UNIVERSITÉ SORBONNE PARIS CITÉ

UNIVERSITÉ SORBONNE NOUVELLE - PARIS 3

École doctorale 268 - Langage et langues: description, théorisation, transmission

LABORATOIRE DE PHONÉTIQUE ET PHONOLOGIE

Thèse de doctorat en Sciences du Langague

DJEGDJIGA AMAZOUZ

Linguistic and phonetic investigations of French-Algerian Arabic code-switching:

Large corpus studies using automatic speech processing

Sous la direction de Martine ADDA-DECKER et Lori LAMEL

Soutenue le /03/12/2019

Comité de jury:

RUDOLPH SOCK	PR. (LILPA/IPS, Université de Strasbourg, France)	Rapporteur
KAMEL SMAILI	PR. (LORIA, Université de Nancy, France)	Rapporteur
BARBARA E. BULLOCK	PR. (University of Texas at Austin, United States)	Examinatrice
RACHID RIDOUANE	DR (LPP,UMR7018,CNRS Sorbonne Nouvelle, France)	Examinateur
MARTINE ADDA-DECKER	DR (LPP,UMR7018,CNRS Sorbonne Nouvelle, France)	Dir de thèse
LORI LAMEL	DR (LIMSI-CNRS Orsay, France)	Co-encadrante

This PhD thesis was prepared in
Laboratoire de Phonétique et Phonologie
CNRS, UMR7018, Sorbonne Nouvelle University
and LIMSI-CNRS, Paris-Saclay University
Under the project ANR-SALSA

Abstract

This thesis investigated French-Algerian Arabic code-switching using automatic speech processing tools. A corpus of 7h30 of code-switched speech from 20 French-Algerian Arabic speakers (5h of spontaneous speech and 2h30 of read speech) has been designed, recorded and annotated. One of the first challenges tackled consisted of developing data processing methods such as language segmentation, code-switching utterance segmentation as well as transcription in French and Algerian Arabic dialect. Automatic speech alignment methods were adapted to process the code-switched data by combining two monolingual alignment systems thus producing time-stamped orthographic and phonemic transcriptions in both languages. An experiment was conducted to automatically detect language switches, however this remains a challenge especially for small speech stretches. A second aspect of this thesis' research studied the influence of the phonological system of language *a* on the second language *b* in code-switched speech, in this case the phonetic productions of French and Algerian Arabic. The annotated corpus was used to carry out phonetic studies on vowel and consonant variation using an automatic ABX-like phone discrimination paradigm. With this paradigm, our results on variation in code-switched speech vowel productions are in line with a priori hypotheses: considering the peripheral /i,a,u/ vowels, higher variant rates are measured in Algerian Arabic (40%) than in French (27%). A comparison with native French control speakers suggests that the bilingual speakers have more conservative vowel productions than natives (34%), at least in code-switched speech. Three types of consonant variation were also explored: gemination, emphatization and voicing alternation. Overall, consonants show similar trends to vowels: 42% variant rates for Algerian Arabic, 30% for French in code-switched speech, compared with 38% for French natives. Future studies using this innovative corpus will contribute to disentangle the complex interplay between phonetic variation and phonological systems in bilingual code-switching speakers.

Résumé

Cette thèse traite du code-switching français-arabe algérien à l'aide d'outils de traitement automatique de la parole. Un corpus de 7h30 de parole de 20 locuteurs bilingues (5h de parole spontanée et 2h30 de parole lue) a été conçu, enregistré et annoté. L'un des premiers défis abordés a consisté à développer des méthodes de traitement des données telles que la segmentation en langues, la transcription du français et de l'arabe algérien. Les méthodes d'alignement automatique de la parole ont été adaptées pour traiter les données du code-switching en combinant deux systèmes d'alignement monolingues, produisant ainsi des transcriptions orthographiques et phonémiques avec des localisations temporelles dans les deux langues. Une expérience a été menée pour détecter automatiquement les changements de langue, mais cela reste un défi, en particulier pour les durées monolingues très courtes. Le second aspect de la recherche de cette thèse porte sur l'influence du système phonologique de la langue *a* sur la deuxième langue *b* dans la parole du code-switching, en l'occurrence les productions phonétiques de l'arabe et du français. Le corpus annoté a été utilisé pour effectuer des études phonétiques sur la variation des voyelles et des consonnes en utilisant un paradigme de discrimination automatique de type ABX. Avec ce paradigme, nos résultats sur la variation de la production correspondent aux hypothèses a priori: considérant les voyelles périphériques /i,a,u/, des taux de variantes plus élevés sont mesurés en arabe algérien (40%) qu'en français (27%). Une comparaison avec des locuteurs de langue maternelle française suggère que les locuteurs bilingues ont des productions de voyelles plus conservatrices que les locuteurs natifs (34%), du moins dans le code-switching. Trois études sur la variation des consonnes ont également été menées: la gemination, l'emphatisation. Globalement, les consonnes présentent des tendances similaires à celles des voyelles : 42% de taux de variantes pour l'arabe algérien et 30% pour le français en code-switching, contre 38% pour les natifs français. De futures études utilisant ce corpus novateur pourront contribuer à démêler l'interaction complexe entre la variation phonétique et les systèmes phonologiques chez les bilingues dans le code-switching.

Remerciements

Il m'est très difficile de remercier toutes les personnes qui ont contribué à l'aboutissement de cette thèse.

Je voudrais tout d'abord adresser de grands remerciements à ma directrice de thèse Martine ADDA-DECKER, pour toute son aide, son encadrement et le savoir qu'elle m'a transmis, je voudrais également adresser de grands remerciements à ma co-encadrante de thèse Lori LAMEL qui grâce à son aide, a permis à ce travail de thèse de voir le jour. Je suis enchantée d'avoir travaillé en leur compagnie, car outre leurs disponibiliés durant ces années de thèse et l'appui scientifique dont j'ai bénéficié, elles ont toujours été là pour me soutenir, me conseiller et m'encourager au cours de l'élaboration de cette thèse. Je les remercie également de m'avoir transmis leur passion pour le traitement automatique de la parole.

J'adresse toute ma gratitude au projet ANR SALSA, au LPP-CNRS de l'université de Paris III et au LIMSI-CNRS de l'université Paris-Saclay pour le financement de recherche qu'ils m'ont attribué, et qui m'a permis de construire les données de thèse, de participer aux manifestations scientifiques et d'effectuer ma recherche dans de bonnes conditions.

Je voudrais remercier le professeur Rudolph Sock de LILPA/IPS de l'université de Strasbourg et le professeur Kamel SMAILI du LORIA de l'université de Nancy, d'avoir accepté d'être rapporteurs de ma thèse et pour le temps consacré à ce travail. Je remercie aussi la professeure Barbara BULLOCK de l'université d'Austin et le directeur de recherche Rachid RIDOUANE pour avoir accepté de participer à mon jury de thèse.

J'aimerais ensuite remercier les membres du LPP, du LIMSI et de Vocapia Research, doctorants et chercheurs, pour leur accueil chaleureux, leur aide précieuse, leur bons conseils et surtout pour leurs remarques très constructives à l'égard de mes travaux de thèse, en particulier Nicolas AUDIBERT, Annie RIALLAND, Jaqueline VAISSIÈRE, Pierre HALLÉ, Rachid RIDOUANE du LPP, Jean-Luc GAUVAIN et Claude BARRAS du LIMSI et Abdelkhalek MESSAOUDI de Vocapia Reaserch.

Mes derniers remerciements vont à Yanis MOHDEB, mon mari, pour son soutien moral quotidien, ses encouragements, mais aussi pour son aide dans la réalisation de ce travail grâce à ses relectures et son travail de traduction. Pour finir, j'aimerais remercier mes parents et toute ma famille et ma belle-famille pour leur grand soutien moral.

Contents

List of Figures

List of Tables

List of abbreviations

AA Algerian Arabic. xxiv, 22, 30, 31, 33, 60, 70, 106, 115

ASR Automatic Speech Recognition. xxiv, 21, 22, 23, 69, 99

BKW Buckwalter Arabic transliteration. xxiv, 61, 62

CNRS Centre National de la Recherche Scientifique, France. xxiv

CS Code-switching. xxiv, 10, 30, 49, 73, 77, 98, 106

ECSP Experience of Code-switching practice. xxiv, 51, 53, 56

FA automatic Forced Alignment. xxiv, 23, 24, 69, 93, 94, 150

FACST The French-Algerian Code-Switching Triggered corpus. xvii, xxiii, xxiv, 36, 49, 50, 52, 53, 54, 57, 62, 65, 67, 68, 70, 74, 77, 79, 80, 82, 88, 89, 98, 100, 111, 115, 123, 150, 155, 165, 193

FR French Language. xxiv, 22, 31, 106, 115

HMM Hidden Markov Model. xxiv, 163

IPA International Phonetic Alphabet. xv, xxiv, 35, 40, 61, 62

LID Language Identification. xxiv, 99

LIMSI Laboratoire d'informatique pour la mécanique et les sciences de l'ingénieur-CNRS, Orsay, France. xxiv, 22, 69, 71

LM Language Model. xxiv

LPP Labotaroire de Phonétique et Phonologie-Sorbonne Nouvelle, Paris, France. xxiv, 52

MSA Modern Standard Arabic. xxiv, 18, 22, 30, 31, 33, 39, 61, 67, 69, 70

NCCFr Nijmegen Corpus of Casual French. xxiii, xxiv, 111, 114, 123, 193, 194

NLP Natural Language Processing. xxiv, 83

OOV Out of Vocabulary. xxiv

POS Part-Of-Speech. xxiv, 84, 88

SAMPA Speech Assessment Methods Phonetic Alphabet. xvi, xxiii, xxiv, 71, 189

VOT Voice Onset Time. xxiv, 19, 21, 39

Introduction

This thesis is a linguistic and phonetic investigation of the French-Algerian Arabic code-switching in spontaneous speech. This research is preceded by several years of naive observation of often heard and read French-Arabic bilingual speech productions in Arabic countries and in France. These observations prompted me to ask questions about the ways this mix of languages is realized.

Also, after annotating and processing a large speech corpus in dialectal Arabic of Maghrebian countries in local medias and entertainment TV shows (under the Orelo project *Origine des REdacteurs et des LOcuteurs*), I noticed that code-switching is frequently used by the invited speakers as well as by the TV show presenters and that code-switching interactions are practiced by a considerable number of speakers. So, I started to have a big interest in the code-switching phenomena with the pair of languages French-Arabic.

It's known that dialectal Arabic in Maghreb countries, especially in Morocco Algeria and Tunisia is the most used and most spread language, and that code-switching with local dialects and French language is very widespread. The historical context of theses countries and the Maghrebian diaspora in France played an important role in these bilingual practices. Nevertheless, the comparison of the code-switching practices in these communities notes a modest literature despite the commonalities of situations and the interesting dialectal differences. So, my first interest on the code-switching studies was a comparison of the different code-switching practises in these three linguistic communities.

The focus on Algerian-Arabic and French code-switching is a result of a quantitative analysis on code-switching productions of the Orelo corpus. The study is based on a comparison of number of code-switching utterances across three pairs of languages: Algerian Arabic-French, Moroccan Arabic-French and Tunisian Arabic-French code-switching speech. The study confirmed that the code-switching is produced frequently by a large number of speakers with each Maghrebian Arabic. However, the implication degree of Arabic dialects and French to produce the code-switching varies in the thee. The Moroccan and Tunisian subsets include 35 and 39 language switches per hour and the Algerian Arabic-French code-switching records 200 switches per hour which is very high. Thus, in my study of code-switching that fits into large scale data of speech, I focused in this thesis on Algerian-Arabic French code-switching which is characterized by potentially dense production.

This research project attempted to provide empirical information concerning the Algerian-Arabic code-switching speech collection and data processing and to study of the phonetic dimension of code-switching speech production. This work aims to give useful information that could benefit many research fields such as sociolinguistics, phonetics, phonology and speech technologies. FACST code-switching corpus, data collection and data process methodology is the core of this research. This thesis describes all the steps from the code-switching spontaneous speech data collecting while focusing on the characteristics of this speech type which remains nowadays a still growing research field.

The phonetic and phonological dimensions of code-switching speech a relatively recent domain. Especially, the influence of language *a* on language *b* in production. So, one of the purposes of this thesis is to study segmental variation in code-switching production.

Code-switching is also a real challenge for automatic speech processing, particularly for automatic speech recognition and language identification, in fact, things get more complex when we deal with lower sourced languages like Arabic dialects. In this thesis, we try to provide more linguistic and phonetical information about Algerian Arabic and how it combines with French in code-switching productions.

List of abbreviations

In the first part of this thesis we will make a theoretical review of bilingual speech and code-switching phenomena, this review will describe the code-switching productions and give information about what studies have been done in code-switching and automatic data processing. This theoretical review of code-switching aims also to highlight the different phenomena that are close to code-switching such as borrowing.

The second part of this work is dedicated to French Algerian Arabic code-switching speech, a detailed description is developed on how these two languages are in contact and how they influence each other to produce one unique speech. This part deals also with the characteristics of this pair languages in contact, such as words association in the code-switching sentences, borrowing and bilingual words. The chapter is concluded with a description of the two systems and gives a description of phonemic inventories of Algerian Arabic and French, focusing on Algerian Arabic which is very low sourced in literature.

The third chapter deals with the different steps necessary to build oral spontaneous code-switching corpus, starting with the speakers selection and records protocols, methods to elicit code-switching during the records are proposed. The chapter describes also the transcription and annotation methods applied to this code-switching speech data. Finally, methods of forced alignment for this bilingual speech is proposed with the use of the automatic speech recognition tools.

The fourth chapter gives a description of the obtained CS data, the chapter is a quantitative review of the data speech words, sentences, language segments, utterances and code-switching utterances, language distribution in the speech and all other code-switching speech components. This review provides first linguistic and phonetic information of code-switching data. Besides, this chapter will also give information about language speech rate that will allow us to evaluate code-switching frequency.

The fifth and last chapter deals with segmental variation in code-switching speech in comparison with monolingual French speech produced by monolingual speakers. Segmental variation in this chapter is studied in vowels and consonants, thus this chapter essentially lays on the influence of the phonological system of language *a* on language *b*.

During this years of work we tried to answer some questions but many other questions raised all along the work, many of them are pertinent questions for future studies. So the conclusion of this work will focus on these new questions and perspectives.

Part I

Theoretical review

Chapter 1

Bilingual speech and code-switching

Chapter contents

THIS introductory chapter deals with the code-switching phenomenon and its related notions: bilingual speakers, bilingual speech and the difference between borrowing and the code-switching. In this literature review, we give various definitions of code-switching through a historical review and we focus on the studies that were conducted on code-switching. Therefore, we addressed the following questions and we try to answer them: what are the fields interested in the code-switching? What are the methods used to study and understand this phenomenon? And what are the material and the data used to process the code-switching speech? The last part of the chapter focuses on the different phonetic and linguistic studies about code-switching, and the recent studies on automatic CS processing. We conclude this chapter by presenting the different existing code-switching corpora.

1.1 Bilinguals, two languages, one speech

Code-switching (CS) is closely related to bilingualism. It is therefore important to emphasize the notions of bilingualism and bilingual practices as well as bilingual speakers. This section is therefore designed to define and distinguish these notions as well as their role in the production of CS.

1.1.1 Bilingual speakers

A bilingual (multilingual) speaker is a person who has the ability to speak more than one language. In simultaneous bilingualism, learning and practicing two languages is done within a parallel frame time. Sequential bilingualism is learning a language L*b* after acquiring a first language L*a*. The language proficiency in sequential bilingualism is a learning continuum which starts with the acquisition of a minimal proficiency and finishes with a maximal proficiency acquisition (Macnamara, 1967). Hence, we can categorize bilinguals according to their proficiency level in each language (From beginner to advanced bilingual). Language proficiency groups different skills such as grammar, syntax, vocabulary and pronunciation. Grosjean (2008) indicated that bilinguals are not double separated monolinguals in one person. So the linguistic and phonological systems of both languages coexist in bilinguals, thinking in L*a* during a conversation in L*b* and CS are the best examples of this linguistic border crossing.

1.1.2 Bilingual and monolingual speech

Monolingual speech is produced by a monolingual speaker or by a bilingual speaker in a monolingual mode which consists in focusing on only one language in the speech production (Grosjean, 2001). Example: French merchant speaking to foreign tourists in English. In the bilingual speech, both languages are triggered, and bilinguals use elements of one language when speaking the other. According to Scotton and Ury (1977); Scotton (2006); Grosjean (2008), in bilingual speech, one language serves as the basis of speech and the other intervenes like a "guest" language. Depending on conversational settings, the insertion of the guest language varies from one word, "borrowing", a utterance or a sentence "code-switching".

Works on the conversational difference between bilingual speech and monolingual speech concluded that CS speech functions have more possibilities of enhancements and of being more numerous compared to monolingual speech, word repetitions in both languages and reported speech in source language when speaking another language (Gardner-Chloros et al., 2000).

1.2 Code-switching, a language contact phenomenon

Authors	Definitions
(Hymes, 1962)	A common term for alternative use of two or more languages varieties of a language or even speech styles.
(Gumperz, 1982)	A juxtaposition within the same speech exchange of passages of speech belonging to two different grammatical systems or subsystems.
(Hoffmann, 1991)	The alternate use of two languages or linguistic varieties within the same utterance or during the same conversation.
(Myers-Scotton, 1993)	The use of two or more languages in the same conversation, usually within the same conversational turn, or even within the same sentence of that turn.
(Gross, 2006)	A complex, skilled linguistic strategy used by bilingual speakers to convey important social meanings above and beyond the referential content of an utterance.

Table 1.1: Definitions of code-switching

1.2.1 Historical review and definitions

Code-switching has been known and studied since the early 20th century. It coincides with high language contact and migratory and economic changes as well as the media.

Before the 19th century studies on languages focused on grammar, lexicography, syntax, semantics and globally on how to master language skills to write correctly. Thus, language switch in production was judged as a lack of language skills in L*b*. CS notion was introduced as a consequence of studies on language mix, and CS studies started with children language learning. In France, Ronjat (1913) recorded the bilingual production of children and used this pioneer data to observe bilingualism evolution in acquisition.

Thus, it is during the 50' that we notice the first greatest interests in bilingualism and code-switching. It was first in psycholinguistic studies that tempted to describe the bilingual production as a mental process (Weinreich, 1953; Uriel, 1953). Thereafter, many grammatical and syntactical studies focused on the CS structures within the sentence. Sankoff and Poplack (1981); Poplack (1980) worked on the CS syntactical categorization and classified the language switch within a sentence "intra-sentential code-switching", language changes at sentence boundaries called "inter-sentential code-switching", the "tag-switching" is the L*b* items in the language base. These classifications are commonly used in recent studies Kebeya (2013); Muysken (2000). The question of the grammaticality of CS utterances has also been the subject of study of linguists. Pfaff (1979) and Woolford (1983) described the switching items combinations and researched for rules in the possible grammaticality alternations between the languages in switch.

With the frequent use of the code-switching linguistic studies on the conversational aspects of the CS appear with Gumperz (1977) and Blom (1972). Researches interested in the conversational strategies involving language switch in order to explain cultural ideas and to conduct conversations to fit a social setting (Auer, 1984, 1995, 1996). Terms like "dialectal CS" is related to the communication in standard languages and their dialectal forms, "situational CS" is related to social settings or the speakers' situation (work, family) , "Conversational CS" refers to the speakers' conversation (colleague, manager, wife, son ...) (Gumperz, 1982).

Scotton and Ury (1977) add an important point to her contribution to CS studies. Indeed, she indicated that society plays an important role in the bilingual speaker language choices. She conducted several works that propose CS model frames of the languages insertions (Jake et al., 2002; Myers-Scotton, 2001) and the semantic and grammatical difficulties that speakers may encounter when switching from a language to another at syntactical and morphological levels (Scotton, 2006; Myers-Scotton, 1990).

CS studies proposed several definitions of the former, hence, the word CS can describe several forms of language contact (code-switching, language mix, borrowings, code mixed speech) Simonin and Wharton (2013). According to CS studies, the bilingual phenomenon definitions and appellations vary, "mixed structure" in sociolinguistic studies (Canut and Caubet, 2001), "code mixing" (Bullock, 2009), "code switching" (Lüdi et al., 1986; Lüdi, 1998). (Muysken, 2000) used the term of CS for the language sequence alternation and used the code-mixing for other L*b* lexical insertions like non integrated borrowings.

CS is characterized by individual choices and individual forms as a placement of adverbs in the beginning or at the end of sentences, the inclusion or deletion of articles/particles at the switch moment and the construction of CS sentences. In intra-sentential CS, speaker may produce an ungrammatical sentence due to the non-correlation of the two grammar codes but the sentences are semantically correct (Tossa, 1998).

1.2.2 Recent studies about code-switching

Many research fields showed their interest in CS, among them, studies on interactional linguistics that studies CS organization focusing on speakers speech turn, CS strategies and the link between CS usage within conversations conducted by bilinguals (Auer, 2010, 2013).

CS was also a subject matter in studies on language acquisition and learning, their main problematic are the necessity to introduce the first language in the practicing and learning process of a second language. The main aim of this approach is to facilitate the acquisition and learning of a second language through translation, explanations, definitions, etc. (Aref and Aref, 2015; Stoltz, 2011; Ezeizabarrena and Aeby, 2010; Moore, 2002).

Syntax and morphology are the research fields that studied the most CS, precisely intra-sentential CS construction. Among the studies that were conducted, syntactical constraints on Spanish-English CS (Poplack, 2001). A study on syntactical variation of different language pairs between Chinese-English CS (Wang and Liu, 2013). Boumans (1998); Ziamari

(2008) studied the syntax of CS Moroccan Arabic-French and the limits of the Matrix language frame (Myers-Scotton, 2001) and described the bilingual words construction with this pair of language.

With the development of the psychologies and the cognitive sciences, some researchers tried to understand CS as a cognitive mechanism before and during the speech production and the CS comprehension by following experimental studies (Basnight-Brown and Altarriba, 2007; Kootstra et al., 2015; Muysken, 2000).

CS is a sociolinguistic phenomenon and a large set of studies has been focused on the role of the social context to elicit CS speech practices. Multiple studies were based on sociolinguistic filed survey in order to surround the social status and the linguistic group influence on the CS production (Blom and Gumperz, 2000; Ziamari, 2008; Beattie et al., 1982; Al-Qaysi and Al-Emran, 2017).

More recently, CS speech has also raised interest in computational linguistics, automatic speech processing research and phonetics. A review of the past studies in these fields and CS questions investigated are in Sections 1.4 1.5.

1.2.3 Code-switching and borrowing

The interest about the borrowing notion started with bilingual speech studies and is inherent to CS (Gardner-Chloros, 2009a). Borrowing is defined like L*b* "transfer of words" in L*a* speech (Weinreich, 1953). Matras (2009) distinguished between two borrowing types: the lexical borrowing and the grammatical borrowing. The first type includes verbs, lexical adverbs, adjectives and nouns and their insertion is adapted to the syntactical functions of the base language and they are subject to syntactical modification like verb negation. The grammatical borrowing constitutes all grammatical units insertion in the basic language: particles, articles, prepositions, discourse markers, connectors... In contrast to lexical borrowing, the grammatical ones contribute in organizing the sentence and introducing switches in the basic language, they usually introduce the switch but without necessarily imposing their syntactic language scheme, example: the FR word "malgré" *despite* requires a noun in FR but both noun and verb in AA can follow this grammatical word, example FR-AA:

"malgrè نَقرَا بَزَّاف" *Despite I work a lot.*

Matras also distinguished between the borrowing and the word insertion in CS speech. Word insertions are part of CS speech, borrowing can overpass the linguistic frame, which leads to the integration of the borrowing at linguistic and phonological levels.

