource conversational speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 5975–5979.

[18] D. Caubet, "Métissages linguistiques ici (en france) et là-bas (au maghreb)," *Ville-école-intégration enjeux*, vol. 130, pp. 117–132, 2002.

[19] H. Saadane and N. Habash, "A conventional orthography for algerian arabic," in *ANLP Workshop 2015*, 2015, p. 69.

[20] D. Amazouz, M. Adda-Decker, and L. Lamel, "Arabic-french code-switching across maghreb arabic dialects : a quantitative analysis," in *Workshop "Corpus-driven studies of heterogeneous and multilingual corpora"*, 2016, pp. 5–7.

[21] C. Barras, E. Geoffrois, Z. Wu, and M. Liberman, "Transcriber: development and use of a tool for assisting speech corpora production," *Speech Communication*, vol. 33, no. 1, pp. 5–22, 2001.

[22] T. Buckwalter, "Arabic transliteration," *URL http://www. qamus. org/transliteration. htm*, 2002.

[23] J.-L. Gauvain, L. Lamel, H. Schwenk, G. Adda, L. Chen, and F. Lefevre, "Conversational telephone speech recognition," in *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*, vol. 1. IEEE, 2003, pp. I–I.