The French-Algerian Code-Switching Triggered audio corpus (FACST)

Djegdjiga Amazouz, Martine Adda-Decker, Lori Lamel

The French-Algerian Code-Switching Triggered audio corpus (FACST)

Djegdjiga Amazouz
Université de la Sorbonne Nouvelle Paris 3

Lori Lamel
French National Centre for Scientific Research

The French-Algerian Code-Switching Triggered audio corpus (FACST)

Djegdjiga Amazouz[1], **Martine Adda-Decker**[1,2], **Lori Lamel**[2]
[1]LPP-CNRS Université Sorbonne Nouvelle Paris-III
[2]LIMSI, CNRS, Paris Saclay University, Orsay, France
djegdjiga.amazouz@univ-paris3.fr, {madda, lamel}@limsi.fr

Abstract

The French Algerian Code-Switching Triggered corpus (FACST) was created in order to support a variety of studies in phonetics, prosody and natural language processing. The first aim of the FACST corpus is to collect a spontaneous Code-switching speech (CS) corpus. In order to obtain a large quantity of spontaneous CS utterances in natural conversations experiments were carried out on how to elicit CS. Applying a triggering protocol by means of code-switched questions was found to be effective in eliciting CS in the responses. To ensure good audio quality, all recordings were made in a soundproof room or in a very calm room. This paper describes FACST corpus, along with the principal steps to build a CS speech corpus in French-Algerian languages and data collection steps. We also explain the selection criteria for the CS speakers and the recording protocols used. We present the methods used for data segmentation and annotation, and propose a conventional transcription of this type of speech in each language with the aim of being well-suited for both computational linguistic and acoustic-phonetic studies. We provide an a quantitative description of the FACST corpus along with results of linguistic studies, and discuss some of the challenges we faced in collecting CS data.

Keywords: Code-switching, bilingual speakers, oral speech data, French, Arabic

1. Introduction

When multilingual speakers communicate using two shared languages in a conversation, they may switch between these languages in a same sentence or utterance. Switching may become a common practice in social groups with strong language contact. This simultaneous use of two language codes can take several forms at the lexical, syntactic and communication levels, and is commonly referred to as code-switching (CS) (Muysken, 2000; Grosjean, 1995). CS may present complex structures in particular at morphological, syntactical and phonetic levels. Hence, CS has interested several research fields such as sociolinguistics and interactional linguistics (Gumperz, 1982; Auer, 2010; Bullock, 2012; Gardner-Chloros, 2009; Sebba, 2012; Tabouret-Keller and Page, 1970; Ziamari, 2008; Piccinini, 2012; Gullberg et al., 2012) with numerous studies describing CS phenomena, explaining the process of its realization and analyzing its productions. Computational linguistics enable the study of large textual and spoken corpora (Çetinoğlu et al., 2016; Schultz and Kirchhoff, 2006). Consequently, experiments have been carried out on language identification (LID) for both oral and written CS as well as automatic identification of CS languages pairs (Lyu et al., 2015; Lyu and Lyu, 2008; Modipa et al., 2013). Furthermore, quantifying CS in large corpora at the word and oral segment levels gives a precise idea about the use of CS in bilingual communities (Amazouz et al., 2016). It also allows the determination of which language is dominant in the CS speech and in the daily CS practices. Applying language processing tools can help to compare CS in different language pairs and to determine how frequent CS is for pairs of languages.

In this paper, we focus on eliciting spontaneous CS and how to collect CS data. And, we question about methods used to build this hybrid oral speech. So, this work describes the design, recording, and annotation of the French Algerian Code-switching Triggered corpus (FACST) being created to support a variety of linguistic studies. Using a methodology designed to elicit CS speech, of 20 bilingual Algerian/French speakers have been recorded in a quiet room.

The paper is organized as follow: presentation of FACST corpus, selection of CS speakers, the recording protocols and the stimuli of CS, and CS data processing.

2. Algerian Arabic dialect, French and CS

The Algerian Arabic dialect (AA) is an oral dialect of North African Arabic dialects group spoken in Algeria and it is the mother tongue of 80% of Algerians. AA is different from Modern Standard Arabic (MSA) at several levels: lexicon, phonetics, phonology, syntax and morphology (Saadane and Habash, 2015; Souag, 2006). MSA is mainly a written language while AA has few written resources. But AA written form becomes more and more widespread especially in social medias. Commonly, AA is written with Arabic characters Table 5 of Appendix. This script is written from right to left. Too, another form of AA script transliterated with Latin characters called "Arabizi transliteration" used heavily on the internet and SMS (Cotterell et al., 2014; Al-Badrashiny et al., 2014; Bies et al., 2014). Thus, AA can be written with two forms.