The borderline between borrowing and CS is considered like a continuum which is difficult to draw (Gardner-Chloros, 1991; Matras, 2009), because if we consider a starting point which is language *b* and an integration procedure that would end with language *a*, we would find in the first step of process CS, then we would find non-integrated borrowings; indeed, they do not sound natural in the target language, so the non-integrated borrowings are schematically far from the target language. Integrated borrowings are the closest to target language, they adopted its form, example: طَابلَة *table*. CS is in a way, the dynamic beginning of the word integration continuum as shown in the scheme 1.1.

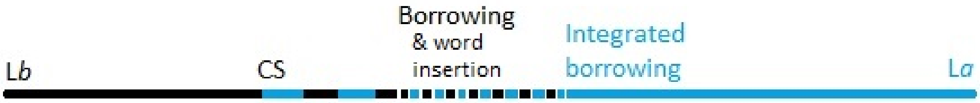


Figure 1.1: Borrowings and code-switching continuum

1.3 Code-switching in conversation

This section details the individual motivations and the collective factors that elicit the CS. The aim of this section is to address the construction of the CS speech and reproduce it in a corpus.

1.3.1 The motivations of code-switching practices

CS is a result of bilingualism and language contact and it emphasizes a communicative skill of two languages. Sociolinguistic factors lead to switching between two languages and the motivations of CS depends on the individual, collective and situational factors. (Gumperz, 1982; Auer, 2013; Gardner-Chloros, 2009b).

1.3.1.1 Individual factors

Individual factors depends first to the speaker's will to code-switch and his language attitude. Indeed, bilingual speakers, in addition to their languages proficiency, accept to use two languages in a same utterance. Also, the individual motivations of CS can be related to the mental accessibility of the language items during the speech. Thus, bilinguals switch when they can find quickly appropriate words or expressions in one language (Grosjean, 1982).

1.3.1.2 Code-switching as a collective "social trend"

Collective factors represent the social group practices and preferences interlocutor in language communication. The role of the participant and the relationship guide the language choice and the speech mode (bilingual or monolingual). Indeed, according to Grosjean (1982), CS practices are encouraged or rejected by bilingual interlocutors, by their attitude toward the speakers and the spoken language, emotion, and the group identity trends. Furthermore, in a social group, speech habits between speakers groups may lead to CS productions. Sociolinguistic studies indicated that CS production varies with social categories and also depends on the common identity of the speakers) (Auer, 2005; Post, 2015). Examples: adolescents use English to code-switch with their La in social networks, code-switching in migrant communities.

1.3.1.3 Situational code-switching

Individual and collective factors influence and guide the CS particles, but also it can be motivated by the conversation situation and the speaker's status. In bilingual or multilingual conversation, some languages are more appropriate than others depending on the topics of conversation and communication register like religion, youth conversations, sport, traditions, work (Ritchie and Bhatia, 2004). It should also be noted that the bilingual mode and the monolingual mode can be chosen in order to establish a type of relationship between two speakers. Example: to establish a friendly relationship a bilingual speaker in a foreign country uses CS with his/her first language and the foreign country language to communicate.

1.3.2 How code-switching is triggered in speech?

Many studies have focused on how CS appears within speech. As shown in the previous section, CS motivation is related to sociolinguistic and conversational settings and to the use of language switches. However, these sociolinguistical settings may be influenced by linguistic elements that facilitate the triggering of CS productions. So, what are the linguistic elements that can introduce and trigger the language switch in speech?

Works on CS triggering go back to the research studies of Clyne (1980, 1987), who is interested in the impact of words that introduce language switch *trigger-words*, in production and in how these words may cause code-switched syntactical forms and influence the language choice. According to Clyne, trigger words are classified in three types. The first type is words that are used commonly by CS speakers and which may be characterized by an ambiguous or shared language affiliation, thus, potentially inviting both languages to continue. So, integrated borrowings of one of the languages in another can contribute to CS and then can be trigger words.

Clyne's study involving 600 German-English bilingual informants and 200 Dutch-English bilingual informants in Australia, showed that 30% of the informants produced CS after *trigger words* in the conversation by the same speaker or the interlocutor. Recent research (Broersma and De Bot, 2006; Mirjam Broersma, 2009; Broersma, 2011) on Moroccan Arabic–Dutch CS, shows that cognate words tend to be followed by the language switch more than non cognates ones.

Furthermore, triggering CS may go beyond the word level. Used shared CS syntactical structures may also facilitate CS productions (Clyne, 1987; Clyne et al., 2003; Kees de Bot and Isurin, 2009) and shared syntax may contribute to the simultaneous activation of both languages to CS (Hartsuiker et al., 2004).

It should be noted that language cognate items vary across languages pairs. Close languages share more cognates than distant languages. For example: the Modern Standard Arabic (MSA)-Arabic dialect pair shares an important part of cognates and French-English pair shares more cognates than French-Algerian Arabic pair or French-Moroccan Arabic. However, Arabic-French witnesses a widespread practice of CS (Ziamari, 2008; Amazouz et al., 2016). So, the triggering of CS with cognate words and syntactical structures mainly concerns close languages.

Besides, regardless of the languages closeness, CS is also produced through interactions and organizes itself through the conversational, situational, and sociolinguistic context. So, a macro-trigger of CS is related to the sociolinguistic conditions (Zentella, 1997, Chapter 5-6). In addition, the motivation to code-switch contributes to trigger the CS even for the most distant languages (Myers-Scotton, 1995). Myers-Scotton (1993) has also shown that the interlocutor's CS behaviour has a direct impact on the speaker, by inviting him to use CS in turn.

To summarize, triggering CS is influenced by conversational context setting. CS also depends on the will of the interlocutors to use it. CS can be triggered and enhanced with *trigger-words* and *trigger-syntax* if the languages pair allows it syntactically. So, the interviewer can opt for a spontaneous CS speech in order to receive code-switched utterances.

1.4 Code-switching and phonetics

When speakers switch from one language to another, all language elements are included in the switch including the phonological systems. Although there are more and more studies on CS, phonetic studies remain limited to this date. We intersect in this part to the addressed phonetic questions and the related research done on the phonetic of CS.

1.4.1 Two phonological systems in one speech

Phonological questions that accompanied CS studies, focused on the presence of two phonological systems in one speech and their impact on production. Hasselmo (1961); Caramazza et al. (1974); Clyne (1987) states that phonological aspects of a language are often preserved when CS occurs. However, this question partially linked to pronunciation skills of both languages, in fact, advanced bilingual speakers are aware of the existence of two phonological systems. Yet, bilingual speakers who do not have a complete proficiency in pronunciation may keep their foreign pronunciation in their second language productions. On the other side, the matter of the influence of a language on another remains complex to study. CS production varies from a pair of languages to another. Indeed, phonological systems of languages can be either close to each other, i.e. they have many phonemes in common and their pronunciation is similar (Spanish and Italian), so the influence may be minor in production. Also, the phonological systems of languages may be very different from each other (French and Arabic), thus, CS productions may be subject to influences and variations in their productions.

In language contact situations and CS speech, two kinds of phonetic production words exist; production that implies a phonological adaptation of L*b* words and borrowings, and production that keeps in target language their initial phonological characteristics. (Grosjean, 2008) gave examples of borrowings that keep their initial pronunciation in target language: "je vais checker" [ʒə vɛ tʃeke] *I will check.* French-English language pair has also adapted borrowings such as: [maʁkœting] *marketing*, [smaxtfɔn] *smart-phone*, etc.

1.4.2 Past acoustic works

Previous studies in phonetics and prosody that focused on the influence of one language on the production of the other one during switching often present diverging results across studies. Indeed, studies assumed that phonetic impact on languages is absent at the switch moment, and others has claimed that the language switch may influence the phonetic production in both languages. Grosjean (2008) in controlled sentences switches in French-English, shown that the switch is not only at lexical and grammatical levels, it also includes the phonetic switch of the words. So, the phonetic impact of the base language on the Embedded language language in CS production would be avoided by the speakers. In the same way, the researcher has performed experiments on /p, t, k/ Voice Onset Time (VOT) values, a comparison between French and English in CS and monolingual context. The results show that VOT values do not change for both languages with the change of context. So, CS does not appear to affect the production in French-English pair. Bullock (2009) study suggests that speakers tend to stick to the spoken languages' standards (e.g. a stop consonant is produced with a typical English burst when speaking English and as a typical Spanish stop during switching to Spanish) with only minor mutual interaction.

Most CS studies involving measurements of the acoustic speech signal are carried out on controlled CS corpora (read sentences including CS, triggered CS speech using the switch at the bip protocol...). Studies focusing on morphological or higher levels of CS speech often rely on transcripts of less controlled, spontaneous CS speech. However, the acoustic signal is mainly used as the compulsory raw material to get transcripts, which then become the true object of study. These works are based on a different corpora (controlled CS speech, speech laboratory, spontaneous CS) and obtained different results.

1.4.3 Phonetic speech variation in CS

Phonetic variation in speech is studying the different phoneme changes in production compared to their canonical pronunciation. Thus, in many languages we observe that within a language itself, variations are observed (Ohala, 1993; Foulkes et al., 2010). A word can have different pronunciation variants: sound change like devoicing, assimilation, lenition, fortition, phone propensity may occur. Segmental reduction (deletion or temporal reduction of phones) like the schwa elision, deletion and the of consonants, for example: /r, ʁ/ in English and French.

Variation factors in speech can be linked with individual speaker factors but also to speech style, for example: media speech, casual and journalistic speech (Adda-Decker and Lamel, 1999; Meunier, 2014). Speech phonetic variation is often studied in large scale corpora studies, thus, speech automatic processing is a helpful method to perform these variation studies. Among those tools; automatic speech alignment with variants (Adda-Decker, 2006; Lamel et al., 2009). It allows to the word pronunciation to be recognized with different pronunciation variants at the phone level (See Sections 5.1.1, 3.7).

Sociophonetics studies in CS studied phonetic speech variations. Among the addressed questions, the phonetic co-influence of the languages in bilingual performance resulting from the languages simultaneous activation during the speech. In comparison between early English-Spanish bilinguals and two groups English-Spanish speakers with inverted L1 L2 status, (Bullock, 2009) showed that the phonetic influence of L1 on L2 in CS is observed in the VOT duration of voiceless stops. The study concluded that the influence can be bi-directional in both L1 and L2 and the variation is observed in each L1/L2 group.

Also, a previous work on VOT variation of stop consonants in English-Spanish CS reveals that bilinguals produce a lower speech rate around the language switches and that phonological modifications of VOT values (Deuchar et al., 2014). Other similar studies reported the reduction of VOT values in Spanish stop consonants in Spanish-English CS productions (Balukas and Koops, 2015; Piccinini, 2016; Botero et al., 2004). Also, Piccinini and Garellek (2014) indicated that English-Spanish bilinguals produce different prosodic contour depending on monolingual and bilingual speech.

Considering that CS is a speech style that would influence the phonetic production in bilingual mode in general and during the language change, variation studies are worth studying especially in CS large scale corpora and to investigate on both consonant and vowel variations. It should be noted that CS variations are few studied even in large scale corpora and spontaneous CS speech corpora.

1.5 Automatic Code-switching processing

With the development of speech technologies tools and the extensive research on language identification and Automatic Speech Recognition (ASR), a large set of research has been devoted to the CS speech (Vu et al., 2012; Yılmaz et al., 2016; Solorio and Liu, 2008; Lyu and Lyu, 2008). Works propose methods to process the switch at clause and word levels and use different cues like acoustic, prosodic and phonetic features. In this section, we report briefly the ASR methods used to process CS that made our studies possible with a focus on automatic speech alignment.

1.5.1 Automatic speech recognition

Automatic speech recognition is developing methodologies that enable the recognition of spoken language into text with the use of computer technologies (Mariani, 1990).

Concerning the ASR system of French language, evaluation campaigns were set up in 1997 and 2000s which contributed to the systems development for this language (Dolmazon et al., 1997; Gravier et al., 2004). The works are based on various speech types: controlled French speech, broadcast speech and spontaneous speech.

The Algerian Arabic system has been developed as an extension of MSA language system named ALASR (Arabic Loria Automatic Speech Recognition) (Menacer et al., 2017). This work concluded that Algerian Arabic (AA) is highly influenced by French language and needs to include French acoustic and language models to MSA. So, a recent interest was focused on the code-switching AA and French Language (FR) ASR (Amazouz et al., 2016, 2017).

The Laboratoire d'informatique pour la mécanique et les sciences de l'ingénieur-CNRS, Orsay, France (LIMSI) ASR system used for this thesis is built on conversational broadcast and telephone speech for both French and Arabic languages (Gauvain et al., 2003, 2002) (See Section 3.7).

The major components of ASR system are: acoustic model, the language dictionary and the language model. Acoustic model is used to classify the predicted phonemes of a language in a given audio input. Acoustic Model uses deep neural networks for frame predictions and the statistical model Hidden Markov Model (HMM) (Rabiner, 1989) to transform the units into sequential predictions. Pronunciation dictionary gives the phonological pronunciation of each word in the language and then it joins the acoustic model to decode the pronounced phones.

The language model consists in assigning probabilities to a sequence of words in order to validate the syntax and the semantics of a given word sequence in a language. This process is realized by modeling the probability for a word occurring in a given sentence and its predecessors.

1.5.2 Code-switching forced alignment

The audio speech segmentation into words and phones and its labelling with phone symbols are important steps for phonetic and ASR research. Obviously, there are a lot of programs that may be helpful during manual segmentation, for example, the PRAAT tool (Boersma et al., 2002; Boersma, 2017), which also enables a vast range of acoustic analyses. However, automatic speech alignment processes large quantities of speech in a very short time when manual processing can be estimated at around 800 $\times$ real-time (Schiel et al., 2012).

Forced alignment (automatic Forced Alignment (FA)) of speech (also called Forced Viterbi alignment (Forney, 1973) or automatic segmentation and labelling of speech (Ljolje and Riley, 1991; Brugnara et al., 1993)) is an automatic method to align the text with the audio signal and consequently segment the audio signal into words and phones. Forced alignment (FA) consists in using an ASR system in a restricted mode: the system is given not only the acoustic signal as input for which it should normally determine the best matching word sequence, but it is also given the reference transcription. The system's only job then consists in locating the word boundaries of these words within the acoustic signal, and more interestingly locating the phone segments within the words in the acoustic signal. Forced alignment thus makes it possible to automatically derive phone transcriptions and segmentation from an orthographically transcribed speech signal. To this aim, the ASR system requires a pronunciation dictionary including all the words occurring in the transcriptions together with a set of acoustic phone models. An interesting aspect of forced alignment is the possibility of introducing multiple pronunciation variants into the dictionary and let the system chose the best matching variant given the acoustic input signal. The FA method, which is illustrated in Figure 1.2, has been validated in several acoustic-phonetic studies dealing with pronunciation variants (e.g. French liaison, schwa, voicing assimilation, word-final devoicing, regional variants).

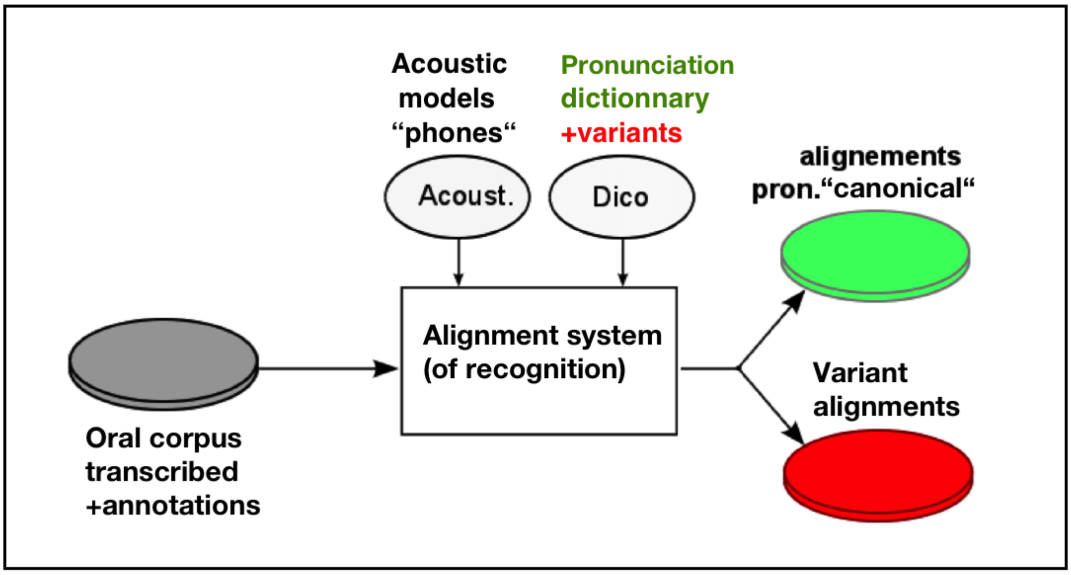


Figure 1.2: Forced alignment process with variants using ASR system (Adda-Decker and Lamel, 2017)

Thus, the data obtained with the FA constitutes a valuable database for linguistic and phonetic studies (Yuan et al., 2013; Adda-Decker and Snoeren, 2011). The methods used in this thesis to realize the CS FA are described in the Section 3.7). The FA of one language is based on its linguistic and phonetic properties: phones(mono-phones, diphones, triphones,) words, syllables and other phonetic classes as the tones. Works on CS automatic segmentation show that the FA of CS data can be realized differently and it depends first to the final goal of the task (phone, syllable, word, sentence, language alignment ...).

1.6 Code-switching corpora

After introducing some CS speech studies and problematic we present in this section a review of audio CS corpora in different languages pairs. We present here below a corpora collection in a chronological order and we summarize their salient attributes (Speech duration, number of speakers, number of sentences and utterances, etc.). We also dedicate part of this section to existing corpora in French-Algerian code-switching in both oral and transcribed (written) speech.

1.6.1 Code-switching speech corpora

- (Chan et al., 2005, 2009) The CUMIX Cantonese-English code-switching speech corpus is developed at the Chinese University of Hong Kong. The data is code-switched speech utterances read by speakers. The database contains 17 hours of read speech recorded from 40 speakers.

- (Lyu et al., 2006) Mandarin-Taiwanese code-switching speech corpus was developed for testing ASR experiments and language identification. The corpus contains 4000 Mandarin-Taiwanese code-switching utterances. The corpus counts 16 recorded participants.

- (Franco and Solorio, 2007) The English-Spanish code-switching speech corpus was compiled at the University of Texas. The corpus contains 40 minutes of transcribed spontaneous conversations of 3 speakers.

- (Ziamari, 2008, 2013) Intra-sentential of French-Moroccan Arabic code-switching of student conversations in university. The corpus contains 11 h speech conversation and 33 speakers. Conversations are recorded from Moroccan students in ENSAM institut. The corpus studies are about the French insertion in Arabic sentences and on Code-switching syntax typology.

- (Solorio and Liu, 2008) A crops of Spanish-English code-switching of 40 mn (∼8K words, 922 sentences) of speech conversation. Part-of-Speech (POS) has been labelled. The data counts 239 switches.

- (Lyu and Lyu, 2008) Designed a Code-switching corpus with Chinese dialects. It contains 4.8 hours of speech that corresponds to 46K bilingual utterances. Language identification and ASR experiments have been performed in this corpus.

- (Lyu et al., 2010; Vu et al., 2012; Lyu et al., 2015) The SEAME is a Mandarin-English code-switching conversational speech corpus developed in Nanyang Tech-nological University in Singapore, and University Sains Malaysia. The data contains 192 hours of transcribed CS speech interviews.

- (Shen et al., 2011) CECOS (Chinese-English code-switching speech corpus) was designed in the National Cheng Kung University in Taiwan. It is a big corpus that contains 121 hours of speech data collected from spontaneous code-switching speakers. the corpus groups 77 participants.

- (Li et al., 2012) Mandarin-English code-switching corpus of 5 hours of CS speech. The corpus groups conversational speech meeting, student interviews in causal speech and text data of on-line news. The corpus duration totalled 8 hours of audio speech. The texts contain 10030 Chinese words and 2688 English words.

- (Ahmed and Tan, 2012) The Malay-English code-switching corpus consists of 100 hours of Malays a Malay-English code-switching speech data from 120 Chinese speakers, 72 Malay and 16 Indian speakers.

- (Imseng et al., 2012) MediaParl is a Swiss accented bilingual database designed with French and German bilingual speech recordings in Switzerland context. The data was recorded at the bilingual Swiss canton Valais. The data contains a considerable set of local accents and dialects.

- (Modipa et al., 2013) A corpus of Sepedi-English code-switching speech corpus was created by the South African CSIR. It consists of 10 hours of radio broadcasts speech and read speech data by 20 Sepedi speakers.

- (Dey and Fung, 2014) A Hindi-English Code-Switching Corpus of (HKUST) university in HongKong University . CS student interviews of 12 speakers are recorded . The corpus counts 400 code-switching words, 30 mn of conversations and collected from 9 speakers . The sentences are tagged with inter-sentential and intra-sentential code-switching.

- (Yilmaz et al., 2016; Yılmaz et al., 2016) FAME! corpus is a Frisian-Dutch code-switching speech corpus of radio broad-cast speech. It was developed at Radboud University in Nijmegen. The recordings are collected from the archives of Om-rop Fryslan, the regional public broadcaster of the province Fryslan. The database covers almost a 50 years time span.

- (Çetinoğlu, 2017) German-Turkish CS of 5 hours of students CS speech. 28 participants male and female have been recorded. The corpus includes a self-evaluation of languages proficiency. The corpus is tagged by sentence types (intra-sentential and inter-sentential CS).

- (Niesler et al., 2018) A South African speech corpus containing four pairs of code-switching languages: English-isiZulu, English-isiXhosa, English-Setswana, and English-Sesotho. The corpus utterances are obtained from South African soap operas by Ewald vander Westhuizen and Thomas Niesler. The soap opera speech is typically fast, spontaneous and may express emotion, with a speech rate higher than prompted speech in the same languages.

- (Hamed et al., 2018) The Arabic-English is recently developed by Injy Hamed, et al., by conducting the interviews with 12 participants.

- (Sreeram et al., 2018) Hindi-English Code-Switching Speech Corpus. The database is composed with CS text and oral speech. Texts come from blogging websites and covers different contexts and topics. The corpus has different speech annotations: speaker variations(pronunciation, accent). It counts 7K CS utterances in total and by 71 participants.

1.6.2 French-Algerian Arabic code-switching corpora

- Abainia (2019) designed Algerian Arabic-French code-switching corpus called DZDC 12 . It is a collection of 2400 Facebook comments and texts in where the scripts are written in Latin characters. This corpus is organized by gender, region and city, and is converted in Arabic script. The corpus also gives translations into Modern Standard Arabic and different annotations allowing multidisciplinary studies: word and sentence annotations, emotions tags.

- Cotterell et al. (2014) An Algerian Arabic-French Code-Switched Corpus from newspaper websites. The text containers dialectal Arabic and French content. The corpus contains discussion on a wide-ranging set of issues including domestic politics, international relations, religion, and sporting events. The data counts 6.949 comments and 150,000 words.

- (Benhattab, 2016) A corpus of code-switching Algerian-Arabic utterance The total of words displayed Mangas is of 6067 words, 131 words uttered in Algerian Arabic and 5936 words in French.

Since CS is a widespread practice among bilingual Algerian speakers, some Algerian Arabic language corpora emphasize Algerian Arabic-French CS data and bilingual utterances. Abidi and Smaïli (2017) in Algerian Arabic Youtube comments (CALYOU corpus) (Abidi et al., 2017) composed 17M of words, notes 82% of the data contains CS items, among French-Algerian Arabic CS.

Chapter 2

French-Algerian Arabic code-switching

Chapter contents

THE AA is one of the Arabic dialects used in informal situations and in daily life contrary to the MSA which is mainly used in formal communication as in education, press news, media and political speeches. AA is the mother tongue of about 40 million 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 (Caubet, 2002). 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* (Saadane and Habash, 2015), 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.

2.1 Two languages in contact

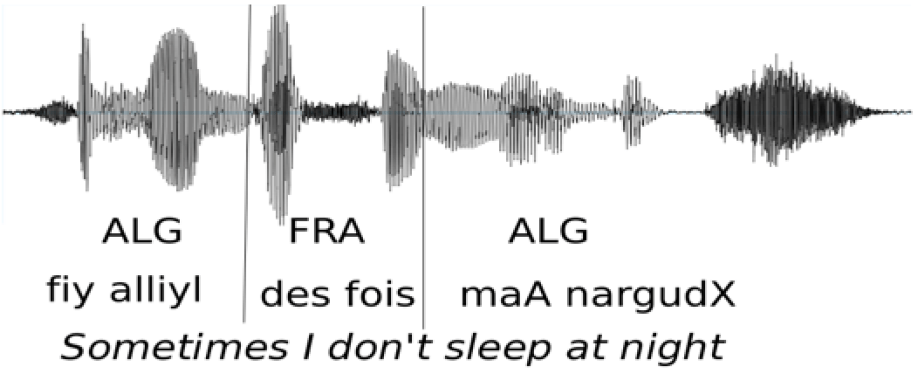


Figure 2.1: Example of a code-switched sentence with an insertion of FR adverb locution "des fois" in AA base

AA is an oral dialect of North African Arabic dialects group spoken in Algeria and it is the mother tongue of 80% of Algerians next to the local Berber languages in Berber communities. AA is different from Modern Standard Arabic AA 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. However, AA written form becomes more and more widespread especially in social medias. Commonly, AA is written with Arabic characters 3.4. This script is written from right to left. Too, another form of AA script transliterated with Latin characters called "Arabizi transliteration" widely used on the internet and SMS (Cotterell et al. (2014); Al-Badrashiny et al. (2014); Bies et al. (2014). Thus, written AA can be found with at least two different script types.

French 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, in France and other French-speaking countries: Belgium, Switzerland, Canada...

Language contact between FR and AA is related to a historical context : Algeria was occupied by the French from the mid 1830 to 1962, even from Independence day to nowadays. French is still taught from primary school as first foreign language. Many university degrees are taught in French which is also present in media and newspaper. In fact, in Algeria we find French and Arabic national newspapers and radio/TV shows.

The coexistence of FR and AA within French-Algerian communities and the use of FR and AA languages in everyday life allows speakers to develop and adopt a bilingual speech strategies whether for direct speech quotation in L*b*, or to express referential subject in L*b* within a L*a* (Richer, 2015; Boyer, 2001; Zeroual et al., 2006), or to converse about a socio-cultural topic with a French base, which implies AA insertions.

This bilingual context and the coexistence of both these languages, FR-AA bilinguals adopt a bilingual speech which implies CS and borrowing forms in daily conversations, medias, social medias, radio TV shows, movies and even in music lyrics (Bentahila and Davies, 2002; Bullock and Toribio, 2009; Wiedemann, 2015). CS motivations emerge from collective and individual choices, in fact, the large bilingual community allows and encourages this kind of speech. Meanwhile, the CS type (intra or extra sentential) depends on the speaker's individual choice.

Finally, bilingual speakers may easily switch within a conversational context and CS becomes a daily habit. In French Algerian Arabic bilingual communities, the two languages are of complementary use in daily life (Zaboot, 2010). The speakers tend to use both FR and AA in their daily conversations resulting in an interdependent bilingualism (Zaboot, 2010, p.208), so the CS is frequent in this pair of languages (Kheder and Kaan, 2016).

2.2 Code-switching speech in media and social media

In this section we will present several works that were conducted on FR-AA CS to describe this pair of language in CS and also the corpora contribution to this field of research.

2.2.1 Code-switching in medias

CS in media is widespread, especially in entertainment shows where speech is less formal and speakers young. FR-AA CS has the highest frequency compared to other similar language pairs such as Tunisian Arabic-FR and Moroccan Arabic-FR CS (Amazouz et al., 2016). FR-AA CS is also present in newspapers, in fact FR inserts were found in AA in Arabic language newspapers that use MSA and AA in their articles (Cotterell and Callison-Burch, 2014). FR-AA CS is also used in newspaper caricatures in order to express social and political humour which follows the bilingual language practices of the community (Hadjadj, 2015).

2.2.2 Code-switching in social networks

With the spreading of the use of computing technologies in communication (social media) FR-AA CS crosses oral borders to develop in another form of written CS. This allowed speakers to have another form of bilingual speech practices and also to develop the oral form of AA which has always been practiced only in its oral form. Hence, many studies and large scale written corpora on FR-AA CS written speech aroused (FR-A CS tweets corpus (Cotterell et al., 2014), YouTube comments corpus (Abidi et al., 2017)).

2.3 Description and comparison of phonemic inventories of French and Algerian Arabic

In this section, we describe first the consonantal and the vocalic system of the languages that constitute the code-switching studied in this thesis work. Also, we compare between AA and FR phonemic system for both consonants and vowels.

2.3.1 Phonemic inventory of Algerian Arabic

The consonantal system of AA is based on Arabic (MSA and classical Arabic). However, following historical situation and language contact with European and local languages, AA has been phonetically and phonologically influenced by Berber dialects, French, Italian, Spanish and Turkish languages. These influences introduced modifications on the consonantal and vocalic system of AA. So, the AA has additional consonants and vowels com-

pared to the MSA and, in return, some of MSA/classical Arabic phonemes are not produced in AA speech (Saadane and Habash, 2015; Saadane, 2015).
Furthermore, it should be noted that the AA groups have several varieties of regional sub-dialects (Derradji et al., 2002; Harrat et al., 2014, 2016) that are marked, in part, by phonetics and phonological differences. Four intra-dialectal varieties are classified (Saadane, 2015). 1) Algérois: is spoken in central north regions of the country (Algiers regions). This sub-dialect represents a large number of speakers. 2) Oranais: is spoken in the Western and Moroccan frontiers (Oran region). 3) Eastern dialects: the speakers are located in the eastern regions of Algeria. 4) Saharan dialect: is the dialect of the south of Algeria population. The classification of theses sub-dialects are based on The AA dialect described in this work is the dialect of the Algiers region.

The following subsections give an overview of the consonants and the vowels produced in AA.

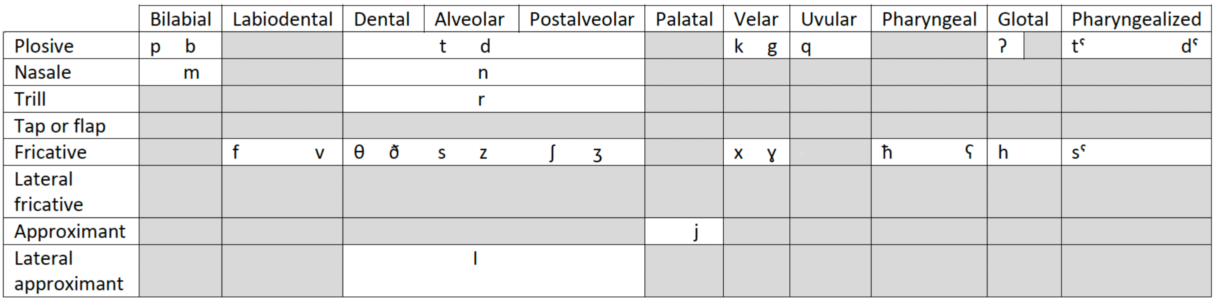
2.3.1.1 Algerian Arabic Consonants

Speech sounds are divided in two categories: the vowels and the consonants. The consonants are produced with obstruction or friction of the air-flow (Ladefoged, 2003; Ladefoged and Disner, 2012). Articulatory, depending on the passage of the air-flow the obstruction can be total (stop consonants) or partial (Approximant consonants). The air-flow can go out through the nose (nasal consonants). To produce the consonants, the lips, teeth, alveolars, palate, velum, pharynx and glottis move or interact with tongue movements. This combination of manners and places of articulations allow to produce multiple distinctive consonants. The consonants can be joined also by the vocal fold vibrations and form voiced consonants. Thus, the voicing gives distinguished consonant pairs with one voiced and another voiceless. A large number of consonantal systems is based on this voicing distinction such as French and Arabic.

The AA has in total 29 consonants and two glides /w, j/ with 25 shared consonants with MSA and 3 consonants /p, g, v/ borrowed from European languages and Berber languages.

The AA allows pronunciation variation for a part of consonants. The consonants /p, g, θ, ð, t/ form minimal pairs like: [paːsˤ a] *he passed* [paːsˤa] *he is compromised*. It can be, at the same time, free variants (See section 3.5.2.2). Examples of free variants: [taːniː] and [θaːniː]/ *also, too*, [qa:l/] [ga:l] *he says*, [plaːsa] [blaːsa] *A place*. Examples of minimal pairs: [qəsˤba, gəsˤ ba] *Casbah, Berber wind instrument*, [qliːl, glliːl] *few, poor/humble*.

According to the articulation gesture, AA consonants have plain consonants with one gesture of articulation and consonants that are realized with pharyngealization which is a secondary articulatory gesture of the plain consonant (pharyngealized consonants). Also, Algerian Arabic, like all Arabic dialects, has geminate consonants which are a doubling articulation of the plain consonants. Figure 2.2 summarizes AA consonants in their manner and place of articulation including the pharyngealized (emphatic) consonants.



	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Palatal	Velar	Uvular	Pharyngeal	Glotal	Pharyngealized
Plosive	p b			t d			k g	q		ʔ	tˤ dˤ
Nasale	m			n							
Trill				r							
Tap or flap											
Fricative		f v	θ ð	s z	ʃ ʒ		x ɣ		ħ ʕ	h	sˤ
Lateral fricative											
Approximant						j					
Lateral approximant				l							

Figure 2.2: AA consonants in IPA classified by articulation mode and manner (Vertical axis) and the place of articulation (horizontal axis). The last column corresponds to the existing pharyngealized consonants. The geminates are not included in the table but symbolized in the corpus with the doubling of consonants. The symbols that appear in pairs, the one to the right represents a voiced consonant. Shaded areas denote articulations that are not observed in AA or the articulation is not possible

Plain consonants: AA has 28 plain consonants and two glides /w, j/ as well as MSA and other Arabic dialects. The language counts height stops with an important part produced with voicing distinction: labials /p, b/, alveolar /t, d/ and the velar /k, g/. The language also includes the uvular stop /q/ and the glottal stop /ʔ/.

AA includes a large set of fricatives. All fricatives are produced with voiced/unvoiced opposition, except the voiced glottal consonant /h/. The fricatives of AA are, ordered by voiced and unvoiced pair: labiodentals /f, v/ dentals /θ, ð/ alveolars and post alveolars /s, z/, /ʃ, ʒ/, velars /x, ʁ/, and pharyngeals /ħ, ʕ/. Figures 2.3 and 2.4 show examples of /p/ and /b/ AA consonants that share the same manner and place of articulation and differ in the voicing. The voicing is visible in the "voice bar" (Fant, 1960) of the consonant /b/ and the devoicing is marked by a silence that is represented with a white band in the spectrogram as highlighted in the figure. The nasal consonants of AA are the labial /m/ and alveolar /n/. The phonological system of this language also includes the alveolar trill /r/ and lateral /l/. The glides of AA are /j/ and /w/.

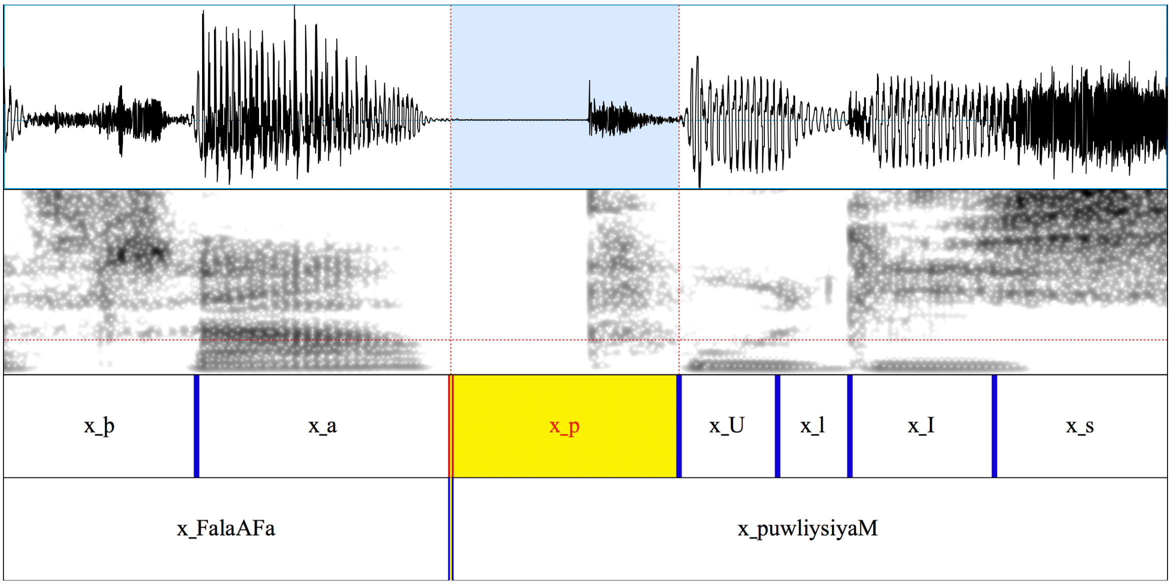


Figure 2.3: Spectrogram (lower graphic part) and oscillogram (top graphic part) of the Algerian Arabic consonant /p/ in word onset in the word "pulissiya" يُولِسِيَّة *police*

Pharyngealized consonants: The pharyngealization is defined as "superimposition of a narrowing of the pharynx" (Ladefoged and Johnson, 2014, p.235-236). This second articulatory gesture can join to plain consonants and forms pharyngealized consonants, also called emphatic consonants in Arabic languages (Ladefoged and Maddieson, 1996; Al-Solami, 2013, p.265-266) (Al-Ani, 1970).

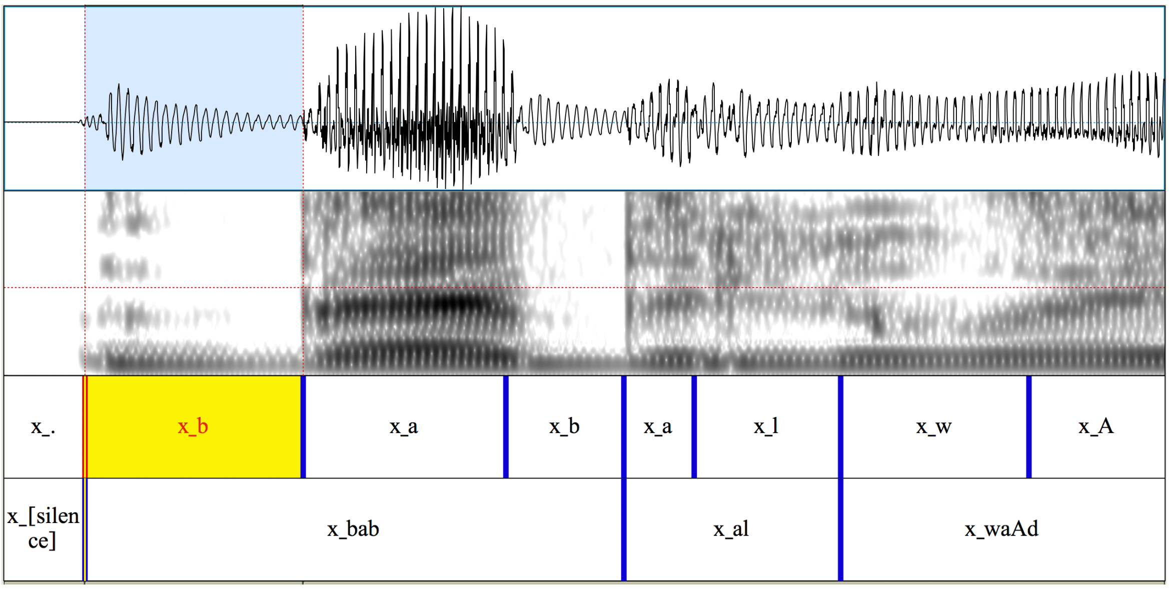


Figure 2.4: Spectrogram (lower graphic part) and oscillogram (top graphic part) of the Algerian Arabic voiced consonant /b/ in the word attack "bab" بَاب *door*.

AA has three phonological pharyngealized coronal consonants: the stops /tˤ, dˤ/ and the fricative /sˤ/. They are counterparts of the plains consonants /t,d,s/. Unlike MSA which counts four emphatics /tˤ, dˤ, ðˤ, /sˤ /, AA speakers use dˤ, ðˤ like free variants. At this time, there are no studies that demonstrate which consonant is the variant of the other and their variation in Algerian Arabic sub-dialects. However, observations of FACST corpus productions (Amazouz et al., 2018) AA speakers indicate that there is no phonological distinction between /dˤ/ and /ðˤ/.

The pharyngealization affects the adjacent vowels in CV syllable and modifies their quality (Ladefoged and Maddieson, 1996). Multiple studies in different Arabic varieties reported about decreasing F2 frequency, which is related the back tong position, of adjacent vowels (Al-Ani, 1970; Mohamed, 2001; Ghazeli, 1978; Boxberger, 1981; Giannini and Pettorino, 1982; Watson, 1999; Al-Masri and Jongman, 2004; Al-Tamimi and Heselwood, 2011; Benamrane, 2013). Also it has been found that F1 and F3 frequency may increase in this context, and then the formants F1 and F2 become closer in emphatic context (Hassan, 1981; Ferrat and Guerti, 2013; Al-Tairi et al., 2016) as shown in the figure 2.5.
In short, acoustic cues for studying consonant pharyngealization are the formants measure-

ments of the adjacent vowels (See also the sections 2.3.1.2 2.3.2.2 about vowels and formants).

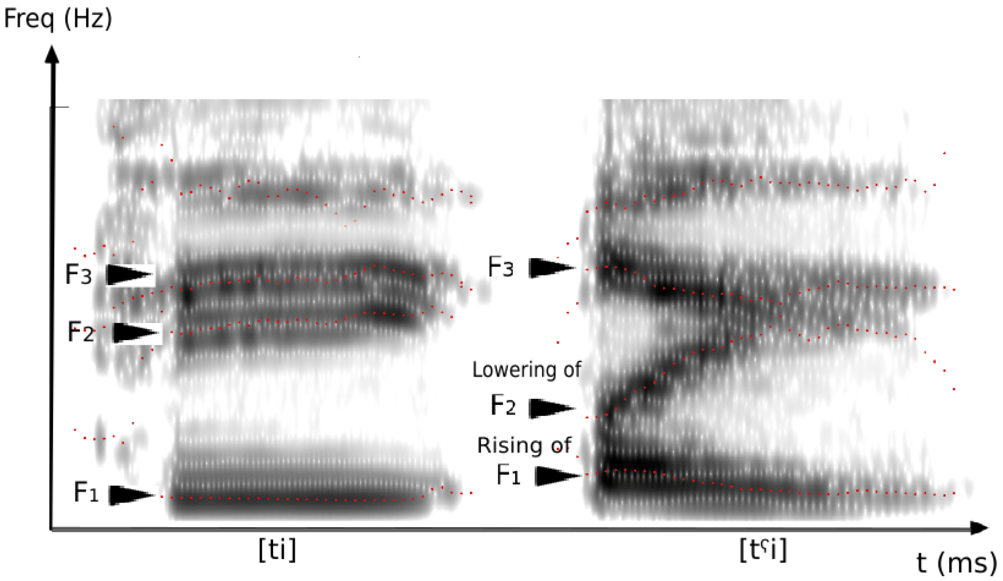


Figure 2.5: Spectrogram of Arabic plain stop [t] compared to its emphatic counterpart [t^{ʕ}] in [*C*i] context.

Geminates and gemination The gemination is the consonant doubling in articulation (Delattre, 1971; Crystal, 2011; Ladefoged and Johnson, 2014). The phenomenon has been observed and studied in a lot of Arabic varieties (Khattab, 2007; Dell and Elmedlaoui, 2012; Ferrat and Guerti, 2017).

All AA consonants can be geminated and gemination is a phonological property in AA (Kiparsky, 2003; Souag, 2006). Indeed, the language contrasts singletons and geminates, and minimal pairs appear in words like: /bat^{ʕ}al/ *hero*, /batt^{ʕ}al/ *break a habit*. Beyond the phonological status of geminates and gemination of consonants in Arabic, they are orthographically marked by a diacritic ّ called *Shadda*.

The gemination of consonants can also be achieved by phonological rules, such as the gemination of coronal consonants followed by the defined article "al" أَلْ *the*, the article's consonant being assimilated to the following coronal.

Algerian Arabic also includes post-lexical gemination (Kenstowicz, 1994; Ridouane, 2010). It is produced when two identical adjacent consonants belong to two different words. Examples: the words /duwwar/ and /raːsuː/ [du**ww**arra:su:] *he tuned his head*. This gemination is phonetic and has not phonological status.

The common acoustical characteristics of geminates is the segmental duration (Hassan, 2003; Zeroual et al., 2008; Ham, 2013; Ferrat and Guerti, 2017). It has been found that the geminate consonants are longer than their singleton counterparts. However, other parameters can be taken into account in order to identify acoustically and distinguish geminates and singleton counterparts as the adjacent vowel duration, the vowel quality in syllable (Al-Tamimi and Khattab, 2011) and stop geminates VOT values (Zeroual et al., 2006).

2.3.1.2 Algerian Arabic Vowels

Referring to the spelling and transliteration of AA Harrat et al. (2016); Amazouz et al. (2018), the language counts six oral vowels contrasting in duration (3 short vowels, 3 long vowels) as well as MSA. The AA vowels are [i, a, u, iː, aː, uː]. however studies about the comparison between AA and MSA in production shows that a large part of vowels in MSA vowels are deleted or reduced to schwa in AA (Mokhtar, 2018) Ex: rasma rsam *draw*, fahima fham *understood*. Some studies in AA have observed a 7th central mid-opened vowel /e/ or /e:/, it called *inclinaison* (Sara, 2007), and (Guella, 2011) have even transliterated in order to facilitate the pronunciation of the words . Indeed, minimal pairs with /a:/ and /e:/ have been observed in AA, in Saadane and Habash (2015); Saadane (2015) works on AA transcription suggested the utility of a distinctive transcription for the vowels /a:/ and /e:/ in especially in minimal pairs: [da:r] *turned* and /de:r/ *he did*.

2.3.2 Phonemic inventory of French

2.3.2.1 French Consonants

The phonological system of French counts 21 consonants (Fougeron and Smith, 1999): seven stops with voiced and invoiced contrast pairs /p, b, t, d, k, g/, three nasals /m, n, ɲ/, six fricatives with voicing contrast pairs /f, v, s, z, ʃ, ʒ/, three central approximates /j, ɥ, w/ and one lateral approximant /l/. It also counts one Allophone of /ʁ/.

Although geminates are not phonologically included in French, some contexts and pronunciation of identical adjacent consonants occur with gemination (Malmberg, 1944; Fagyal et al., 2006; Delattre, 1971). We distinguish in French two forms of gemination: lexical gemination and post-lexical gemination.