French language (FR) is the first foreign language spoken in the Algerian community and is for the most part the second language for Algerians. This bilingual community has tens of millions of bilingual speakers who live in Algeria and in France. CS is increasingly part of their daily communications.

Code-switching is a general term referring to language change within a given conversation or an utterance. Code-switching can take many different forms, such as inter-sentential CS (language changes at sentence boundaries), intra-sentential CS (CS within sentences) (Kebeya, 2013), Code-mixing (insertion of L2 word within L1 utterance) (Muysken, 2000), borrowing, and bilingual verbs (L1 verb inserted in L2 form) (Muysken, 2000).

CS is also characterized by individual choices and individual forms as a placement of adverbs in the beginning or at the end of the sentences, the inclusion or deletion of articles/particle at the switch moment and the construction of CS sentences. In intra-sentential CS, speaker may produce in ungrammatical sentences due to the non-correlation of the two grammar codes but the sentences are semantically correct (Tossa, 1998). In FACST corpus, we used this general term CS for these transcodic marks.

3. FACST corpus

Table 1: Compact FACST presentation

Label	French Algerian Code-switching Triggered (FACST) corpus
Languages	French (FR), Algerian Arabic (AA)
Speakers	20 speakers: 10 male, 10 female Ages: 23-39
Duration	Recordings ranging from 15 to 40 minutes/speaker Total: 7 h 30 of speech
Content	Read speech and stimulated spontaneous speech.
Year	2016-2018

Building an oral CS corpus requires careful attention, in particular on oral languages as Algerian Arabic dialect (AA). Indeed, comparing French language (FR) and AA, AA is an oral language with low written resources unlike French which is standardized with more written resources and more steady in grammar. CS corpus involves selection of speakers who represent CS speech of the linguistic community. It requires too an appropriate methodology to ensure that the recordings contain sufficient instances of CS. Corpus annotation is challenging, requiring guidelines for segmentation of CS and transcription of dialectal speech. In the FACST corpus, these particularities are taken into account in the essential steps of oral corpus construction, that is, segmentation, annotation, transcription and alignment.

3.1. CS speaker selection

We selected participants using a socio-linguistic online questionnaire named *Experience of Code-switching practice* (ECSP). The ECSP form includes questions about linguistic autobiography of the potentially recruted speakers, their bilingualism, the environment in which French and AA languages are practiced, language acquisition/learning, CS habits. . . The participants were selected and recorded in 2016-2018 and were (young) adult speakers aged between 23 to 39 years, with an equal number of men and women. We selected speakers who stated that they used to code-switch in their daily lives and that they use CS at least in two domains among studies, work, family, friends. All the speakers have lived a part of their life in Algeria and another part in France. They all have studied in university. At this date, we transcribed and aligned 13 speakers.

3.2. Recording protocol

The speakers were recorded in a soundproof room at LPP (Labortoire de Phonétique et Phonologie) of Sorbonne-Nouvelle University in Paris or in a very calm room. The corpus is intended to serve multiple CS research purposes including phonetic, prosodic and lexical studies as well as more automatic processing challenges such as language identification and language boundary detection. The main challenge of recording the CS conversations in a lab environment was to meet two antagonistic requirements: ensure the collection of informal speech with a high quantity of CS and at the same time guarantee a high acoustic quality of the signal.

To this aim, each recording session started with a preliminary unrecorded conversation with the speaker, to get her/him in a relaxed setting, practicing both languages in the same interaction. We started the recording with the first task of reading texts in both French and AA. This was followed by discussions designed to elicit spontaneous speech with CS production. Figure 1 shows the corpus organization in two parts: controlled speech with reading texts task and spontaneous speech triggered by questions.