The gemination within lexicon follows the double consonant letters in spelling e.g irresponsable *irresponsible*, illisible *illegible*, grammaire *grammar*. However, unlike AA, the spelling in FR is a relative indicator to gemination. Indeed, a large set of words are written with double consonants but the consonants are not pronounced like geminates: lettre *letter*, appel *call*. The lexical gemination in French is produced according to rules (Fagyal et al., 2006). Examples: /ʁʁ/ in verbs requires geminate pronunciation in order to time distinction mourait *he/she is dying* (past form) /muʁɛ/ and mourrait /muʁʁɛ/ (conditional form). Gemination is also produced when a noun that starts with r, l, m, or n is preceded by an antonym prefix that has the same letter at its end to double the initial first letter of the noun (irrégulier,illisible, immature, innombrable).

Post lexical gemination occurs in word boundary situations such as "il l'aime", *he loves her*, which is different from "il aime" *he loves*), il l'a dit *he said it*.

Despite its distinctive role, lexical gemination is not necessarily in production, Yaguello reported that lexical gemination is produced by highly educated persons.

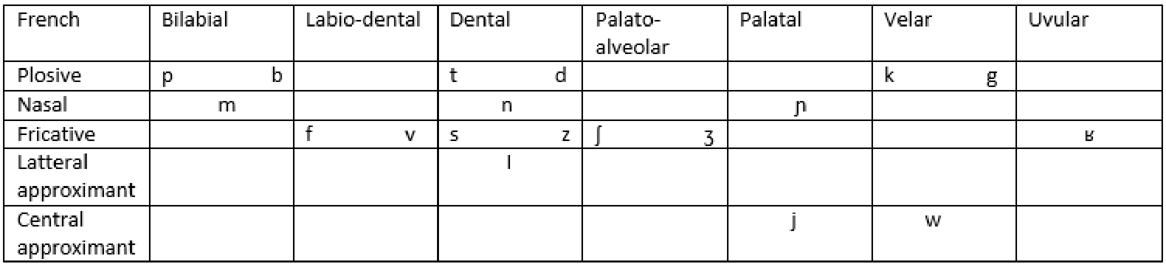
2.3.2.2 French Vowels

Nowadays, the French vowel system tends to be described using eleven oral vowels /i, e, ɛ, y, ø, œ, a, ɑ, ɔ, o, u/, three nasal vowels /ɛ̃, ɑ̃, ɔ̃/ and one schwa sound ə (Houdebine, 1981; Derivery, 1997; Fougeron and Smith, 1999). In some regional varieties, a fourth nasal vowel can be found ([œ̃]).

2.3.3 Comparison of phonemic inventories of French and Algerian Arabic

2.3.3.1 Consonants

By comparing the Algerian-Arabic and FR, we notice that the AA has a consonal system with more consonants than French and the /y, w/ are considered as consonants since they can be geminated. French consonants are limited to plain consonants with one nasal that does not exist in AA system /ɲ/. The consonants /p, v, g/ are produced in AA, mostly in borrowed French and other European languages words. Thus, Algerian Arabic and French share 20 consonants. The AA phonological system covers all FR consonants excepted the nasal /ɲ/. The figure 2.6 illustrates the consonant in IPA table.



French	Bilabial	Labio-dental	Dental	Palato-alveolar	Palatal	Velar	Uvular
Plosive	p b		t d			k g	
Nasal	m		n		ɲ		
Fricative		f v	s z	ʃ ʒ			ʁ
Latteral approximant			l				
Central approximant					j	w	

Figure 2.6: Shared French and Algerian Arabic consonants

2.3.3.2 Vowels

The figure 2.7 illustrates the vowels of French and Arabic. The principal difference between the languages is the French vowels have two types of vowels: nasal vowels /ɛ̃, ɑ̃, ɔ̃ and oral vowels in contrast to AA that counts only oral vowels.

The figure also shows that French has a richer vocalic system than AA. The FR includes opened and mi-opened vowels /e, ɛ, o, ɔ/ and rounded front vowels /ø, œ/ (Delvaux et al., 2002). AA vocalic system is limited to the extreme vowels /i, a, u/. Another difference between FR and AA that should be noticed is that AA has short and long opposition in all of its vowels.

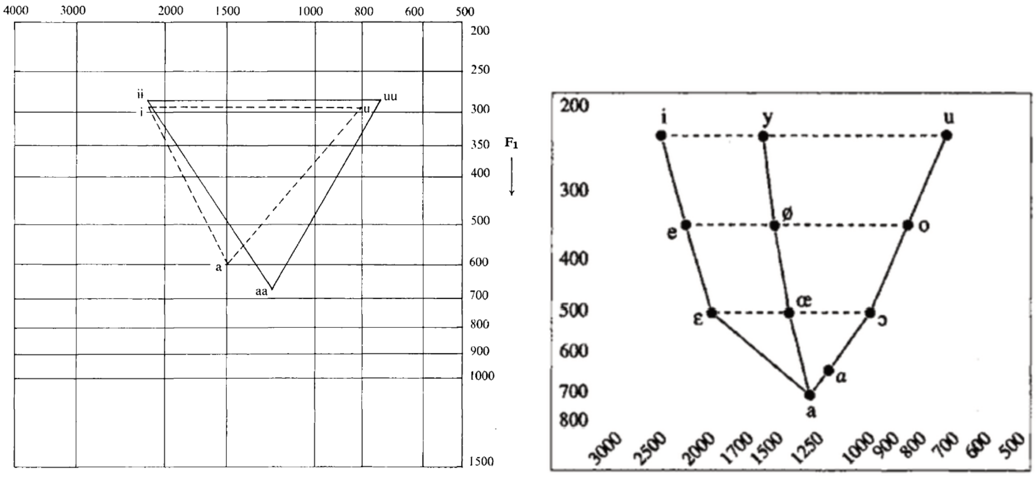


Figure 2.7: Arabic vowel diagram (Al-Ani, 1970, p.25) in the left and FR vowel diagram (Delattre, 1966) in the right

2.4 Research questions of the thesis

This state of the art review tells us that CS is highly practiced in bilingual communities and it is a growing research field and that there are still many paths to be explored and many questions to be answered. In fact, large scale corpora allow the research to study a big amount of data and thus, to obtain precise and detailed results in all linguistic research fields (phonetics, sociolinguistics, socio-phonetics, morphosyntax and language technologies...).

One of the questions that we try to answer in this thesis is how can we collect an important amount of spontaneous CS data in order to investigate this data in different linguistic research fields. The CS that we try to study is a spontaneous CS, we try to get a spontaneous CS speech without limiting the research to language change. In fact, it is important to distinguish between language change which is a language switch that is independent from the conversational context and that is triggered with elements that are external to the sociolinguistic setting, example: experiments conducted with instructions to switch at precise moments (image switch comments, switching following clues, etc.). The CS we aim to obtain is a spontaneous production that is triggered spontaneously by reproducing all the elements of a sociolinguistic setting. We enhanced the space setting by recording the conversations in a soundproof room to get the best quality sound recordings in order to enhance the precision of our data results.

Collecting a spontaneous high sound quality data is a thing, but once this data is collected and analyzed, to what scale or reference it should be compared to allow us to answer our questions about variation for example. Besides, CS literature showed that both languages interactions may have an influence at a phonetic level and thus produces variations in production.

The first research question that arises here is; does CS have an impact on the phonetic productions of vowels, knowing that the vocalic systems of FR and AA are different? Thus, to what extent this difference can be considered as a production variation at the moment of CS speech?

The hypothesis that we propose is that vowel production may vary in FR because of the influence of AA which has a low number of vowels and that may push the speaker to only use the minimum of necessary vowels. The second hypothesis that we propose is that FR may influence AA vowel production, FR which has a phonological system which is richest to the AA one may influence AA vowels production.

Phonetic variation can also impact consonants, starting with the hypothesis stating that the AA phonological system has geminates and a second articulatory gesture emphatics, thus we asked the question of the influence of the presence of two phonological systems in FR and AA CS speech and what are the gemination and emphatization consonants variation that we can observe in this context?

The hypothesis that we establish states that FR consonant my be affected by the gemination and the emphatization in CS speech. Also, AA geminates and emphatics may vary to simplification due to FR impact in CS.

Finally, the literature showed that voicing variation in consonant are produced in a lot of languages. So, in CS speech what are the voicing consonant variation in FR and AA and what are the most affected consonant by this variation. How the voicing and devoicing opposition in AA and FR can influence speech variation and consonant variation of both

languages in CS speech?
The hypothesis that we consider here states that voicing variation may be produced in both languages in CS but the difference between the involved consonants may be different from language to another according to their respective impacts and influences. Furthermore, voicing may vary with the phoneme position within the word, and CS speech may have an impact on the voicing of each phoneme in each word position.

In the forthcoming we try to answer those questions by following large scale corpora methods of treatment using automatic language processing and automatic speech recognition tools.

Part II

Realized works

Chapter 3

The FACST corpus

Chapter contents

An important part of this thesis has been devoted to design an oral CS corpus. FACST was created in order to support a variety of studies in phonetics and natural language processing. The first aim of the FACST corpus is to collect a spontaneous CS speech 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 chapter describes the 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 linguistics and acoustic-phonetic studies. We provide 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.

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

Table 3.1: Compact presentation of FACST corpus

3.1 Reflections about code-switching corpus design

The corpus reports about spontaneous speech of speakers guided by linguist questions. It should be precised that the corpus is not a recovering of existing spontaneous code-switching practices like conversations in medias, web speech data or recordings of exchanges among friends, colleagues or family members. Also, it is not a stimulated CS with speech tasks like the switch at the beep for example or comment images and switch at the image change. The aim of FACST data is to reproduce recorded CS conversations as it is practiced in natural and spontaneous situations. In addition, the speakers are informed that they are recorded. That may change the scale of the productions' spontaneity. So, record CS speech in this corpus is a challenge that follows multiple conditions of CS production. The design of FACST corpus started with a reflection that tempts to answer the following main question: how to obtain spontaneous and natural speech with a considerable amount of CS in a soundproof room of a laboratory? Or briefly, how to create CS speech environment to collect CS data?

In recorded speech corpus, the speakers choice is an important step. So, the first step to build a corpus is selecting code-switching profiles. It should be noted that all CS speakers are bilinguals, but not all bilinguals use CS in their speech productions. So, the speakers selection as shown the Section 3.2 is based on language proficiency the personal motivations to code-switch with two languages.

Besides, creating a natural environment for CS practices requires a reflection about the conversational context of CS and it is related to the speakers' motivations in order to encourage the use of two languages in a same interaction. Thus, developing recording protocols (See Section 3.3) with getting closer to the conversation topic of the speakers is one of the methods that we describe in the following sections.

This corpus interests also to the methods to eliciting CS and illustrates the role of the linguist interviewer to contribute to produce the code-switched speech and collect the data. Section 1.3.2 addresses the methods that we used and the show how the date has been collected.

To summarize, the constitution of the FACST corpus is designed on reflections that are shared between the CS speakers profiles, the methods to elicit CS in spontaneous conversations and the collection of spontaneous CS data in laboratory.

3.2 Code-switching speakers participants

Following the studies about the Factors Influencing Language Choice (Grosjean, 1982, p.136) and social and individual CS practices motivations (Gumperz, 1982; Gardner-Chloros, 2009a; Bullock, 2009), we created first a sociolinguistic online questionnaire named Experience of Code-switching practice (ECSP) to select CS profiles (see Section C.1 and here the questionnaire link). In this questionnaire, the potentially recruited speakers to answer the questions about three spheres [1]. The first one is a form about personal information (sex, age, education, country of birth, country in which the speakers live). The second is about the spoken languages, their bilingualism and their linguistic autobiography. The speaker answers to questions about the mother tongue, second language, the manner and the age of learning/acquisition of the languages. The last part is about the practices of the CS. We put in place questions about their opinion about language mix and the situations/factors favouring the CS speech.

The principal aim of this questionnaire is to select young adult bilinguals who can reproduce natural and spontaneous CS during the recording. The selection was based on the answers of the bilingual speakers about several criteria. The first criterion is the bilingual proficiency. The selected speakers declare to have the same or similar language skills of L*a* and the L*b*. Also, all selected speakers have lived a part of their life in Algeria and a part of their life in France. Also, they studied all at university and most of them did their education in both countries. We focalized too on the young adult category, the speakers are aged between 23 to 39 years, with an equal number of men and women. Finally, all speakers declare to practice frequently CS with other bilingual speakers in their entourage. The context of CS usage varies between work, family, friends, studies, work, friends conversations, life in

[1] The questions of the ECSP questionnaire are in French. The language proficiency level is not tested. The selection is based on the declarations of the speakers.

both countries and each speaker uses CS at least in two domains.

In short, the ESCP online questionnaire was made to select a homogeneous group of speakers to record of French-Algerian Arabic CS. It is the first step to create the CS speech corpus. 20 CS speakers have been selected, 10 males and 10 females, following sociolinguistic criteria (age, education, bilingualism proficiency, language acquisition, life and study in two countries). The details of the ESCP questionnaire and selected speakers' feedback are in The Appendix section C.1.

3.3 Speech data Records protocols

The speakers were recorded in 2016-2018. The audio records have been realized in a soundproof room at Labotaroire de Phonétique et Phonologie-Sorbonne Nouvelle, Paris, France (LPP) or in a very calm room. 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.

For that purpose, each recording session started with an unrecorded preliminary conversation with the speaker, to get her/him in a relaxed setting and encourage a spontaneous speech productions. Indeed, starting to practice both languages in the same interaction before records helps to create CS conversational context and a spontaneous environment. This unrecorded conversation also helps to get more information about the speaker's production like language choice, language preference (Auer, 1995; Myers-Scotton, 1995) and let the linguist be attentive to the spoken language negotiation during the interactions (Auer, 1996). These recorded conversations, contribute, with the conversation guide of the linguist, to elicit CS during the records.

The recording session starts with a first task of reading texts in both French and AA (see section 3.3.2. The Reading texts are followed by discussions designed to elicit spontaneous speech with CS production. Figure 3.1 shows the FACST corpus organization. We resume it in two parts: controlled speech with reading texts task and spontaneous speech triggered by questions.

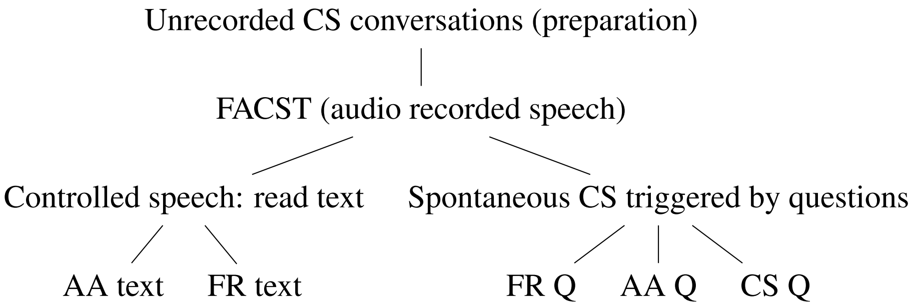

Unrecorded CS conversations (preparation)
FACST (audio recorded speech)
Controlled speech: read text
Spontaneous CS triggered by questions
AA text
FR text
FR Q
AA Q
CS Q

Figure 3.1: FACST data speech organization.

3.3.1 Eliciting Code-switching

As indicated in Section 1.3.1, various sociolinguistic and individual factors affect the quantity and the type of the CS. Also, the CS can be triggered by linguistic elements (words, utterances, sentences ...) Kees de Bot and Isurin (2009); Mirjam Broersma (2009) and it can be a continuation of a pre-existing conversation in CS.

In this corpus, the methods to elicit the CS in conversation are based on a question guide (see the question guide in the Appendix A.2). In addition to creating a relaxing and a spontaneous environment to CS and choose topics that interest the speaker (hobbies mentioned by the speakers in the ECSP for example), short open questions in French and AA are asked to the speakers with insertions of *trigger-words*. We can classify the question types in three parts: (1) monolingual questions in French, (2) monolingual questions in AA, (3) code-switched questions in each French and AA as base language [2]. The conversations are lead by the context conversations to trigger the CS. So the main common questions for most of speakers are related to the following topics and distributed in the following languages :

1. AA questions about their studies and work in France.

2. FR questions about their lives and studies in Algeria.

[2] We note that the base language in this study is the language that presents more grammatical and lexical elements in the sentence or into the segment. In FACST corpus there are code-switched questions where it is hard to define the basic language. Example: " **wiyn** ils seront **mlaAH**?" *where they will be comfortable?*

3. FR/CS: code-switched questions using French as a base language, questions about education comparison in both countries.
4. AA/CS: questions using AA as a base, question about differences in life between France and Algeria.

The open questions allow the speakers to speak freely, discuss at length in the conversations and allow the linguist speech to occupy a very short part in the records. It's important to note that the question guide represents just the main lines to conduct a conversation and the conversations take place spontaneously during the records.
The table 3.2 resumes the CS speakers answers as follows : most of the language productions are in French code-switched with AA. In the spontaneous CS triggered with balanced questions, we notice that CS frequency is considerably high (see the quantitative review of FACST in the Section 4. In CS FR/AA triggered, the language base can change within a same sequence of conversation and the speaker can change to the Embedded language several times in a list of successive sentences (Mondada, 2007).

This method aims to gain insight into CS practice by analyzing how CS may be induced and elicited. The goal of this protocol is to obtain a natural and spontaneous data speech with a maximum of CS productions.

	Language questions	Trigger questions (examples)	CS type gathered
1	AA monolingual questions	wiyn qriyti? *Where did you study?* kunti taxadmi fiy EljzaAyar? *Did you used to work in Algeria?*	1- inter-sentential CS in a French interaction with AA sentences: 2- intra-sentential CS with French base 3- intra-sentential CS with AA base
2	FR monolingual questions	qu'est ce que tu faisais pour remedier à ce problème? *How did you manage to resolve this problem?*	1-inter-sentential CS in a 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 as a base	cajbak al texte? *Did you like the text?*	1- inter-sentential CS in a French interaction 2- intra-sentential CS with French as a 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 a French interaction 2- intra-sentential CS with French base

Table 3.2: Examples of questions to trigger code-switching in interviews. The last column shows the types of the code-switching at the sentence level.

3.3.2 Read speech tasks

The speakers were also asked to perform oral readings of two texts, one in AA and another one in French, at three different speech rates (slow, normal, fast) which correspond to a hyper-articulation, fluent articulation and speed articulation of pronunciation. For French, the text was an excerpt from "Le Petit Prince" ('The little Prince' by De St Exupéry (1943)). For Algerian Arabic, we used an excerpt from an Algerian movie scenario "Bab El-Oued City" Allouache (1994) transcribed in Arabic letters. The mean duration of a read session is

∼6mn. The controlled read speech recordings are summarized in Table 3.3.2.

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

Table 3.3: Read speech in FR and AA. Number of words including repetitions and average reading times in seconds (slow-medium-fast)

The first goal of the read-speech recordings was to obtain a controlled monolingual speech 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 read (controlled) speech data helps to apply acoustic analysis and serves to beacon for phonetic observations in CS spontaneous speech. The total duration of the recorded task is 2h.

3.3.3 Code-switched speech

The CS data is the main part of the corpus, 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 as shown in Section 3.3.1. The principal questions were inspired by the speakers' 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 (see Appendix A.2). 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.

In these stimuli, the linguist tries to implicitly suggest the use of AA and French in a same conversation to obtain CS spontaneous speech. The recordings of this task lasted from 18 to

25 minutes for each speaker. The total duration of the CS recordings is 5h30, they represent 70% of FACST corpus data.

The details about the CS speech data productions, the volume of speech and the frequency of the CS speech in FACST corpus are detailed in Section 4.

3.4 Speech segmentation and language annotation

In the following, we present the major speech processing steps, prior to manual orthographic transcription. The steps are: CS speech segmentation and languages annotation.

3.4.1 Oral speech segmentation

Oral CS data processing requires an organization that takes into account the oral speech particularities and language change within the speech. In this step, speech segmentation concerns word groups like oral phrases and word groups delimited by a language switch. The segmentation is done manually and at three levels. The first one is based on speech turns, the segmentation is carried out at each speakers change. Thereafter, the oral speech is segmented by breath groups or group of an utterance or an oral sentence that is marked by long pauses and formed meaning items with a minimum of syntactical items. Finally, the segmentation follows the language change in the code-switched utterances. So, a new segment starts at the language change and every segment contains one language. In order to distinguish between the phonetic segments (segmentation at the phoneme level, Section 3.7) and the speech segmentation, we called these segments "Speech stretch". Figure 3.2 shows that different levels of the oral speech this speech segmentation: speakers turn, breath pauses or clean "sentence" ending and language change (in CS sentence).
The Speech stretch can be composed by:

- Sentences
- Part of sentences
- Particles/grammatical items

- Word insertions

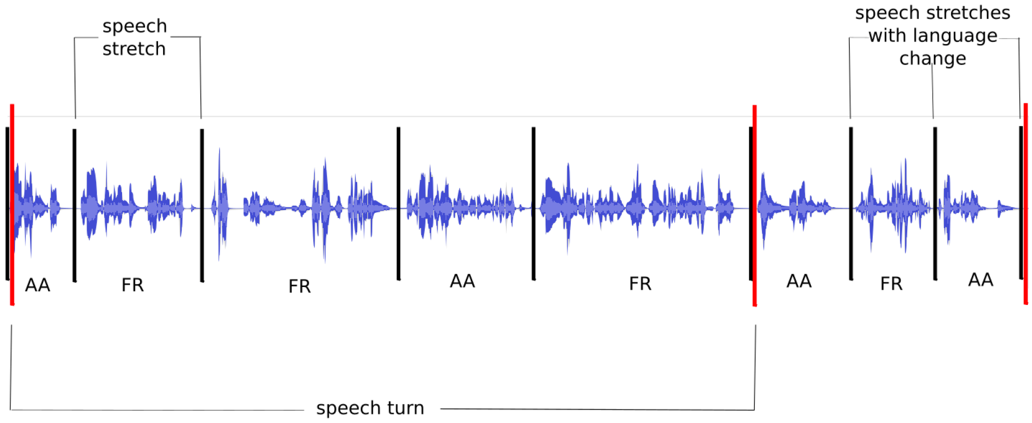


Figure 3.2: Speech segmentation of FACST signal audio. The speech stretch corresponds to the breath or rhythmic group (oral phrase). The speech stretch is also delimited by the language change. The "FR" and "AA" notations correspond to the language of the segment. The speech turns correspond to single speaker speech

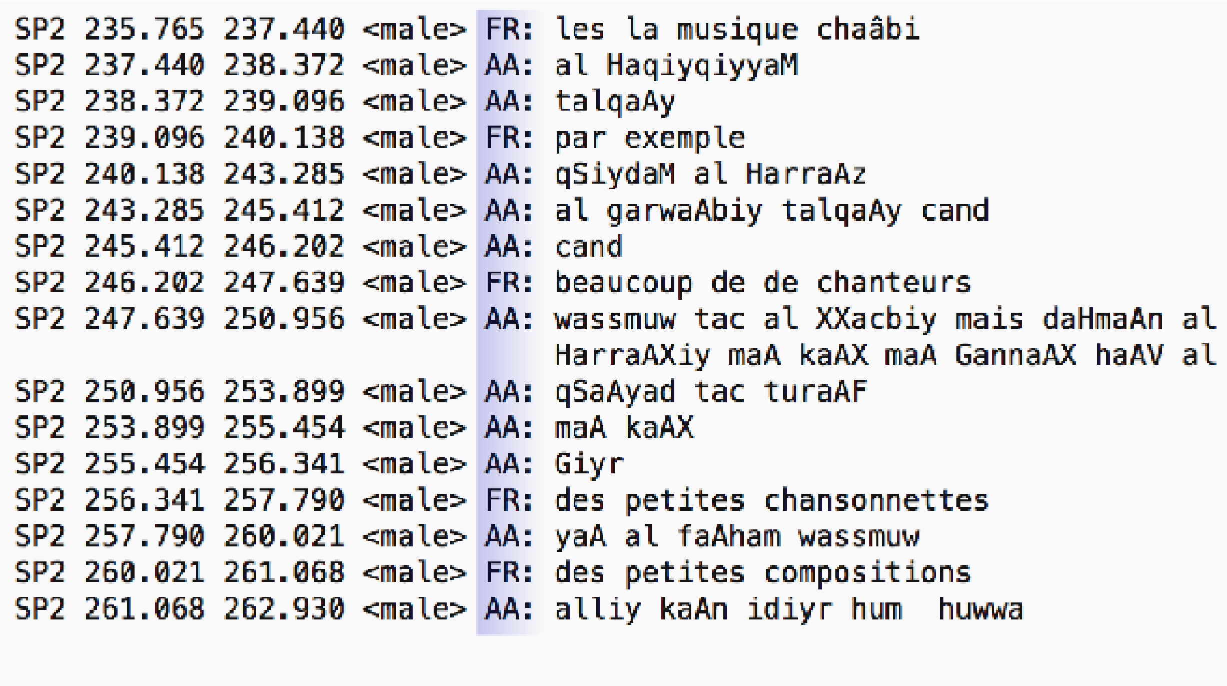
The length of the Speech stretch is quite variable, ranging from very short segments (less than 1s) to longer ones (6s), with an average segment length of 4s. The short stretches corresponds to short words insertions of L*b* in L*a* in CS utterances such as an article or a particle of embedded in phrase. The aim of this segmentation is first to get segmented speech units easy to handle for automatic processing and linguistic analysis. Also, this segmentation allows to get boundaries for each language and label them "AA" for Algerian Arabic segments and "FR" for French segments (See Section 3.4.2). This type of segmentation and annotation helps thereafter when processing and manipulating the data and the language switches. We used Transcriber program (Barras et al., 2001) to draw and fix the time boundaries of each segment and annotate the data.

3.4.2 Annotation of code-switched speech

With the help of Transcriber program we did three types of annotations on the speech stretches. First, these segments are labelled by their language. The aim of this annotation is to manipulate the speech stretches and especially processing the language switches.

The speakers are also annotated by speaker's number or speaker's name and the gender. Figure 3.3 shows an example of speech segmentation with the segment-level and their annotations.

It should be noted that the read speech has only monolingual segments. So, languages annotation is not present in this part of the corpus.



```
SP2 235.765 237.440 <male> FR: les la musique chaâbi
SP2 237.440 238.372 <male> AA: al HaqiyqiyyaM
SP2 238.372 239.096 <male> AA: talqaAy
SP2 239.096 240.138 <male> FR: par exemple
SP2 240.138 243.285 <male> AA: qSiydaM al HarraAz
SP2 243.285 245.412 <male> AA: al garwaAbiy talqaAy cand
SP2 245.412 246.202 <male> AA: cand
SP2 246.202 247.639 <male> FR: beaucoup de de chanteurs
SP2 247.639 250.956 <male> AA: wassmuw tac al XXacbiy mais daHmaAn al
                               HarraAXiy maA kaAX maA GannaAX haAV al
SP2 250.956 253.899 <male> AA: qSaAyad tac turaAF
SP2 253.899 255.454 <male> AA: maA kaAX
SP2 255.454 256.341 <male> AA: Giyr
SP2 256.341 257.790 <male> FR: des petites chansonnettes
SP2 257.790 260.021 <male> AA: yaA al faAham wassmuw
SP2 260.021 261.068 <male> FR: des petites compositions
SP2 261.068 262.930 <male> AA: alliy kaAn idiyr hum  huwwa
```

Figure 3.3: Example of segments annotation by speaker code (SPx), time-codes in seconds (columns 2-3), gender, language code highlighted in blue (FR/AA) and transcriptions.

3.5 Transcription of FACST data

Manual transcription has been used a lot for automatic speech processing, even more for low resourced languages. The transcription of written resourced language like English and French is more stable than the low written resources language like Arabic. Indeed, AA presents different forms of spelling following the pronunciation variation and it requires special attention. In this section, we detail the transcription of FR and AA in FACST corpus.

The transcription conventions vary according to the speech style (journalistic speech, read speech, informal conversations, children speech ...), they also vary depending on the goals of the study and corpus analysis (phonetics, syntax, discourse analysis, language disorders ...). To study our CS spontaneous speech, we choose one principal convention of transcription and two different transcription manners. We used One for French speech, another for Algerian Arabic speech. The following subsections describe the conventions we used and the specificities of each language.

3.5.1 Transcription of French

The French speech stretches were manually transcribed using an orthographic transcription. To take into account the major speech phenomena which are related to spontaneous speech in general and to oral French in particular, we adopted the following conventions. As a general principle, every produced sound is transcribed, either by an existing word or by a specific label indicating speech noise for example. With respect to disfluencies, at the word level, repetitions are explicitly transcribed, like "le le ..." *the the ...* or oral auto-corrections or restarts. The unfinished words are transcribed as they are pronounced, for example: "les remerci-" for "les remerciements" *acknowledgement*. The Word truncations (apocope) is very recurrent in French oral speech. Apocopes are included in dictionaries as a possible forms of the word. They are written as they are pronounced, for example: "manif" for "manifestation" *the demonstration*, "prof" to mean "professeur" *teacher*.
Spoken language is characterized by a syntax which is somewhat different from written language (Cooper and Paccia-Cooper, 1980; Zribi-Hertz, 1988; Kurdi, 2016). In oral French, some syntactic elements may be omitted, for example: the negative form of the verbs with-

out "ne" are transcribed as they are produced. "je sais pas" is an oral form of the written one "je **ne** sais pas" *I don't know*. Besides, much of interrogative forms are marked just by a rise in intonation like: "tu vas au lycée?" instead of "vas-tu au lycée" or "Est-que tu vas au lycée?" *Do you go to high school?* So, besides dots and commas, question and exclamation marks are explicitly written, in order to make the transcription easy to read.

In addition to the main speaker's production transcriptions, backchannel annotations are marked. Hesitations, interjections and onomatopeia like *euh, ah...* are annotated.

Phonetic phenomena giving rise to specific pronunciation variants are not marked in the transcription, like annotation of the lengthening of the vowels and speech reduction (like schwa or consonant deletion...).

3.5.2 Transcription of Algerian Arabic

3.5.2.1 Method

Algerian Arabic speech presents a number of challenges to manual transcription. First of all, it is an oral language and there is no established tradition to write in AA. When writing down the content of oral productions, we cannot rely on AA-specific standards. In Algeria, public written documents tend to make use of Arabic MSA. Personal written communications exist, however they are very tolerant to variation as the written forms may be either close to MSA or to the actual oral production in AA.

Next, the Arabic script presents specificities which are not easy for automatic speech processing in general, and phonetic segmentation in particular. Important graphic and phonetic information is given by diacritics (vowels and consonant gemination) which are not necessarily consistently reproduced in the transcriptions. Words are not necessarily separated by blanks: the syntax in the Arabic script requires attached characters between two grammatical units, for example: the article with the noun, the particle with a substantive, attached possessives. Moreover, the Arabic script is written from right to left.

To transcribe AA speech, we did not use the Arabic script, but a transliterated script was used. It is inspired by Buckwalter Arabic transliteration (BKW) (Buckwalter, 2002) and modified according to AA phonetic and syntactical specificities.

The first aim of the transliteration is to get scripts with the same characters in both languages and get transcriptions which are written in the same direction. Also, 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. The Table 3.4 in the Section 3.5.2.2 illustrates the characters chosen for FACST and the corresponding symbols for each characters in Arabic script, BKW convention and IPA symbols.

As French speech transcription, the AA one also includes pauses, repetitions, hesitations, speech backchannels and various linguistic disfluencies. In Arabic script, the articles and some particles are attached to the word: اَلبَاب (بَاب + **اَل**) *the door*, لِلبَاب (بَاب + اَل + **لِ**) *for/to the door*. For these words, 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 for two reasons. First it was applied in order to process easily the AA units and its apparition in CS speech. Indeed, as shown in Chapter 2, the CS in French and Algerian Arabic can be produced only with particles and articles of AA: "je coupe **بِ** les ciseaux" *I cut **with** the scissors*. Also, this separation helps to count the number of words and to compare speech production at word level in CS FR-AA. So, in FACST transcription the word اَلبَاب is transcribed "al baAb" *the door*, the utterance لِلبَاب is transcribed "li al baAb" *for/to the door*. We used this method 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, example: attached objects "jabt**hum**" *I brought **them** back.*

3.5.2.2 Convention of transcription

Because of the low written resources of AA and due to the low literature about its oral speech, we tempt in this section to describe with more details the oral production of the AA phonemes and their transcriptions in Arabic letters. We link this description to the convention of transcription proposed to transliterate the AA part of FACST corpus.

Table 3.4 shows the consonants of AA in IPA and illustrates the characters chosen for FACST and the corresponding symbols for each character in Arabic letters, BKW convention and IPA symbols transcription in FACST corpus. The Abjad symbols correspond to the Arabic letters. The table also shows the corresponding BKW symbols and it includes the AA written and pronunciation particularities in the comments column. Table 3.5 shows the transcription convention of the short and long vowels of AA including the orthographic forms of the vowel /a:/. Simplified conventions of transcription with annotations for both AA and French are in Appendix B.

This transcription convention has been created in order to facilitate the use of the manual transcriptions and realize phonetic analysis and studies. Also, this convention facilitates the transcription of Arabic vowels in letters instead of the diacritics ■_ ■َ_ ■ُ_ ■ِ_ used in Arabic script and then it allows an automatic alignment of the transliteration with the audio signal (See Section 3.7). However, this convention was made to keep at the same time the possibility to convert the transliterated characters to Arabic characters in future studies, and it also gives the choice of readability of the AA transcriptions in both characters. Figures B.1 and B.2 in Appendix show an example of CS transcription with two versions. The first one with transliterated AA and the second one with an automatic converted characters of AA in Arabic characters.

	IPA	FACST symbol	Abjad symbol	BKW symbol	Examples Abjad-FACST	Trans-lation	Comments
Plosives	p	p	پ	—	پلَاصَا plaSaA	place	foreign origin
	b	b	ب	b	بَاطل bATal	free	
	t	t	ت	t	تَبَّع tabbac	follow	
	t	M	ة	p	بَڨرَة bagraM	cow	Ortho form of /t/ at the end of words

Continued on next page

	IPA	FACST symbol	Abjad symbol	BKW symbol	Examples Abjad-FACST	Trans-lation	Comments
	tˤ	T	ط	T	طِبِيب Tbiyb	doctor	
	d	d	د	d	دَبزَة dabzaM	punch	
	dˤ/ðˤ	D	ض	D	ضَحكَة DaHkaM	smile	/dˤ,ðˤ/ no phonological distinction
	k	k	ك	k	كتَاب ktaAb	book	
	g	g	گ	g	گَاع gaAc	all	
	q/g	q	ق	q	قَرعة qarcaM	bottle	/g/ can be a free variant of /q/
	ʔ	E	أ	>	أمَل Emal	hope	
Affri-cates	dʒ/ʒ	j	ج	j	جُوع juwc	hunger	
Nasals	m	m	م	m	مَاضِي 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	/t/ can be a free variant of /θ/
	ð	V	ذ	*	هَذَا haVaA	this	/d/ can be a free variant of /d/
	s	s	س	s	سفَر safar	travel	
	sˤ	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	

Continued on next page

	IPA	FACST symbol	Abjad symbol	BKW symbol	Examples Abjad-FACST	Trans-lation	Comments
	ɣ,ʁ	G	غ	g	غرِيب Griyb	foreign	
	ħ	H	ح	H	حيَاة HyaAT	life	
	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	
Appro-ximants	w	w	و	w	وَردَة wardaM	rose	
	j	y	ي	y	يَد yad	hand	
Gemi-nates			■ّ		سَنَّة sannaM	tooth	Doubling of of consonant
Tan-ween	an	aN	اً	F			Syntax declination

Table 3.4: Pronunciation and transcription of AA consonants with word examples and notes about usage and particularities

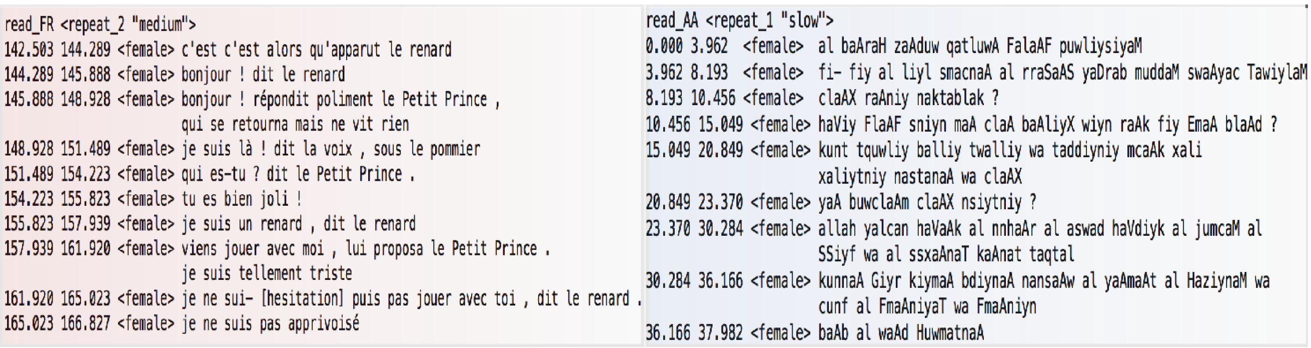
3.5.3 Read speech transcription

Read speech transcriptions are mainly based on the chosen texts. However, in this re-transcription, repetitions, hesitations, omissions and all the elements of the convention of transcription are added in the final transcriptions. For French, the same convention and orthographic transcription of the French part of CS speech has been applied (See Section 3.5.1 and Appendix B). For AA, the original read texts are written in Arabic letters and the read speech was transliterated in Arabic FACST symbols (See Section 3.5.2.2 and Appendix B). The three repetitions with different speech rates are managed and labelled with Transcriber. The speech stretches correspond mostly at prosodic groups that vary with the three speed of the reading speech. Figure 3.4 shows examples of transcription of read speech excerpts in

	IPA	FACST symbol	Abjad symbol	BKW symbol	Examples	Translation	Comments
vowels	i	i	اِ	i	ضِيَّع Diyyac	lose	
	iː	iy	اِي	iy	خشِين xXiyn	bold	
	u	u	اُ	u	جُملَة jumlaM	sentence	
	uː	uw	اُو	uw	غُول Guwl	monster	
	a	a	اَ	a	فَرجَة farjaM	show	
	aː	aA	اَا	aA	مَاكلَة MaAklaM	food	
		Y	ى	Y	علَى claY	above	Orthographic form of /a:/
	ə						Absence of vowel or schwa

Table 3.5: Pronunciation and transcription of AA consonants with word examples and comments about usage and particularities

FR and AA.



```
read_FR <repeat_2 "medium">
142.503 144.289 <female> c'est c'est alors qu'apparut le renard
144.289 145.888 <female> bonjour ! dit le renard
145.888 148.928 <female> bonjour ! répondit poliment le Petit Prince ,
                         qui se retourna mais ne vit rien
148.928 151.489 <female> je suis là ! dit la voix , sous le pommier
151.489 154.223 <female> qui es-tu ? dit le Petit Prince .
154.223 155.823 <female> tu es bien joli !
155.823 157.939 <female> je suis un renard , dit le renard
157.939 161.920 <female> viens jouer avec moi , lui proposa le Petit Prince .
                         je suis tellement triste
161.920 165.023 <female> je ne sui- [hesitation] puis pas jouer avec toi , dit le renard .
165.023 166.827 <female> je ne suis pas apprivoisé
```

```
read_AA <repeat_1 "slow">
0.000 3.962  <female>  al baAraH zaAduw qatluwA FalaAF puwliysiyaM
3.962 8.193  <female>  fi- fiy al liyl smacnaA al rraSaAS yaDrab muddaM swaAyac TawiylaM
8.193 10.456 <female>  claAX raAniy naktablak ?
10.456 15.049 <female> haViy FlaAF sniyn maA claA baAliyX wiyn raAk fiy EmaA blaAd ?
15.049 20.849 <female> kunt tquwliy balliy twalliy wa taddiyniy mcaAk xali
                       xaliytniy nastanaA wa claAX
20.849 23.370 <female> yaA buwclaAm claAX nsiytniy ?
23.370 30.284 <female> allah yalcan haVaAk al nnhaAr al aswad haVdiyk al jumcaM al
                       SSiyf wa al ssxaAnaT kaAnat taqtal
30.284 36.166 <female> kunnaA Giyr kiymaA bdiynaA nansaAw al yaAmaAt al HaziynaM wa
                       cunf al FmaAniyaT wa FmaAniyn
36.166 37.982 <female> baAb al waAd HuwmatnaA
```

Figure 3.4: Read speech transcription texts in FR (highlighted in red) and in AA (highlighted in blue). The transcription includes word repetitions, hesitations. The headers of the texts show the language, the number of the repetition and the rate of the read texts

3.6 Remarks on Code-switching annotation and transcription

The transcription of the spontaneous speech presents several hesitations and difficulties. The first that we cite is also the first that any transcriber meets, it is the perception of the audio

speech. Between perception hesitation and "wrong perception", the transcriber needs mostly more than one listening of a Speech stretch or words to transcribe them. So, the replay of speech segments is necessary for an accurate transcription, and it means that it takes a lot of time to transcribe dozens of hours of speech. The challenges of spontaneous speech transcription is not only to perceive of the pronounced speech, but also to realize the speech reductions and omissions in order to understand the statement and transcribe it correctly.
CS speech adds more transcription and annotation difficulties to the language perception, especially in the homophones units and the common words. One of the major difficulties of language annotation in CS FR-AA is articles **al** /l/ in AA and **l'** /l/ in French at the switches. Both articles are pronounced identically but it is very difficult to identify the language of this word at the switch moment. Examples: (fiy **l'**école fiy-AA, **l'**-FR, école-FR) or (fiy **al** école fiy-AA, **al**-AA, école-FR).
Besides, the CS FR-AA is characterized by the Bilingual words. CS verbs can take a base in one language but are redesigned with the other language form. An illustration is the following CS example of AA-FR: `**y**partaAj**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 the segmentation of spontaneous speech stretches of FACST, some segments are very short (less than 1 second) because of the speed of the language change. So it is difficult to segment them manually and align them automatically. These segments generally correspond to particles and articles of both languages. The most frequent units that could be noted are the AA particle "fiy" "فِي" [fi:], [fi], [f] (with a very reduced pronunciation) *in, at* in French utterance and the French conjunction "donc" [dɔk], [dɔ], *so* in AA utterances. Figure 3.5 illustrates an example of CS with the duration of the AA particle "fiy" embedded in French utterance.

Another difficulty noted in FACST transcriptions is the orthographic forms of words. In fact, French language contains a lot of homonyms like: "ces, ses" *these, their*. Also, the plural mark "s" at the end of words is not defined in the oral speech and words are pronounced

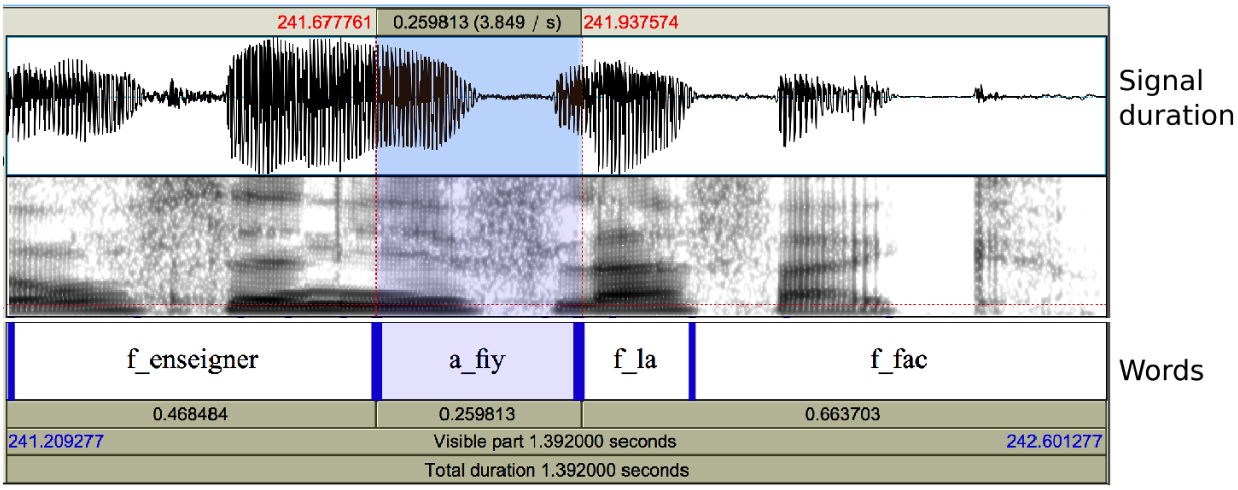


Figure 3.5: Example of the particle duration **fiy** *in* (0.25s) in CS utterance. The words in FR are preceded by "f_" and the words in AA are preceded by "a_"

identically in some cases like "d"étude" and "d'études" *of studies, of a study*. Concerning the Arabic speech, the transcription of an oral language with low written resources leads to hesitation in orthographic form transcription. Indeed, multiple studies show that written outlines of AA in social media note different forms of spelling. Theses differences are due to the non standard orthography and the pronunciation variation in different regions of AA (Saadane and Habash, 2015; Abidi and Smaïli, 2018). Besides, referring to the MSA to transcribe AA speech gives a strong guidelines. However, while AA and MSA sharing a significant part of words together, they are pronounced differently and then the transcription requires a particular attention at phonetic levels. For example: the transcription of the long vowel in the AA word [maʕ**a:**h] "معاه" *with him* which does not exist in MSA pronunciation and orthography [maʕ**ah**] "معه".

3.7 Automatic alignment of code-switched speech

When it comes to two languages in one continuous acoustic signal, the methods are adapted to process two or more languages and taking into account the switches (Lyu and Lyu, 2008; White et al., 2008; Lyu et al., 2015). In order to obtain boundaries and labels of the words and phones of CS, we combined two alignments and we used different ASR systems in parallel. A first alignment for French Speech stretch and a second alignment for AA. The following subsections, 3.7.1 and 3.7.2, present the system and the language models used to the Forced alignment of FACST data for each language, and the results of the segmentations.

3.7.1 Forced Alignment of French

The French Speech stretch, labelled "FR" (See 3.4.2), has been aligned using LIMSI speech recognizer in a forced alignment mode (Gauvain et al., 2003; Laurent et al., 2016), assigning word and phone level time codes. Speech files were automatically segmented into acoustically homogeneous segments, which ideally corresponds to speaker turns and/or to a given language or stable acoustic conditions (broad band/telephone band...). This mode of ASR system uses an acoustic model and a dictionary of pronunciation that make possible to match the audio signal with the orthographic transcription at phone and word levels. The pronunciation dictionary of standard French is used for this alignment. It contains canonical pronunciations and pronunciations with variants, like speech reductions, realization or deletion of the schwa (Adda-Decker and Lamel, 1999; Adda-Decker et al., 1999a; Adda-Decker and Lamel, 2017). So, the French alignment gives lexicon pronunciation of the transcribed speech (canonical pronunciation) and can allow the pronunciation variants if they exist in the pronunciation dictionary and if they are recognized into the speech signal. Example: the lexicon "autre" *other*, following the speech style, has three pronunciation variants [ctʁə], [ctʁ], [ct]. This identification of word pronunciations is particularly useful for phonetic and phonological studies and it gives a more precise vision about the natural speech production (Adda-Decker, 2006). The language model for the forced alignment is based on the given French manual transcriptions.