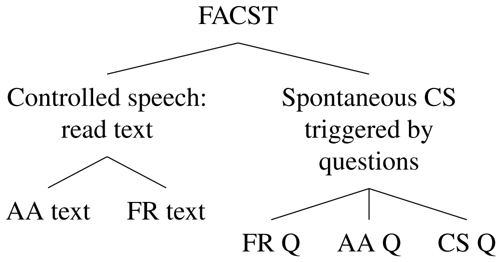

FACST
Controlled speech: read text
Spontaneous CS triggered by questions
AA text
FR text
FR Q
AA Q
CS Q

Figure 1: FACST data speech organization.

3.3. Reading tasks

The speakers were asked to perform oral readings of two texts, one in AA and another one in French, at three different speech rates (slow, normal, fast). For French, the text was an excerpt from "Le Petit Prince" (’The little Prince’, by A. de Saint-Exupéry). For Algerian Arabic, we used an excerpt from an Algerian movie scenario "Bab El-Oued City" (by M. Allouache) transcribed in Arabic letters. The controlled read speech recordings are summarized in Table 2. The first goal of the read-speech recordings was to obtain a controlled monolingual speech

Table 2: Read speech in French and AA. Number of words including repetitions and average reading times in seconds (slow-medium-fast)

Language	# words	Ave. reading times for 3 rates
FR	185	92s - 60s - 55s
AA	102	50s - 37s - 30s

corpus in AA and FR for the bilingual speakers before proceeding to the bilingual speech. Second, the recordings can serve to highlight potential pronunciation differences of consonants and vowels in each language separately. Third, studying the productions at the three speech rates provides data to investigate rate-related differences in the realization of consonants and vowels in each language. So, the controlled speech data helps to apply acoustic analysis and serves to beacon for phonetic observations in CS spontaneous speech.

3.4. Elicited CS speech

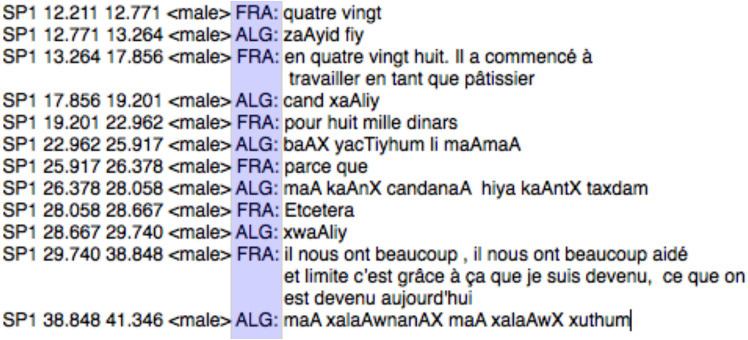
In this step, we recorded dual conversations between the linguist (who is a bilingual speaker of both languages) and the speaker. The CS conversations are triggered by questions. The principal questions were inspired by the feedback in the ECSP questionnaire of each speaker. Most of the subjects covered by the speakers were about describing and comparing life in both countries, studies in both countries and conversations about language and bilingualism practices. Other sub-themes were also addressed in the conversations following a free speech approach. So, the role of the linguist in this task is to ask the questions and let the speaker answer freely, making spontaneous use of CS. The recordings of this task lasted from 10 to 25 minutes for each speaker. To elicit CS, we asked four main types of questions in each speech sequence. These questions are summarized as follows:

1. AA questions about studies and work.
2. FR questions about life and studies in Algeria.
3. FR/CS: code-switched questions using French as the base language.
4. AA/CS: code-switched questions using AA as a base.

In these stimuli, the linguist tries to implicitly suggest the use of AA and French in a same conversation to obtain CS spontaneous speech.