3.7.2 Forced Alignment of Algerian Arabic

3.7.2.1 Acoustic models

The LIMSI Arabic ASR system used to align this part of CS speech was trained on several hundreds of hours of Arabic speech, predominantly Lebanese, MSA and an Algerian Arabic dialect from a large number of speakers (Lamel et al., 2008, 2009). The Algerian models are the result of adapting the multi-dialect Arabic system with about 300 hours of data from Algerian speakers. The AA dialect is a consonantal language with a greater number than the MSA (30 consonants, see the phonemic inventory of AA Section 2.3). So, the system was adapted for the 3 foreign consonants /p,v,g/ which are frequently used in the AA speech. Then, the ASR system used a set of 37 phones and processed the 39 symbols (Illustrated in Section 3.5.2.2 and resumed in Appendix B.3). It includes all emphatic consonants of MSA (27 consonants), the glottal stop "Hamza" and 3 foreign consonants (/p,v,g/). The system also used the 6 short and long vowels of Arabic /i, iː, a, aː, u, uː/. The AA vowels /ɛ, ɛː/ described in the phonemic inventory of the previous chapter are not taken into account in this alignment because of limited resources (orthographic transcription of only 6 vowels, requirements of monolingual data to study in detail the vowels, need to train the system with more data ...). However, a dictionary of pronunciation (See Section 3.7.2.2) was designed to allow the pronunciation variants of the minimal pairs and other words with /a, i/ and /ɛ/ or /aː, iː/ and /ɛː/ noted in FACST data. Example: the entries دَار and دَار have a same orthography but pronounced differently and have distinct meanings, /daːr/ *house* and /dɛːr/ *he/she did*. Allowing this variation gives only the multiple pronunciations of the word but they are associated to one written form and we should be reminded that the FA can be realized with or without this pronunciation variants. In short, the question of these vowels would be one of our future filed of investigation in AA Arabic automatic processing.

A geminate version for each consonant is included in the system. A set of geminate symbols was designed to correspond to the double consonants into the signal and in the transcriptions.

3.7.2.2 Pronunciation dictionary of Algerian Arabic

To the grapheme-phoneme conversion, a pronunciation dictionary was designed specifically for Algerian Arabic pronunciation. It lists phonemic transcription of all the words of FACST by adding the various pronunciations for a large number of the most frequent words.

The relationship between the word spelling and the sound in Algerian Arabic is highly corresponding. The AA is similar to the MSA and other languages like Spanish and Finnish at this level. The oral characteristics of this dialect make it possible to have a very simplified orthography and then the mapping of the letters and sounds does not present a large number of exceptions and irregularities. Hence, the dictionary was written automatically by mapping one letter-to-one phone. The conversion also allows to the different orthography forms of one letter that represents one sound to be converted in one phoneme such as "al hamza" أ ـِءً, ءو, ءي /ʔ/. However, some exceptions require to convert character sequence in one phone. Indeed, in order to distinguish between the long vowels /uː, iː/ and the glides /w, j/ which are written with the same letters ي (y), و (w), the conversion will be done with the sequences (uw, iy, aA) for the sounds /uː, iː, aː/ and the rest of the tokens was processed as one letter-to-one phone. It should be noted that a large part of end of word letters are not pronounced and then not mapped like أ "A" and ة "M". The exceptions of some words are processed manually like the words قُضَاة ، حَيَاة /ħajaːt, qudˤaːt/ *life, judges*. Also, since the declinations are not pronounced in AA, they are not transcribed and not mapped for FACST data. However some words still keep "Tanween" in pronunciation (See Section 3.5.2.2) and mapped on the transcription of it, like the words, which are frequent in the corpus, شُكرًا *thank you*, أصلًا *basically*.

AA also contains speech reduction and other phenomena like lenition and fortition, reductions of the vowel duration, the deletion or pronunciation of schwa. These phenomena change the pronunciation of the words and their new forms are manually processed in a second time within a dictionary behind the canonical pronunciation (Adda-Decker et al., 1999b)

3.7.3 Combined forced alignment for code-switched speech

As described in Section 3.7, the forced alignment of CS speech is the result of a combination of two alignments of each language. Figure 3.6 resumes the use of the LIMSI ASR systems to align each language speech data. The alignments also generate speech segmentation at word and phone levels. Figure 3.7 shows an example of CS sequence[3] alignment displayed with Praat program. Three segmentation levels are applied in the alignment outputs: segmentation at the language level, segmentation at word level and segmentation at the phone

[3] The English translation of CS utterance: "même si talqaAyah très fort" even you **find him** pretty good. The text in bold corresponds to the AA item.

level.

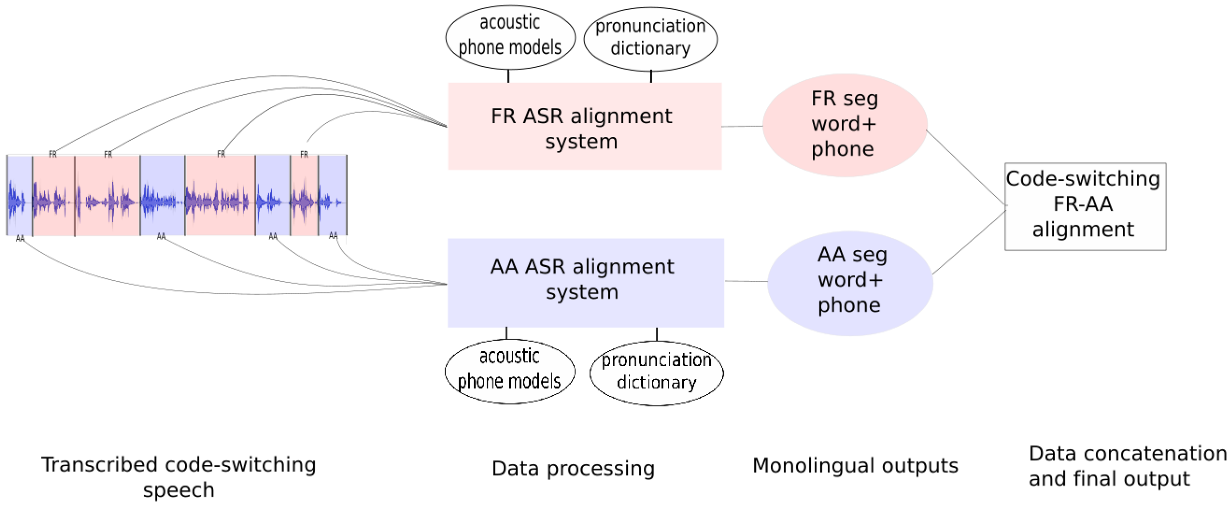


Figure 3.6: The automatic speech alignment of code-switched speech using two ASR systems, one for FR and one for AA. The monolingual alignments (phone and word segmentations) are thereafter concatenated to reconstruct final alignment for CS speech

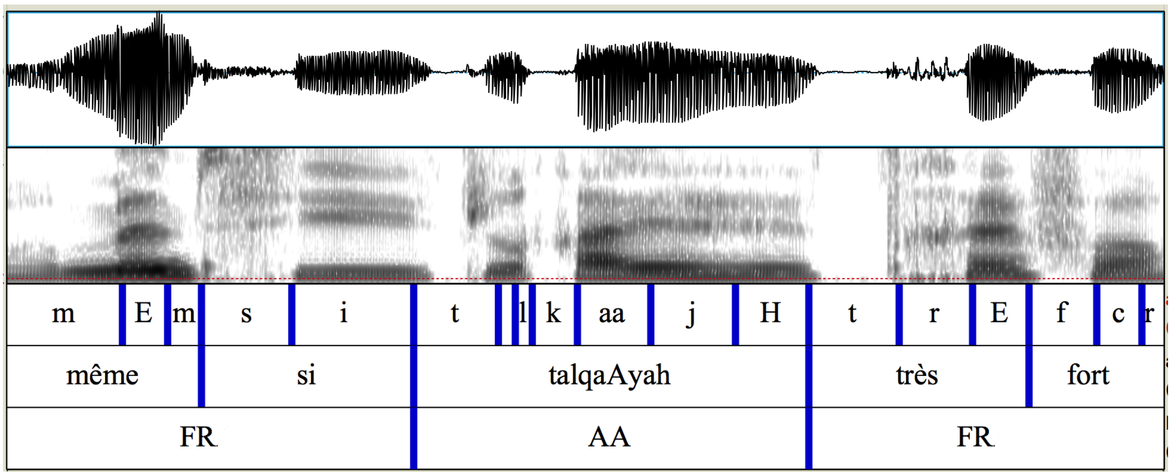


Figure 3.7: Example of FACST alignment on intra-sentential CS audio segment with spectrogram. Three levels of segmentation and transcription are applied. From top to bottom of tiers: phoneme segmentation with phonetic transcription (SAMPA convention), word segmentation with orthographic transcription, language segmentation and annotation.

3.8 General discussion

In this chapter, we presented the steps of the FACST corpus design and we defined methods to record, process, organize and align the CS data.

Even if CS seems to be produced spontaneously in French-Algerian Arabic bilinguals, CS data construction is a complex procedure. CS speech can be collected only if there is a favorable setting. There are two important conditions, that of the speakers and the conversational context. The first one is the profile of the speakers, both bilingual speakers must be able to accept producing CS speech, and the CS must be triggered by one of the speakers. The linguist's role is thus to conduct a conversation that allows the speakers to produce CS. Therefore, the will of the speaker is necessary to obtain spontaneous CS data. The second one is the communicative situation, besides the speakers will, the topic of the discussion plays an important role in CS spontaneous productions. In fact, topics such as professional life, studies, work in both countries helps the speaker to use both languages and thus produce CS speech. It should be noted that read data was recorded to have the opportunity to compare between monolingual productions and CS productions.
In this chapter, a special processing and resourcing of AA speech needed to construct this corpus was described. A transliterated transcription convention was designed to facilitate the automatic processing of CS speech data, in fact, AA isgenerally written from right to left in contrast to FR, and is written with the MSA alphabet which is tricky in computational processing.

CS data processing and CS corpus building relies on the language annotation method and the determination of language boundaries in order to process language switches either syntactically or phonetically. This kind of difficulty cannot be encountered with monolingual corpora. Speech stretch segmentation in FR-AA CS is a task that needs a particular attention, in fact, CS can be produced on equivocal units, example: language segmentation of shared words that are identically produced "stade" /stad/ *stadium*. It can also be difficult to put language boundaries with homophones that have a similar grammatical function such

as article "al" "l’" /l/ *the*.
Automatic alignment of CS data also rely on language segmentation and labeling. In fact, CS data alignment has been processed with two combined alignment. The first one is FR forced alignment and the second one is AA forced alignment.

To conclude, the FACST corpus and its related data (questionnaire, phone and word automatic alignments, transcribed lecture speech) were designed for multipurpose studies and can be used in other research fields such as automatic speech processing, phonetics, phonology, general linguistics and sociolinguistics. This corpus can also be augmented with more annotations such as CS sentence type annotation (intra and extra sentential), AA central vowel /e:/ annotation and language accent evaluation tests.

Chapter 4

Quantitative review of FACST data

Chapter contents

QUANTITATIVE review of what a corpus of CS can contain at linguistic levels and what is the distribution of the consonants and the vowel in the speech will be detailed in this chapter. The aim of this review is to give quantity observations about CS productions. This review also helps to adapt the analysis measures to the differences and similarities between the languages quantities and compare between the languages. Thus, we show first the types and the frequency of the CS speech utterances in the corpus. Furthermore, we present the distribution of the words in the CS speech in terms of quantities and word occurrences. This chapter also presents a phonetic review of CS speech with counting consonants and vowels and their distribution in the speech. Finally, speech technologies experiment for Code-Switching utterances detection are applied on FACST CS data.

4.1 Code-switching speech utterances

One of the first questions that comes when getting CS data is about the construction and the forms of the CS speech. So, to analyze the Speech stretch in the CS speech we answer the following questions: what are the quantities of language switches in the data? What are the proportions of both languages in CS speech? How these language stretches are distributed in the speech and what is their duration?

First, we counted the Speech stretch of the CS speech part of the corpus independently from their syntactical position in the speech (in the beginning of the sentence, at the end of the sentence, inside L*a* or inside L*b* sentences). The aim of this work is to know what languages represent in the CS speech in terms of quantity and know what is the language intervention in the corpus. We recall that these Speech stretch are delimited by breath groups and they can be delimited too by language switches as described in the Section 3.4.1. Thus,

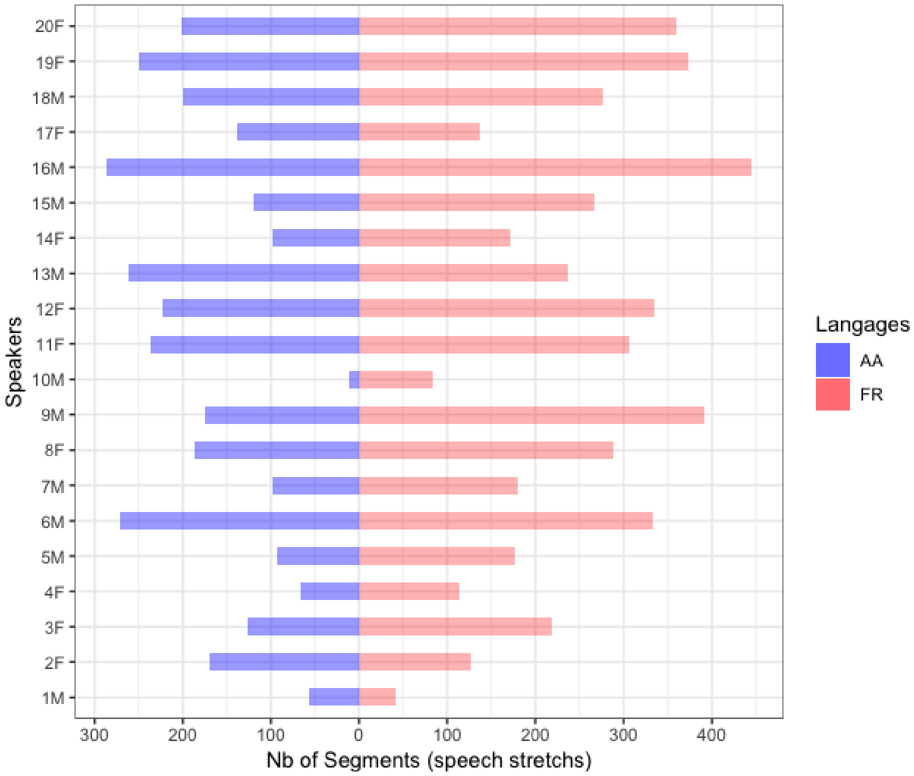


Figure 4.1: Number of segments (Speech stretch) in FR (in red colour) and AA (in blue colour) for 20 CS speaker of FACST (in the *y* scale). The labels "M" and "F" correspond to the genders Male and Female

the number of segments represents the number of acts of languages with the CS speakers. The figure 4.1 shows the number of Speech stretch annotated by language in the CS speech, these segments are classified by language (AA in blue and FR in red). We notice first that while the variation of the number of segments for each speaker, the French segments are higher than the AA ones with most speakers with a total of 49k segments of FR behind 33k of AA. Then, the average of segments for each speaker is 2.4k for FR and 1.6k for AA in ∼18m of the main duration of the CS speech. So, this language distribution explains that both languages are highly present in the CS speech production, however the numbers indicate clearly that the FR Speech stretch dominates the CS data.

The number of segments gives the quantities of languages in CS speech but counting them only is not sufficient to get a precise view of the speech. So, in order to get more information about the use of the languages in the corpus, we counted the total of the Speech stretch and we classified them by their duration. We also complete the study relating the number of words with the segments and their duration in Section 4.3.

The figure 4.2 shows smoothed curves that represent the number of segments and their duration in both FR and AA languages. We observe that the duration of the segment is related to their number for both FR and AA. Indeed, we observe a highest part of the segments is between 0.5s and 3s and the number decreases according to the longest duration. However, the major difference between FR and AA is the segment duration, the longest duration of the FR segment exceeds 10s and for AA the duration stops at 8 seconds. Besides, a considerable part of segments with 6s and 9s of FR are higher than the AA which means that rhythmic groups in FR may have a longer duration than in AA during the CS speech. To sum the count of the figure, we can divide the duration of the segments in three parts. The first part contains the shortest but the most numerous segments for the two languages (between 0s and 3s). This part represents more than 50% of the segments. The second one contains the medium-length segments (3s-6s), it is less numerous than the first part, it represents 35% of the total of speech segments. The last part groups of the longest segments (from 6s at 12s) and at the same time presents the lowest quantity. The distribution of the languages takes mostly similar curves (correlation between numerous short segments and long seg-

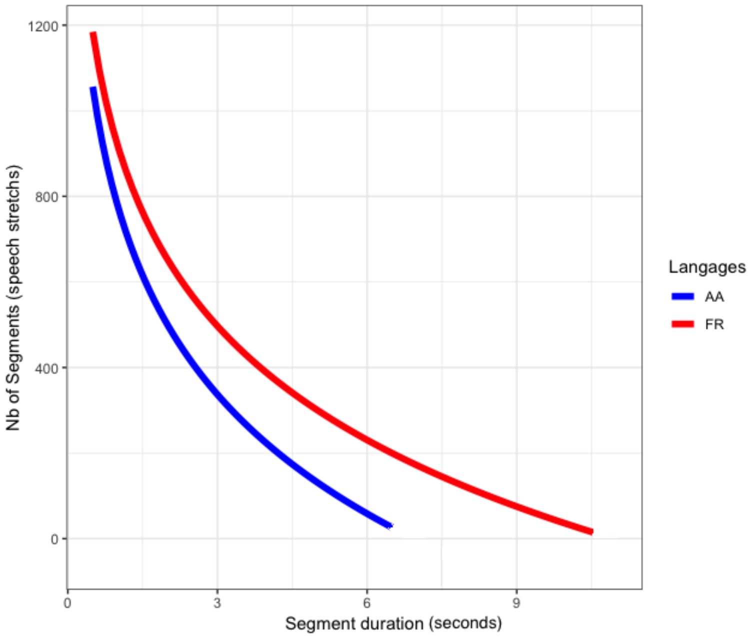


Figure 4.2: Number of segments (language stretches) and their duration (smoothed lines)

ments with low quantity) but they are not equal. Indeed, the last part of segments concerns only the FR language, AA segments do not record segments longer than 7s.
The figure supports too that the Speech stretch distribution in the CS speech varies in FR and AA and it joins the results of figure 4.1 that both of languages note a significant presence but the FR language is dominant in the CS speech. Linking the results of this Speech stretch counts with their duration highlights that the CS in FR AA is composed with a big part of very short switches (from 0.03s to 3s) and it also refers to think about investigating the intra-sentential CS in and its syntactic organization in FACST corpus.

To summarize, these figures give an overview of the Speech stretch quantities of FR and AA that compose the CS speech of FACST, the French segments are more numerous than the AA ones with all of the speakers. We conclude that the speakers use highly both languages to produce CS speech, however the French language shows more presence and dominance in these productions. The corpus FACST counts a big part of very short segments in CS, that means that the language switch is very short and the FR-AA CS is composed of very small language units.

4.2 Read speech utterances

The segments of read speech texts are monolingual and they are delimited following one major criterion: the intonation groups that joins the reading text prosody and the speed of the reading. Since the texts are read with three speech rates (slow-medium-fast) (See the sections 3.3.2 and 3.5.3), the length of the segments of one repeated text in different rates varies with the same speakers. We recall that the aim of the segmentation in Speech stretch in the read speech is to get an "oral sentence" of this speech which helps to process the speech and align the data at the word and the phone levels. It should be noted too that the selected texts in FR and AA have a similar reading duration and a similar length but they do not have exactly the same lengths. The FR text contains a few more sentences and then a few more words.

Table 4.1 shows the total number of segments in each language and their duration average. The three repetitions of FR text contain more segments and words than the AA one. The segment duration and the distribution of the number of words in a segment are slightly different in both languages. FR counts more segments than AA, however the duration of FR segments is shorter than the AA. At the same time, the average of the number of words by segment is higher in FR. Observing the number of words and the average number of words in a segment, we can say that the AA language takes more time to read with 6.7 words/segment in 3.6s than FR with 7.9 words/segment in 3.4s. This analysis leads to reflect on the question of the duration of the words and their morpho-syntactical composition in AA and FR and to take into account these characteristics in CS languages evaluation and comparison. Indeed, the Arabic languages are known to have attached possessives, personal pronouns are written within the verb, objects are attached to the verbs(See Section 3.5.2) which is not the case for FR. So, word counts are different in both languages and we need to investigate about word counts to understand more the CS speech. The next section investigates about the words in the CS speech.

Lang	Nb of segments	Average seg duration	Nb of words	Average word/segment
AA	1001	3,6s	6746	6,7
FR	1355	3,4s	10804	7,9

Table 4.1: Total counts of read segments (Speech stretch) in AA and FR languages with the average duration and the number of words with the average of the number of words by segment. The count includes the three repetitions of read texts

4.3 Words and lexicon

This section investigates about the place of the words in the CS speech in multiple aspects. First is to investigate about the word counts in FR and AA and link the observations to the phrases composition and the Speech stretch segmented during the corpus processing. Furthermore, this section gives an overview of the words occurrences in CS and the frequency usage of the words and the lexicon. This section interests finally to the construction of the CS production and the words that trigger the language change. The aim of this study is to describe the CS FR-AA at the word and sentence level and show the common regularities of the CS production with all of the speakers.