3.5. Segmentation and annotation

First, manual segmentation of the recordings is based on language change, breath groups and speaker turns. There are two types of segments: language segments which correspond to one language, and breath group segments which correspond to a rhythmic group of an utterance or an oral sentence. The lengths of the segments are quite variable, ranging from very short segments (less than 2s) to longer ones, with an average segment length of 6s. Indeed, a switch can be limited to a very short word such as an article or a particle. The aim of this segmentation is to get boundaries for each language and label them "ALG" for AA segments and "FRA" for French segments. This type of segmentation and annotation helps thereafter when processing and manipulating the data. We used Transcriber program (Barras et al., 2001) to segment and annotate FACST data. Figure 2 shows an example of segmentation with the segment-level annotations.



```
SP1 12.211 12.771 <male> FRA: quatre vingt
SP1 12.771 13.264 <male> ALG: zaAyid fiy
SP1 13.264 17.856 <male> FRA: en quatre vingt huit. Il a commencé à
                               travailler en tant que pâtissier
SP1 17.856 19.201 <male> ALG: cand xaAliy
SP1 19.201 22.962 <male> FRA: pour huit mille dinars
SP1 22.962 25.917 <male> ALG: baAX yacTiyhum li maAmaA
SP1 25.917 26.378 <male> FRA: parce que
SP1 26.378 28.058 <male> ALG: maA kaAnX candanaA  hiya kaAntX taxdam
SP1 28.058 28.667 <male> FRA: Etcetera
SP1 28.667 29.740 <male> ALG: xwaAliy
SP1 29.740 38.848 <male> FRA: il nous ont beaucoup , il nous ont beaucoup aidé
                              et limite c'est grâce à ça que je suis devenu,  ce que on
                              est devenu aujourd'hui
SP1 38.848 41.346 <male> ALG: maA xalaAwnanAX maA xalaAwX xuthum
```

Figure 2: Example of segments annotation by speaker code (SPx), time-codes (columns 2-3), gender, language code (FRA/ALG) and transcription.

3.6. CS Transcription

The segments were manually transcribed using an orthographic transcription for French and a transliterated transcription for AA inspired by Buckwalter Arabic transcription (BKW) (Buckwalter, 2002) and modified according to AA specificities. The aim of the transliteration is to get a scripts with the same characters in both languages and to get a script which is written in the same direction (Arabic script is written from right to left). This transcription convention has been created in order to facilitate the use of the manual transcriptions, without special characters, for phonetic analyses while keeping the possibility to convert the transliterated characters to Arabic characters in future studies. Table 5 of Appendix illustrates the characters chosen for FACST and the corresponding symbols for each characters in Arabic letters, BKW convention and IPA symbols. For AA and FRA languages, the transcription also includes pauses, repetitions, hesitations, speech backchannels and various linguistic disfluencies. 13 of the speakers are transcribed at this date. Table 3 shows statistics about the number of segments and words counted in CS speech of the transcribed speakers.

In Arabic script, the articles and some particles are attached to the word: اَلبَاب (**اَل** + بَاب) "the door", لِلبَاب (**لِ** + اَل + بَاب) "for/to the door". For AA segments, we transcribed the articles and a large number of particles, placed initially at the beginning of the words, separately from the word. This separation is applied in order to count the number of words and to compare speech production at word level in CS FR-AA. So, in FACST transcription the utterance لِلبَاب is transcribed "li al baAb" "for/to the door".We used this to readily separate the languages in intra-sentential CS. Example: "liy" in AA, a mark placed at the end of a word refers to pronoun suffixes, conjugation morphemes, and number and gender marks.Thus, due to the morphological construction of AA (Souag, 2006), we did not apply separation for attached morphemes at the end of words in this corpus, examples: attached objects "jabt**hum**" *I brought **them** back.*

With the help of these transcriptions, we counted too the

Table 3: FACST spontaneous speech: number of segments and words counted for each language and for each speaker (13 speakers)

Spks#	FR seg	FR wrd	AA seg	AA wrd
SP1 M	28	112	36	108
SP2 F	126	583	170	734
SP3 F	219	1283	126	369
SP4 F	113	720	67	230
SP5 M	175	1619	93	348
SP7 M	180	1157	98	243
SP8 F	288	829	186	277
SP10 M	84	793	11	45
SP11 F	307	1746	236	948
SP12 F	334	1960	223	598
SP16 M	419	2075	277	1108
SP19 F	372	2122	248	710
SP20 F	335	1906	202	1172

highest occurrences of words in AA and FRA in order to know the words and the utterances which appears frequently in this pair of languages. Especially, getting a precise overview of AA insertion as an embedded language in FRA as a matrix language. The lists bellow present words and groups of words occurrences > 20 for AA and > 60 for FRA.

AA: waAHad, allah, bazzaAf, gaAc, waAluw, in, kaAmal, bacd, hnaAyaA, kiymaA, wiyn, kaAyan, laAzam, claY, hnaA, kunt, liy, mca, baAX, iyh, min, li, kaAn, lak, taAc, anaA, waAX, dzaAyar, kiy, alliy, laA, fiy, maA, wa, al, salaAm.

FRA: c'est-à-dire, un petit peu, il y a , deux, non, dans, très, par, beaucoup, quand, voilà, enfin, fait, plus, ils, tout, même, oui, tu, pour, donc, on, qui, parce que, mais, une, ça, en, j'ai, un, des, il, à, et, le, la, les, c'est, pas, que, je ...

A lot of AA words introduce CS, they start AA segment after FRA segment or end AA segment to switch to FRA. Mostly this words are particles:"fiy" in, "li" for/to, "taAc" for, "wiyn" where, "claY" on, "alliy" ... They connect the two languages and introduce a significant part of switches with a high number of occurrences as shown in table 4.

word	#occurrences	#occ introducing CS
fiy	208	112
wa	224	54
alliy	60	21
li	34	14
taAc	32	25

Table 4: AA word occurrences introducing CS > 14 occurrences

The following examples illustrate the dispositions of theses words in CS speech.

"FRA: internet
ALG: wa tXuwf al **fiy**
FRA: les forums"
"FRA: une licence
ALG: **fiy** al carbiyyaT
FRA: français et anglais"
"FRA: et du coup
ALG: **fiy**
FRA: la faculté"
"FRA: il avait des notes très très basses
ALG: **fiy**
FRA: français"
"FRA: le bouchon
ALG: **alliy** kaAyan **fiy**
FRA: le réservoir"

3.7. Data alignment

The data has been aligned using LIMSI speech recognizer in a forced alignment mode (Gauvain et al., 2003), assigning word and phone level time codes. The alignments made use of acoustic models from different ASR systems (Gauvain et al., 2003; Laurent et al., 2016) in parallel: a French system, and an Algerian Arabic (dialect) system. The alignment is composed by two separate parts. First alignment for French segments using pronunciation dictionary of standard french.The second alignment for ALG segments with pronunciation dictionary of AA created on the basis of MSA model and adapted on AA with specific phonemes used in this dialect as /p, v, g, ʒ/. Thereafter, both alignments are combined to the speech signal as shown in Figure 3.

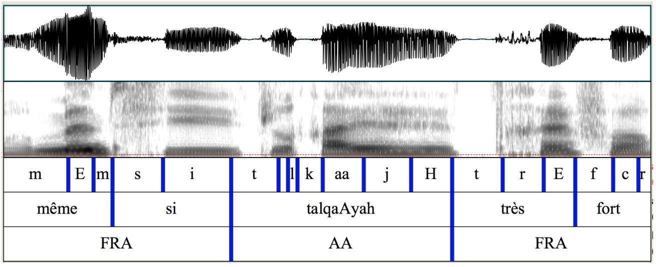


Figure 3: Example of FACST alignment on intra-sentential CS audio segment with spectrogram. Three levels of segmentation and transcription are applied. From top to bottom of tiers: phoneme segmentation and phonetic transcription, word segmentation and orthographic transcription, language segmentation and annotation.

4. Discussion and future work

In spite of a high CS production in the FR-AA bilingual community, recording speakers in laboratory requires a careful preparation of the communicative situation and of the setting of CS usage (Gumperz, 1982). The role of the linguist, in this experience is, to be an interactive speaker who can adapt himself to the daily CS usage of the participant in order to incite a CS conversation.

We found that frequent CS could be elicited by different questions, but the number of segments in FR and AA is not equal. FR has a higher number of segments and of produced words than Algerian for almost all speakers Table 3 shows clearly the quantity of AA and FRA distribution in CS spontaneous speech. So, we can say that in the FACST spontaneous speech corpus, FR can be considered as the dominant language and AA as dominated one. In the segmentation of spontaneous speech portions of FACST, some segments are very short (less than 1 second) because of the speed of language change. These segments generally correspond to particles and articles of both languages. One of the major difficulties of language annotation in CS FR-AA is the articles **al** /l/ in AA and **l'** /l/ in french at the switches. The two articles are pronounced identically but it is very difficult to identify the language of this word at the switch moment. E.G.: (fiy **l'**école fiy-AA, **l'**-FR, école-FR) or (fiy **al** école fiy-AA, **al**-AA, école-FR)

CS verbs can take a base in one language but are re-designed with the other language form. An illustration is the following CS example of AA-FR:

ypartaAj**iy** يپَرتَاجِي **/j**parta:ʒ**i:**/ "he shares"

This type of neologism is not easy to classify as French or as AA because it doesn't conventionally belong to none of the two languages. The root of the verb in bold ***ypartaAj****iy* is in French *partager* "to share". The prefix and suffix *y* - *iy* are in Arabic: the present form of the verb with the pronoun "he".