4.3.1 Word counts

With the help of the orthographic transcription, we counted all the words produced in the CS speech and we classified them by language. Figure 4.3 gives the word numbers for each language and for each speaker. The first observation that we can notice in this figure is that the number of words varies with the speakers. This variation depends on the quantity of the speech data collected for each speaker. But, despite this speaker quantitative variation, FR words are clearly high compared to the AA counts for most of the speakers except speakers 1, 2 and 12. French words in CS speech represent 65% beside 35% for AA words, so, 28k of FR words and 13k of AA words. In comparing theses words quantities with the number of segments of the figure 4.1, we notice that the quantities are higher in FR language at the words level (Speech stretch: 59% for FR and 41% for AA and words: 65% FR 35% for AA). Although counting the number of words does not represent an absolute equivalence of

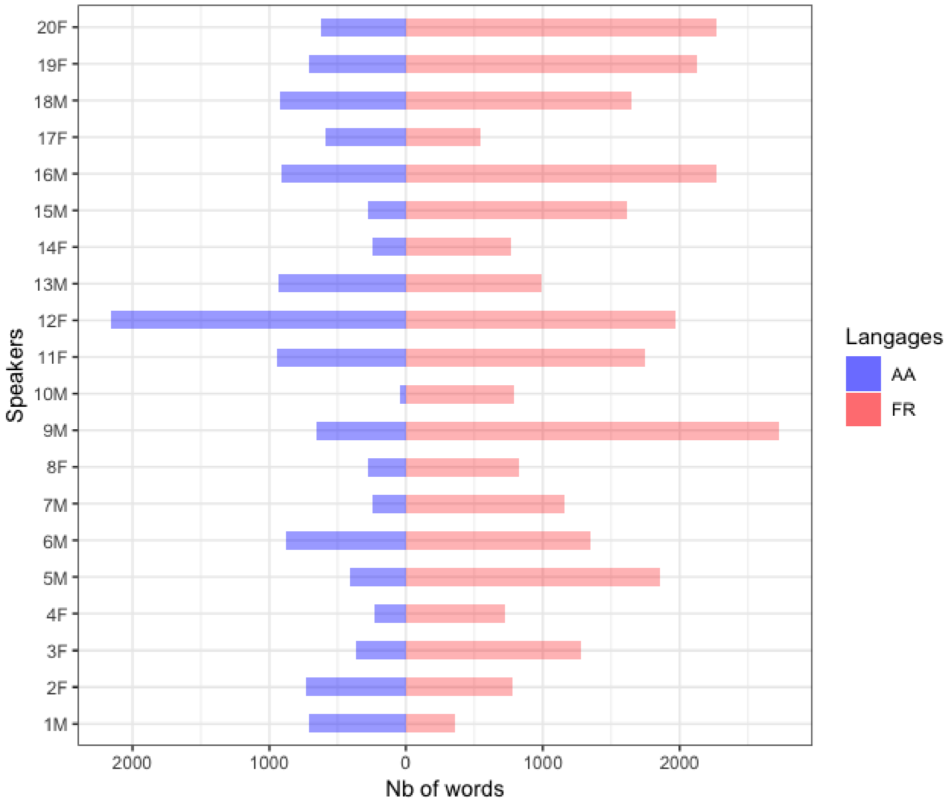


Figure 4.3: Word counts of code-switching speech of FACST corpus: number of words in FR (in red colour) and AA (in blue colour). The y scale represents the speakers and the labels "M" for male speakers and "F" for female speakers

the quantities of the two languages, we observe that the FR language is the dominant language in terms of words and segment quantities in the CS. The comparison also means that AA segments are produced with fewer words than the other language. Table 4.2 resumes statistics about the number of segments and the number of words counted in CS speech for the 20 speakers recorded. The total of words and segments represents an average of 3,9 word/Speech stretch for AA and 5,7 word/Speech stretch for FR. So, the production of words by segments in the CS is superior in the FR language.

4.3.2 Words frequency in code-switching speech

In Natural Language Processing (NLP), analyzing word frequency of a given speech allows first to know the common language use, to explore also how such a word-level features best

Spks#	FR seg	FR wrd	AA seg	AA wrd
#1 M	28	112	36	108
#2 F	126	583	170	734
#3 F	219	1283	126	369
#4 F	113	720	67	230
#5 M	175	1619	93	348
#6 M	333	1348	271	880
#7 M	180	1157	98	243
#8 F	288	829	186	277
#9 M	392	2726	174	654
#10 M	84	793	11	45
#11 F	307	1746	236	948
#12 F	334	1960	223	598
#13 M	237	986	261	928
#14 F	171	771	98	247
#15 M	266	1619	119	278
#16 M	419	2075	277	1108
#17 M	137	545	137	580
#18 M	276	1644	200	917
#19 F	372	2122	248	710
#20 F	335	1906	202	1172

Table 4.2: Number of segments and word tokens of FR and AA languages in CS speech. The words and segments are counted for each speaker. The speakers are labelled by a number and gender: "M" male, "F" female. The total average of words by segment is 5,7 for FR and 3,9 for AA

used to characterize the speech. In CS speech, the tasks of extracting and analyzing the word frequency allow to characterize the use of both languages at the same time to produce this type of bilingual speech not only at the word-level but also at the syntactic and semantic levels.

With the help of computer tools, we divided the work on CS word frequency in two parts: a general word frequency analysis in the CS speech and an analysis of the most frequent words introducing CS utterances.

4.3.3 Word frequency

The first part is to count the highest word frequency in FR and AA and classify them by Part-Of-Speech (POS): grammatical items (preposition, particles, conjunctions, pronouns), and lexical items (nouns, adjectives, verbs and adverbs). The aim of this task is to compare the use of FR and AA words to construct the CS and observe the global word choice to code-switch. We also compare the equivalences and the complementarities of this high frequency words in FR and AA in CS.

The table 4.3 presents the most word frequency of the languages. They are organized in two lists. The First column shows the highest frequencies of FR with >80 occurrences. The second column gives height frequencies that are > 40 occurrences for AA. The reason for fixing two different limit frequencies for each language is the difference in the amount of data obtained in each of these languages. AA presents less words than FR and lower word frequencies. These differences in limits allow to compare in a relevant way the use of both languages.

In these lists of words frequency, the word quantities in the most of categories are getting closer in the languages. The grammatical words are the highest portion of word occurrences with 60% of the total number of the occurrences number listed in this table. In these grammatical words, the languages share an important part of the words that express the same meaning as the definite articles, the negation with "pas" in FR and "maA" in AA, the addition and consequence express with "et" for FR, "wa" for AA, "pour" for FR and "li" for AA, and particles that express the manner, time, location, accompaniment "à, en, avec" in FR and "bi, fiy, mca" in AA. Among these lists, we also note the discourse marker "donc" in FR list but we do not notice a word equivalence in AA in this list. So, referring to these lists, the grammatical words in CS present a considerable usage in both languages in terms of quantities. The lists also show similar or equivalent of most of words in each language. So, the CS production is based globally on grammatical words alternate of both languages. The second part of the word frequency contains the lexical words. They form 40% of the total counted words in this list. As the grammatical words, FR and AA share a significant part of lexical words, like the comparison adverbs "comme" in FR and "kiyma" in AA, the

	POS	FR >80 #Occ		AA >40 #Occ	
Grammatical words 60 % of total Nb of Occ	Prep Prtic Conj Pron	de, des	pas	al	li, lak
		que	pour	wa	claY
		le, la, les	donc	fiy	min
		et	avec	maA	mca
		il		alliy	hnaA
		à		tac	mca
		un		anaA	
		ça		kiy	
		en		bi	
Lexical words 40 % of total Nb of Occ	Adv	beaucoup		kiymaA	bazzaAf
		plus		waAX	
		enfin		bark	
		non		laA	
		comme		wiyn	
		après		iyh	
		aussi		baAX	
		alors		kiyfaAX	
		juste		mbacd	
		comment		kiyf	
	Noun Adj	bien	langue	HaAjaM	waAluw
		deux	temps	carbiyyaM	kaAmal
		français	exemple	dzaAyar	XGul
		bon	tout	XwiyyaM	
		même		waAHad	
	Verb	est	avait	kunt	dart
		fait	peut	laAzam	yacniy
		était	peux	candiy	qriyt
		suis	faut	jiyt	quaAl
		sont	dire	raAniy	
		a	vais	kaAn	

Table 4.3: High word frequency list of FR and AA in CS speech. The grammatical words present ∼60% of the total number of all counted occurrences. The lexical words present ∼40% of the total counted word occurrences

quantity adverbs "beaucoup" in FR and "bezzaAf", and other adverbs like "après" in FR and "mbacd" in AA *after, later*, "comment" in FR and "waAX, kiyfaAX" *how*. For the adjectives and nouns, the equivalence in both languages is less present than the previous parts, but we note the adjectives "même" in FR and "kif, kif-kif"*same, same thing*, "tout" in FR and "kaAmal" in AA *all*. AA and the FR share the same most common verbs produced in the speech "être, avoir, faire, aller, dire, falloir" in FR and "kaAn, raAniy, candiy, daAr, qaAl, laAzam" in AA.

Discourse markers are the one of the most used words in the speech, though they are different from one language to another (Schiffrin, 2001). According to these lists of CS speech, we observe that languages have words that can be used as discourse markers. These discourse markers are different in both languages and they belong to different word categories and have different meanings. We notice that among FR discourse markers in the table, "alors, bon, enfin, donc", and only the verb "yacniy" and the noun "XGul" in the AA list.

4.3.4 Words introducing CS

An important interest was centred on the words produced at the language switch (words followed or preceded by a language switch), especially the most frequent words that introduce the switch. The goal of this work is to investigate about the most frequent words that arouse and predict a possible CS and get information about these concerned words. In order to achieve this aim, we extracted from the CS speech the most frequent words of both languages that are followed or preceded by a language switch. The table 4.4 shows the AA words followed or preceded by a FR Speech stretch and the table 4.5 shows the FR words followed or preceded by a AA Speech stretch or AA words.

In the table 4.4, AA words introducing the CS speech are grammatical words that have a function of linking between two nouns and verbs "wa" *and*, introducing consequence "li" *for/to*, expressing situation "fiy" *in*, expressing relationship between two parts "taAc" *of*, pronoun "alliy" *that*. These words have a function of linking two parts of a sentence. It is indicated that these words contribute to form an intra-sentential CS and they serve as a link to combine the two languages in one sentence.
The number of words introducing the CS represents mostly a high part compared to their total of word occurrences and indicate that the inter-sentential CS is highly produced with theses words. Indeed, except the particle "li" that represents 14% of "li" occurrences that introduce CS, we observe that the percentage of AA words introducing the CS is between

50 % and 62% of the total of occurrences.

In the table 4.5, the most frequent FR words introducing CS vary in the POS categories, we notice that the conjunctions that link two sentences or utterances "parce que" *because*, "donc" *so*, "mais" *but*. The list also contains the adverb "déjà" *already, ever, yet* and the formula "c'est" *it is* that allows to start a sentence and introduce lexical items. As AA words that introduce CS, all FR words in the table, except the formula "c'est", link two items of two different languages in the same sentence. So, these words contribute to producing the intra-sentential CS. The percentages of these words used in CS utterances are between 15 % and 50%, they are therefore used to code-switch less than the AA words in the table 4.4. "c'est" presents the lowest item used in CS context in this list with 15% of the total of word frequency. Also, by computing the number of "c'est" that appears only in AA sentence ("c'est" in intra-sentential CS), example: sentence starts with "c'est"+ AA, the number of occurrences is only 12. However, the grammatical words indicate a considerable usage to code-switch to AA with 50% and 45% for "déjà" and "parce que".

word	# total occ	#occ introducing CS	% occ introducing CS
fiy	467	271	58
wa	505	251	50
taAc/tac	158	99	62
alliy	91	54	56
li	156	23	14

Table 4.4: AA words introducing CS > 20 occurrences and their total of word frequency in the speech. The percentage column illustrates the rate of the words used to produce CS utterances (intra-sentential and inter-sentential CS)

The following excerpts of CS speech transcriptions from FACST illustrate how the most frequent words "fiy, alliy" in AA and "parce que, mais, dèjà" in FR introduce language switches and contribute to producing the CS speech.

AA words:

word	#total occ	#occ introducing CS	% occ introducing CS
c'est	763	116	15
mais	330	113	36
parce que	241	109	45
donc	316	82	26
déjà	31	15	50

Table 4.5: FR words introducing CS > 15 occurrences and their total of word frequency in the speech. The percentage column illustrates the rate of the words used to produce CS utterances (intra-sentential and inter-sentential CS)

```
      (fiy) FR: internet
            AA: wa tXuwf al fiy
            FR: les forums

 (fiy, wa)  FR: une licence
            AA: fiy al carbiyyaT wa
            FR: français, anglais

      (fiy) FR: et du coup
            AA: fiy
            FR: la faculté
            FR: il avait des notes très
            très basses
            AA:fiy
            FR: français

    (alliy) FR: le bouchon
            AA: alliy kaAyan fiy
            FR: le réservoir
```

FR words:

```
     (déjà) AA: yakuwnuw yajawzuw
            FR: déjà deux ans

     (déjà) AA: tasskun fiy cnnaAbaM
            FR: déjà
            AA: maA nacrafhaAX

(parce que) AA: nridiyji hum
            FR: parce que
            AA: maAzaAl maA bdiytX

(parce que) AA: bi al cqal
            FR: parce que
            AA: Habbiyt nsaqsiyk bark

     (mais) AA: macruwfiyn
            FR: mais
            AA: al baAqiy djaznaA cliyh

     (donc) FR: c'est un système donc
            AA: yssannaA bazzaAf min cand
```

4.4 Code-switching frequency and code-switching sentences

In a study of CS corpus, counting the switch frequency and counting the most used CS sentences by the speakers of these languages is an important step to describe the data. We therefore analyzed to determine the frequency of CS in speech and to highlight the frequent CS forms and sentences of the speakers. This section is divided into two parts, we present first an overview of the CS frequency and the number of switches in FACST. Also, a processing of the CS sentences and utterance frequency was made for the purposes of highlight the sentence forms regularities of the CS FR-AA.

4.4.1 Code-switching frequency

The segmented data speech provides Speech stretch which are labelled by language (See Section 3.4.1). The language changes in each speaker turn is considered as a CS production. So, counting the CS speech frequency includes both of intra-sentential and inter-sentential CS and words inclusion in L*b* utterance.
The CS average frequency of the continuous speech is calculated by the sum of the number of the Speech stretch language changes in each speech turn of the speaker record divided by the sum of speech turn times. The formula (1) is used to compute the CS frequency rate in the speech for all the speakers.

$$CS\ frequency\ (f) = \frac{\sum language\ switch\ in\ each\ speaker\ turn}{total\ CS\ speaker\ turn\ duration} \tag{1}$$

The table 4.6 gives the results of the CS frequency rate computed as well as the number of segments and the sum of duration of the Speech stretch.

#Seg	# Switch	total dur of seg (m)	Segs/m	CS/m
8130	3012	308	26.3	9.7

Table 4.6: Summary of Speech stretch (segments) and word counts with the average of the number of CS per minute (CS frequency rate) highlighted in gray

The number of switches in this table includes all of the intra-sentential the inter-sentential CS and word insertions in L*b*. The number of the language switch in the corpus represents approximately 40% of the total number of the segments as shown in the table. It also means that the CS is produced every 2 or 3 Speech stretch. The language switch represents $\sim$10 per minute in the speech corpus for the total of speakers. This frequency indicates that the CS is produced highly in this corpus.

4.4.2 Code-switching sentences

The previous sections shown that the language switch is highly frequent in the speech. Thus, we investigate whether in this frequency of CS, typical code-switched phrases are existing and CS phrases are used frequently and regularly by the speakers. So, in this section, CS phrases are extracted from the speech and sorted by number of frequency. We observe in these lists the most common combination of words of FR and AA and the phrase characteristics in the CS speech.
The methods used to extract the CS utterance frequency are the following:

1. Automatic selection of 3 Speech stretch which contains at least one language switch in the continuous speech for each speaker turn.
2. Classify the segments by language switch order in the following four groups of CS utterances:
   AA-FR-AA   AA-FR-FR /AA-FR   FR-AA-FR   FR-AA-AA/FR-AA
3. Numerical sorting of the occurrences in each group.
4. Extract the most frequent CS utterances in each group >4 of the number of occurrences and >4 for the number of speaker for each occurrence.

A total of 2100 CS utterances has been analyzed (combining all CS utterances groups) and 13 utterances are extracted as the most CS formula produced in number of occurrences. The extracted utterances correspond to a part of code-switched sentences with a number of occurrences higher than 4. The highest occurrence of CS utterance extracted is 30. In order to observe the most common CS FR-AA speech productions by the speakers, the occurrences are selected also by the number of the speakers that produced the utterances. To this aim, the utterance selection is based on at least 4 speaker's production. The Table 4.7 shows the CS utterances with their number of occurrences and the number of the speakers producing the utterances.

The list indicates that CS speakers use regular forms of CS utterances. The most frequent utterances in the selected list vary from 30 to 4 of the number occurrences and they also vary

CS utterances	# Occ	# Speakers producing the utterances
candhum beacoup de...	30	16
maA bacd j'ai/ils ont fait	16	11
DurkaA je pense mca	16	8
parce que fiy/fiyh/fiyhum	15	10
Le fait que ydji /yeddiy	14	8
yahdar bien	10	7
mca les études/le travail	6	5
HnaA on fait	6	4
DurkaA je pense mca	6	5
claY/fiy 1-12 mois	5	5
wa laA la même chose	5	5
Durk par rapport li bakriy	5	4
ça fait des mois min alliy	4	4

Table 4.7: CS utterances with their number of occurrences and the number of speakers that produced. The text in blue corresponds to the AA speech and the text in red corresponds to the FR speech

from 4 to 16 of the number of speakers. The selected utterances represent most of language switch orders. However, the utterances with one language switch (the suites of FR-AA and AA-FR) are the most present utterance groups in the list. 7 utterances starting with AA and followed by FR and 3 utterances starting with FR and followed by AA. This regular CS utterances are a part of verbal or nominal sentences. Examples: "wa laA la même chose" ***or** the same thing*, "Dork par rapport li bakriy" ***now** compared to **formerly***, "Le fait que ydji/yeddiy" for the fact that he come/take.

These most frequent CS utterances are all in intra-sentential CS ans they represent mostly short switches with AA and FR particles (li, fiy, min, mca, le fait que, parce que ...). The first utterance that stands out from the set is the highest occurrence "candhum beacoup de..." ***they have** a lot of...* with 30 occurrences and produced by most of speakers (16 speakers of a total of 20 speakers).

Concerning the words that used to produce this bilingual utterances, we notice that words introducing CS listed in the section 4.3.4 contribute to produce a part of the CS utterances, e.g: *parce que, fiy, alliy*. So, these CS introductory words helps to produce CS and may influence to produce regular forms of CS utterances.

Finally, it is important to note that the regular forms of CS utterances and their repetitive use in the speech can be a result of a practice of regular sentences between CS speakers of this language pair. However, this CS utterances may also depend on conversation topics. Indeed, the topic conversations chosen to obtain the recorded CS data are related to study, jobs, personal opinions on life in two countries (See section 3.3.1). Thus, CS utterances noticed in the table as the most frequent in the speech are partly related to the mentioned topics. Examples: mca les études/le travail ***with*** *study/work*, DurkaA je pense mca ***at the moment*** *I think that* ***with***....

4.5 Phone occurrences in code-switching speech

This section gives an overview of the phonemes quantity produced in the CS speech of both AA and FR languages. The aim of these counts is to highlight what are the most frequent phonemes produced and the least phonemes. This overview also gives quantitative information about the phonemic inventory of AA which is less resourced than the FR language. Furthermore, phone frequency results help to focus the acoustico-phonetic and segmental analysis that we plan to approach in the following chapter. The work is divided in three parts, the first part is devoted to the consonants in FR and AA and the second part is about the vowels. A comparison between the languages in consonants and vowels is realized in the third part.

4.5.1 Consonant occurrences

To extract the phonemes produced in CS speech, automatic forced alignment FA at the phone level with canonical pronunciation is used for the data. The figure 4.4 shows the quantitative distribution of the FR consonants. The x axis represents the consonants and the y axis represents the number of occurrences. AA consonants are represented in two figures. The figure 4.5 shows the AA simple consonant counts (simple and emphatic consonants). The figure 4.6 displays the AA geminate consonants produced[1].

[1] The figure presents only the produced consonant in the speech. The consonants that are not produced in FACST corpus are: P, G, þþ, ðð, V, ʁʁ, ʕ

The consonants in the CS speech represents globally 40% of the phones with 37% for the FR. All FR consonants are produced in CS of FACST corpus. The fricative and lateral approximate consonants record the highest occurrences with > 7K for /s ,r, l/. The voiceless plosive consonants present also a high production with > 5K of occurrences for the consonants /p, t, k/, however the voiced plosives are less produced in FR /b,g/. The least FR produced consonants in the CS speech are the nasal /ɲ/ and the plosive /g/ with < 30 occurrences.

The consonant frequency in AA represents 43% for AA counted phones. In the simple consonants, a large number of occurrences is concentrated in the plosives consonants /t, k, b/ with > 1K, the nasal consonants /m, n/ >1.6K, the pharyngeal and glottal fricatives /ħ, h, ʕ/. The group of consonants that record the most occurrences in the corpus is the approximant and the lateral approximant /l, w, j/ with 4K, 1.5K. The figure shows also that the fricative /r/ is frequently produced in AA with 1K occurrences.
The geminates consonants are less produced than the simple, the figure 4.6 shows that the geminate occurrences follows globally the distribution of their simple version in the figure 4.5. Indeed, as the simple consonants, the most frequent geminate consonants are /ll, rr, dd, ww, nn, ss/. In the Arabic language, all consonants can be produced as a geminate, however, a part of geminates are not produced in the CS speech like the consonants of foreign origin /pp, vv, gg/, the glottal stop /ʔʔ/, the fricatives /þþ, ðð, ʁʁ, ʕʕ/. The simple version of theses consonants have mostly low occurrences in the AA simple consonants figure.

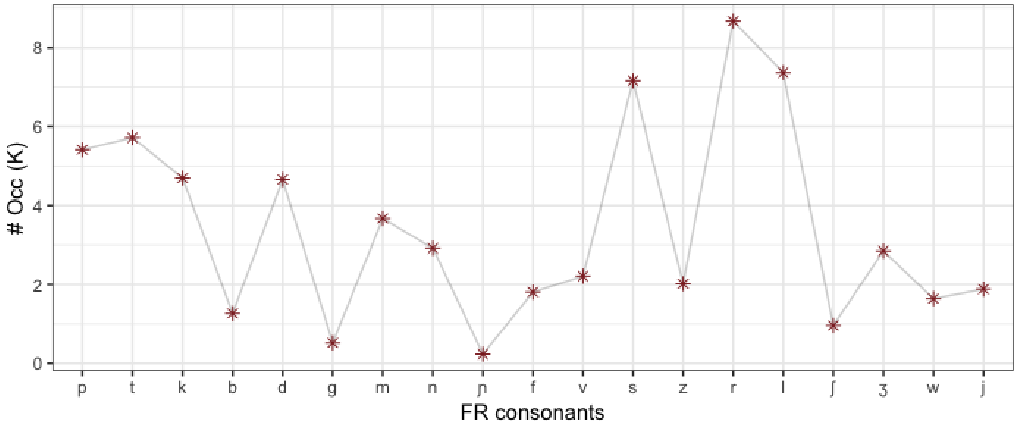


Figure 4.4: Consonants occurrences frequency of FR language in CS speech

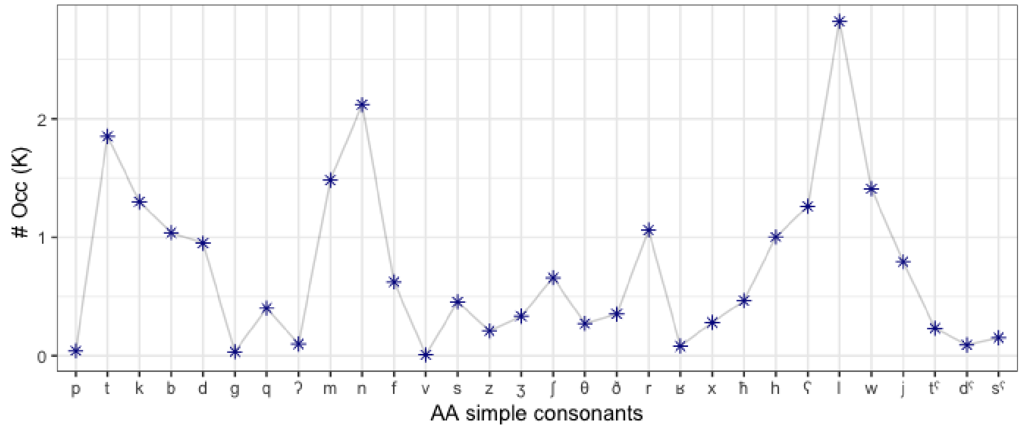


Figure 4.5: Simple consonants occurrences frequency of AA language in CS speech.