In summary, this paper describes the steps used to construct a CS corpus containing spontaneous speech and read bilingual speech. We present the methods employed to obtain CS in a dual conversation between a linguist and a speaker. We developed a methodology to annotate the CS corpus and we proposed transcription conventions for dialectal Algerian Arabic speech. An acoustico-phonetic study will be planned to analyze CS phenomena, in particular, phonemes at the moment of switches and the question of articles mentioned above.

5. Acknowledgements

This work 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 hand by the French Investissements d'Avenir - Labex EFL program (ANR-10-LABX-0083). The study was conducted at LPP-CNRS (UMR 7018) Paris-III Sorbonne Nouvelle University and LIMSI-CNRS Paris-Orsay University. Many thanks to the 20 participants recorded who contributed to realize the FACST corpus.

6. Bibliographical References

Al-Badrashiny, M., Eskander, R., Habash, N., and Rambow, O. (2014). Automatic transliteration of romanized dialectal arabic. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 30–38.

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

Auer, P. (2010). *Language and Space: An International Handbook of Linguistic Variation. Theories and Methods*, volume 1. Walter de Gruyter.

Barras, C., Geoffrois, E., Wu, Z., and Liberman, M. (2001). Transcriber: development and use of a tool for assisting speech corpora production. *Speech Communication*, 33(1):5–22.

Bies, A., Song, Z., Maamouri, M., Grimes, S., Lee, H., Wright, J., Strassel, S., Habash, N., Eskander, R., and Rambow, O. (2014). Transliteration of arabizi into arabic orthography: Developing a parallel annotated arabizi-arabic script sms/chat corpus. In *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*, pages 93–103.

Buckwalter, T. (2002). Arabic transliteration. *URL http://www. qamus. org/transliteration. htm.*

Bullock, B. E., (2012). *Phonetic reflexes of code-switching*, chapter 10, pages 163–181. Cambridge University Press, Cambridge.

Çetinoğlu, Ö., Schulz, S., and Vu, N. T. (2016). Challenges of computational processing of code-switching. *arXiv preprint arXiv:1610.02213.*

Cotterell, R., Renduchintala, A., Saphra, N., and Callison-Burch, C. (2014). An algerian arabic-french code-switched corpus. In *Workshop on Free/Open-Source Arabic Corpora and Corpora Processing Tools Workshop Programme*, page 34.

Gardner-Chloros, P. (2009). *Code-Switching*. Cambridge University Press, Cambridge.

Gauvain, J.-L., Lamel, L., Schwenk, H., Adda, G., Chen, L., and Lefevre, F. (2003). Conversational telephone speech recognition. In *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*, volume 1, pages I–I. IEEE.

Grosjean, F. (1995). A psycholinguistic approach to code-switching: The recognition of guest words by bilinguals. *One speaker, two languages: Cross-disciplinary perspectives on code-switching*, pages 259–275.

Gullberg, M., Indefrey, P., and Muysken, P., (2012). *Research techniques for the study of code-switching*, chapter 2, pages 21–39. Cambridge University Press, Cambridge.

Gumperz, J. J. (1982). *Discourse strategies*, volume 1. Cambridge University Press.

Kebeya, H. (2013). *Inter-and intra-sentential switching: are they really comparable?* Ph.D. thesis, Kenyatta University.

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

Lyu, D.-C. and Lyu, R.-Y. (2008). Language identification on code-switching utterances using multiple cues. In *Interspeech*, pages 711–714.

Lyu, D.-C., Tan, T.-P., Chng, E.-S., and Li, H. (2015). Mandarin–english code-switching speech corpus in south-east asia: Seame. *Language Resources and Evaluation*, 49(3):581–600.

Modipa, T. I., Davel, M. H., and De Wet, F. (2013). Implications of sepedi/english code switching for asr systems. *PRASA 2013, Johannesburg, South Africa*.

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

Piccinini, P. E. (2012). Cross-language activation and the phonetics of code-switching. *Speech Communication*, 52(11-12):930–942.

Saadane, H. and Habash, N. (2015). A conventional orthography for algerian arabic. In *ANLP Workshop 2015*, page 69.

Schultz, T. and Kirchhoff, K. (2006). *Multilingual speech processing*. Academic Press.

Sebba, M. (2012). On the notions of congruence and convergence in code-switching. In Almeida Jacqueline Toribio et al., editors, *The Cambridge handbook of linguistic code-switching*, chapter 3, pages 40–57. Cambridge University Press, Cambridge.

Souag, M. L. (2006). Explorations in the syntactic cartography of algerian arabic. Master's thesis, School of Oriental and African Studies (University of London).

Tabouret-Keller, A. and Page, R. B. L. (1970). L'enquête sociolinguistique à grande échelle: Un exemple: Sociolinguistic survey of multilingual communities. part i, british honduras survey. *La Linguistique*, 6(2):103–118.

Tossa, C.-Z. (1998). Phénomènes de contact de langues dans le parler bilingue fongbe-français. *Linx. Revue des linguistes de l'université Paris X Nanterre*, 1(38):197–220.

Ziamari, K. (2008). *Le code switching au Maroc: l'arabe marocain au contact du français*. Collection Espaces discursifs. L'Harmattan.

7. Appendix

Transliteration convention for AA consonants and vowels

	IPA	**FACST symbol**	Arabic symbol	BKW symbol	Examples	Transl
Plosives	p	**p**	پ		پلَاصَا plaSaA	place
	b	**b**	ب	b	بَاطل bATal	free
	t	**t**	ت	t	تَبَّع tabbac	follow
	t	**M**	ة	t	حيَاة HyaAM	life
	$t^{ʕ}$	**T**	ط	T	طبِيب Tbiyb	doctor
	d	**d**	د	d	دَبزَة dabzaM	punch
	$d^{ʕ}$	**D**	ض	D	ضَحكَة DaHkaM	smile
	k	**k**	ك	k	كتَاب ktaAb	book
	g	**g**	گ	g	گَاع gaAc	all
	q	**q**	ق	q	قَرعة qarcaM	bottle
	ʕ	**E**	أَ	>	أمَل Emal	hope
Affrc	dʒ	**j**	ج	j	جُوع juwc	hunger
	dz	**dz**	جز		جزَايَر dzaAyar	Algérie
Nasals	m	**m**	م	m	مَاضي maADiy	past
	n	**n**	ن	n	نُوم nuwm	slumber
Fricatives	f	**f**	ف	f	فُوق fuwq	on
	v	**v**	ڤ		ڤِيلَا viylaA	villa
	θ	**F**	ث	v	ثُوم Fuwm	garlic
	ð	**V**	ذ	*	هَذَا haVaA	this
	s	**s**	س	s	سفَر safar	travel
	$s^{ʕ}$	S	ص	S	صُور Suwr	wall
	z	**z**	ز	z	زِيت ziyt	oil
	ʃ	**X**	ش	$	شَاف XaAf	saw(you)
	x	**x**	خ	x	خرِيف xriyf	autumn
	ɣ,ʁ	**G**	غ	g	غرِيب Griyb	foreign
	ħ	**H**	ح	H	حِيط HiyT	wall
	ʕ	**c**	ع	E	عِين ciyn	eye
	h	**h**	ه	h	هُومَ huwma	they
Laterals	l	**l**	ل	l	لِيل liyl	night
	r	**r**	ر	r	رَاس raAs	head
Spirants	w	**w**	و	w	وَردَة wardaM	rose
	j	**y**	ي	y	يَد yad	hand
vowels	i	**i**	اِ	i	ضِيَّع Diyyac	lose
	u	**u**	اُ	u	جُملَة jumlaM	phrase
	a	**a**	اَ	a	فَرجَة farjaM	show
	a	**Y**	ى	Y	علَى claY	above
	iː	**iy**	اِي	iy	خشِين xXiyn	bold
	uː	**uw**	اُو	uw	غُول Guwl	monster
	aː	**aA**	اَا	aA	مَاكلَة MaAklaM	food

Table 5: Transliteration of AA consonants and vowels in FACST corpus with corresponding symbols in IPA, BKW, Arabic symbols and examples of words. FASCT symbols represent the phonological pronunciation of AA (Algiers region)