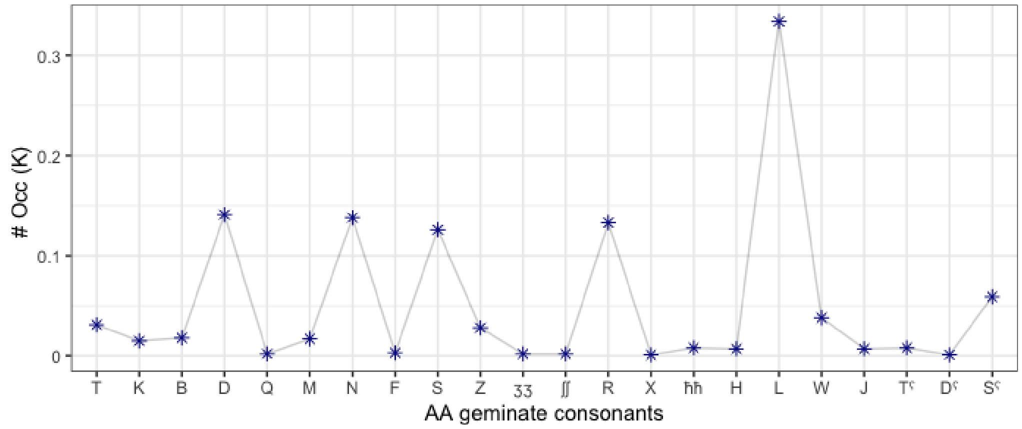


Figure 4.6: Germinate consonants occurrences of AA language in CS speech. The consonants are illustrated with capital symbols or doubling consonants symbols

4.5.2 Vowel occurrences

The vowels are counted from the phones FA as well as the consonants. The figure 4.7 shows the vowels in FR and their occurrences in CS speech. The production of the vowels in FR is height in the most of the front vowels part with 10K for the vowel /a/ as a highest vowel occurrence and > 6K for /i, ɛ, e/. The schwa is also among the most produced vowels with 6K. The nasal vowels and posterior vowels constitute a minor part of the occurrences with 2K, 2.5K and 3.3K for /ɛ̃, õ, ã/.

The figure 4.8 shows the occurrences of produced vowels in AA language. The number of occurrences between long and short of each vowel are getting closer /a, aː/ with 6.5K and

4K, /u, uː/ with ∼1K for both vowels. The exception observed in this corpus is the vowel /i/. It is weakly produced with only 40 occurrences. However, the long vowel /iː/ is highly produced with 2.5K occurrences.

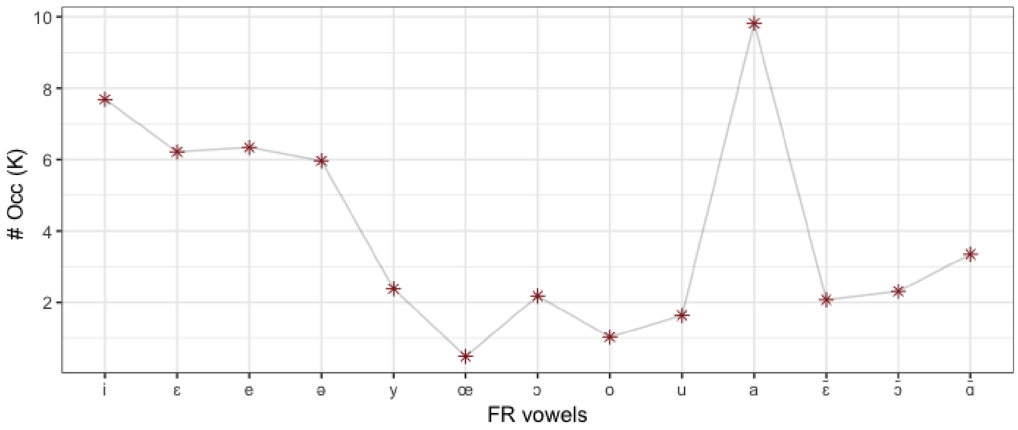


Figure 4.7: Vowels occurrences frequency of FR language in CS speech

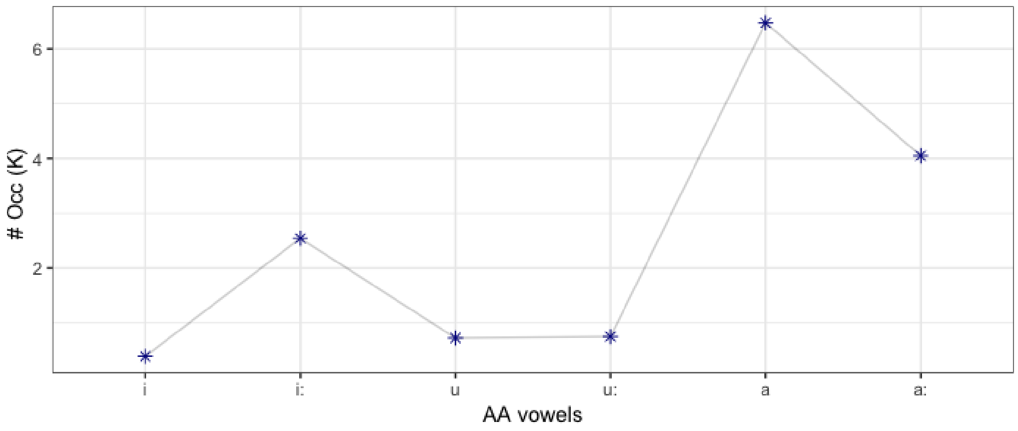


Figure 4.8: Vowels occurrences frequency of AA language in CS speech

4.5.3 Comparison of phoneme occurrences in FR and AA

According to the occurrences production of the vowels and the consonants in FR and AA, this section compares the phone's production of the languages in the CS speech and tempts to describe the particularity of AA and FR phones in CS production.

Consonant frequency comparison

We can note that although the number of the occurrences in FR is higher than AA, the consonant occurrences of the FACST CS corpus shows that the FR and AA share a similar distribution in the speech of a considerable number of consonants as the plosives /t,k,d/ high number of occurrences of the plosives /t,k,d/ they share the same distribution of.

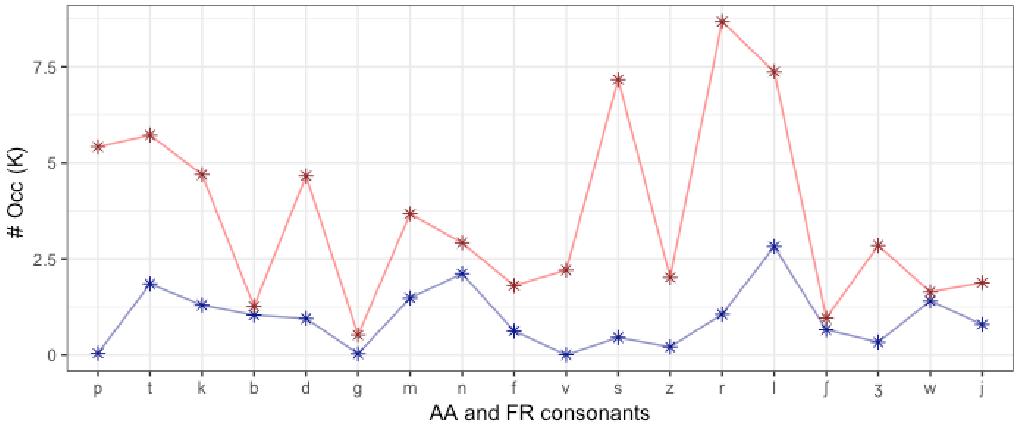


Figure 4.9: Comparison of shared consonant occurrences in AA (blue points) and in FR (red points) in CS speech

Vowel frequency comparison

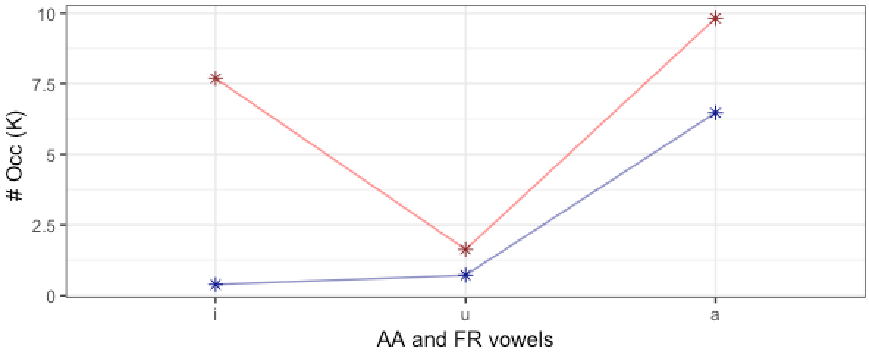


Figure 4.10: Comparison of shared vowels occurrences in AA (blue points) and in FR (red points) in CS speech

After comparing the vowel quantities of both languages we notice that vowel /a/ has a high score in both languages and the /u/. The vowel /i/ has a high score only in FR and has a relatively low score in AA. The distribution of occurrences in both vowels is different because FR vowels are numerous in contrast to AA vowels since AA has only three vowels with long and short versions.

4.6 Language identification test in code-switching speech

This study focuses on only CS speech part of FACST 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 FR-AA as compared to the neighboring countries Morocco and Tunisia (Amazouz et al., 2016)‘. We report on some initial studies to locate French, Arabic and the code-switched speech stretches, using ASR system word posteriors for this pair of languages.

4.6.1 Experiment

In the following, we report on an experiment to detect CS using automatic speech processing tools. 5 speaker’s CS speech are selected and the speech corresponds to 4 hours. 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 band/telephone band...). These segments were then automatically transcribed using different ASR systems Gauvain et al. (2003); Laurent et al. (2016) 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.

4.6.2 Results

Figure 4.11 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 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).

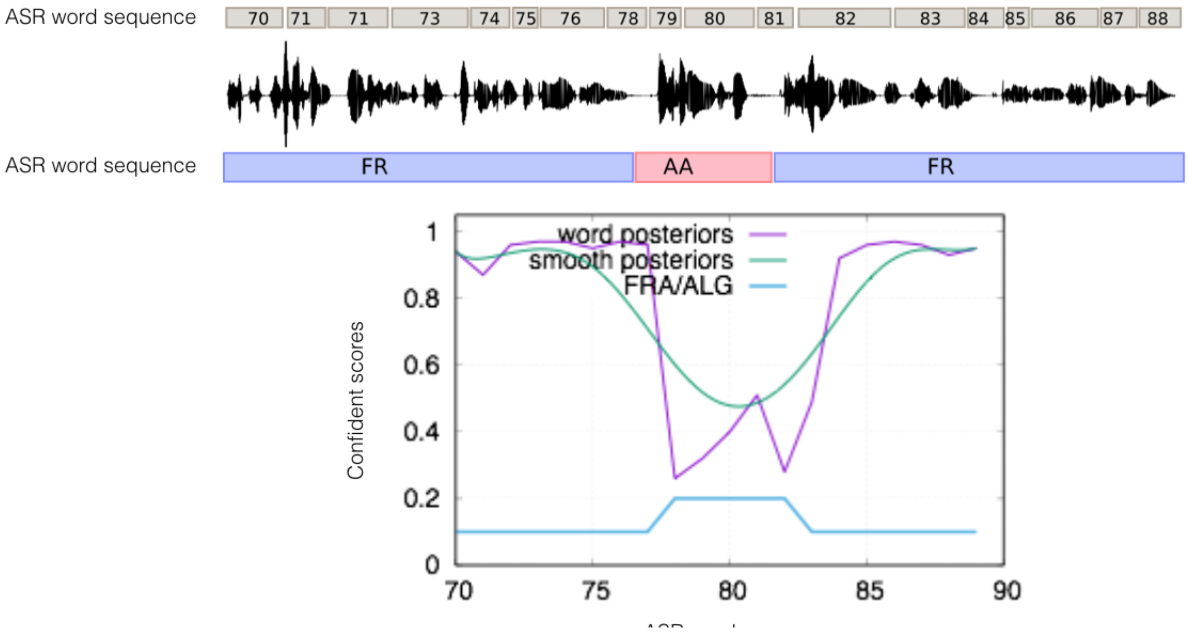


Figure 4.11: ASR word posteriors of the French transcription systems (raw scores and smoothed). The X-axis corresponds to an excerpt of speech: words numbered from 70-79 and 85-89 are in FR, words numbered from 80-84 correspond to an AA code-switch. The lowest curve (blue) denotes FR (0.1) or AA (0.2).

Table 4.8 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.

CS language → **Speaker**	**Fre**	**Ara**	**Fre**	**Ara**
	French ASR		**Arabic ASR**	
Speaker 1	**0.74**	0.72	0.54	**0.56**
Speaker 2	**0.79**	0.57	0.58	**0.65**
Speaker 3	0.74	0.78	0.51	**0.54**
Speaker 5	**0.78**	0.61	0.56	**0.58**
Speaker 6	**0.62**	0.59	0.59	0.56
Overall	**0.76**	0.64	0.56	**0.59**

Table 4.8: Average word level posterior scores by the speaker with the French (left) and Algerian (right) ASR systems, for words of the CS segments manually annotated as French and Arabic.

4.6.3 Discussion

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 on Language Identification (LID) experiments and developing ASR methods for FR-AA CS speech. 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 at acoustic, phonetic and prosodic levels.

4.7 General discussion

In this chapter, we presented a quantitative review of the FACST data. We studied the following points: description and frequency of the words, phones and lexicon in CS. The frequency of CS sentences and the number of switch. Comparison between the FR and AA at sentence, words and phone levels. Also, we presented a test of automatic identification of CS speech.

The quantitative review given in this chapter reveals that FACST CA data have more

FR speech compared to AA in terms of the number of speech stretches and words. This allows us to conclude that FR dominates the productions. However, being the dominant language in a corpus does not mean that this language is the basic language of the sentences produced by the speakers. In fact, the CS productions contain AA statements that contain FR items/utterances. This quantitative review also gives information about CS composition and organization with FR-AA languages. The switch between languages relies on lexical borrowings, on switches of grammatical items and shows that the language utterances are constructed from sentences using both languages. The protocols followed in Chapter 3 in order to get the data allowed us to obtain a CS spontaneous speech with a high switch frequency, but also with various CS types (inter and intra-sentential CS, L*b* word insertions, borrowings, etc.).

The approach, based on the automatic processing of speech large data allows us to conclude that the words that introduce CS are lexical and grammatical words. Both of them are found in intra-sentential CS such as linking two units in a sentence in L*a*. Borrowings in CS are seen to be inserted in both directions. The results obtained on words introducing CS push us to think about regular forms and clues to signal a language switch, and the possibility of predicting a switch in this language pair. Thus, these results of regular forms of switch can provide more information for automatic CS identification and then automatic CS speech recognition.

This review concludes that one of the major characteristics of FR-AA CS is switching between very short segments (particles, word, short Speech stretch). These short switches are very challenging for human annotations and automatic speech recognition.

To sum up, this quantitative review of CS data gave a overview of the data quantities in the corpus for linguistic, phonetic and speech technologies studies. It highlighted also the CS FR-AA organization and the particularities of this pair of languages in CS speech.

Chapter 5

Segmental variation in code-switched speech

Chapter contents

ALGERIAN Arabic-French bilingual speakers show phonetic variation with respect to vowel timber and consonant production in both languages. In this chapter, we investigate speech variation in vowels and consonants as produced in CS speech. Our investigations are divided into two parts: the first one focuses on vowel variation and the second part deals with consonantal variation for bilingual FR and AA and native monolingual French speakers. The vowel and consonant investigations each have into studies. Concerning vowels, we first compare the observed variation in French and Arabic vowel productions from the same set of bilingual FACST speakers. We then focus on French vowels only and compare the bilingual productions to those of native (monolingual) speakers of French. Finally, we investigate vowel centralization in both French and Algerian Similarly, for consonants, our investigations are divided into three sub-parts: the first one deals with geminates and gemination variation in CS speech, the second one deals with pharyngealization (emphatization) in FR and in AA, and in the last part we deal with the subject of voicing and devoicing of consonants in CS speech.

The methodology we adopted to explore production variation relies on automatic forced alignments permitting specific variants in the pronunciation dictionary. This approach can be seen as an automated approximation of an ABX-like design widely used in perceptual studies to measure the discriminability between categories. Beyond studying variation using forced alignment with variants, we also carried out acoustic analyses in order to supplement, complement and reinforce the alignment experiments.

5.1 Methodology and experimental design

In order to give a glimpse of what is carried out during forced alignment using specific variants, it can be useful to think about the well-known ABX discrimination paradigm. In this paradigm, given two tokens with category labels A and B respectively, one would like to know which of the two categories a third unknown token (labeled X) is most similar to. In psycholinguistics, this measure of similarity is carried out perceptually using a population of native listeners.

In our case, we study production variation through an automatic ABX-like forced choice paradigm, but the implementation is somewhat different from what is described above. We do not operate on triplets of acoustic tokens, to carry out the discrimination task. As we make use of acoustic models during automatic forced alignment, the A and B categories are represented by their corresponding acoustic models. The adopted ABX-like paradigm then translates as follows: given the two categories (acoustic models) A and B, is the X token better explained (matched) by the A model or by the B model? This binary forced choice paradigm can be easily extended to an arbitrary number of categories.

5.1.1 Automatic alignment with variants

Method Figure 5.1 gives a schematic overview of forced alignment, which can be considered as a sub-process of an automatic speech recognition system. The top part of the figure illustrates the basic task of forced alignment: given the speech signal and its transcription as input, the forced alignment process locates words and composing phones in the signal, thus providing the time stamps of their hypothesized boundaries. If pronunciation variants are proposed for a given word, the forced alignment process will also carry out the task of choosing the best matching variant. This is illustrated in the bottom part of Figure 5.1. In particular, the illustration highlights the variants paradigm as we propose to use it in the following investigations and which we sometimes refer to as *parallel variants* paradigm. For a given sound category a, add as an alternative the (competing) sound category b in all positions where the a category appears, and let the system decide for the incoming x signal which one of the two a (target) or b (competing) categories match best. In the illustration of Figure 5.1 (bottom), all occurrences of a in the input stream are added b,and the first occurrence of sound category a in the output stream is replaced by category b to exemplify that b was found to be more similar to signal x in this position.

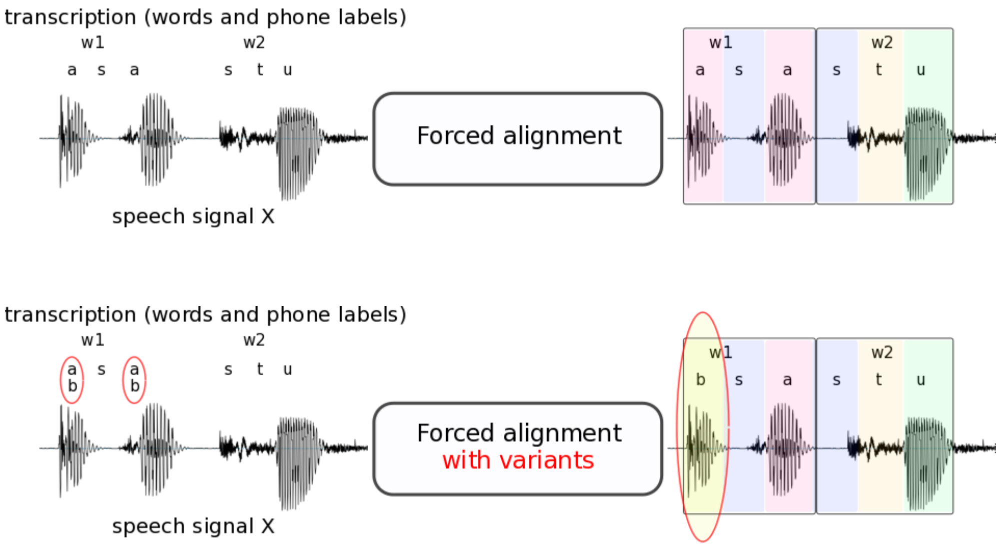


Figure 5.1: Schematic representation of the forced alignment process. Input includes signal, transcription and phonemic representation. Output gives time stamps to words and phones Top: no variants in phonemic representation. Bottom: with local variants, highlighted by red circles. In addition to time stamps, the system chooses the best matching acoustic phone model among proposed variants a or b.

Implementation The automatic forced alignment of the parallel variants across all speech corpora and for each experiment was realized using a set of position-independent monophone acoustic models similar to those described in (Gauvain et al., 2002; Lamel et al., 2004; Gelly et al., 2016; Lamel et al., 2009). This setup was preferred to context-dependent acoustic models, as previous studies showed that these large sets of context-dependent models typically used in speech recognition systems, tend to capture very specific co-articulation variation which may reach beyond a simple segment location. For example, in French spontaneous speech, shortened and devoiced high vowels may be typical in some obstruent contexts (as for /y/ in French *tu sais* "you know", typically produced as [tsɛ] without [y] segment between [t] and [s]). We therefore prefer using context-independent phone models, as they average and represent the spectral characteristics of all occurrences of a given sound, rather than only a subset extracted from a specific left/right context (Adda-Decker and Lamel, 1999; Mareüil and Adda-Decker, 2002). The alignment system locates word and phone boundaries using orthographic transcriptions and the best matching pronunciations chosen among the pronunciation variants that are included in its dictionary. For technical reasons, the segmentation resolution is limited to 10 ms and the minimum duration of a segment is 30 ms. The phone labelling is not really phonetic, but rather phonological or phonemic (corresponding in most cases to standard word pronunciations).

Typical pronunciation variants in French are due to optional liaison consonants and schwa vowels (which may be described as sequential variants), allowing for one more or one less phone symbol in the pronunciation i.e. the word *facile* (easy) might provide the following choices to the system: [fasil], [fasilə], (and for example, in the case of a vowel /i/ centralization experiment, additional [fasəl] and [fasələ] variants which allow for parallel [i,ə] variants). Other typical variants are due to word-final consonant cluster simplifications as in the word *autre* (other) which provides the following choices to the system: [otr], [ɔtr], [otrə], [ɔtrə], [ot], [ɔt]. This particular example combines both parallel ([o, ɔ]) and sequential (optional [r] and [ə]) variants. However, in general, most lexical entries tend to be described by their canonical (full form) pronunciation.

For the vowel variation experiments, the automatic alignment system makes use of standard French acoustic models. As the French inventory contains more vowels than are present in Algerian Arabic, the French acoustic models allow us to quantify what happens in a larger number of smaller vocalic locations than if we made use of the Arabic acoustic models. In particular, the larger set of French vowel acoustic models allow us to quantify whether the Algerian Arabic vowels are realized in a similar way as the corresponding French vowels or whether they tend to be shifted and if so in what direction. It is noteworthy to remind that the use of French acoustic models should not lead to interpretations such as aligned variants correspond to realizations of French phonemes, but rather that the realization of an Arabic vowel is acoustically close to that of a given French vowel.

For the consonant variation experiments, the automatic alignment system is based on Arabic acoustic models, as Arabic has a larger inventory of consonants than French and all studied French consonants have a counterpart in the Arabic model set. As for French, the Arabic acoustic models consist of position-independent monophone acoustic models similar to those described for French.

Discussion The forced alignment with specific variants method may be considered as globally more objective than human annotation, as exactly the same acoustic models and the same decision measure is applied in all positions over time. Above all, this method is extremely more time-efficient than human labelling. However the variant choices are categorical and limited to the options offered by the variants included in the pronunciation dictionary, which motivates the proposed term of ABX-like categorization for this variant alignment approach. In order to explore gradual variations, additional analyses such as formant measurements and perceptual tests using human listeners are necessary. In this work some automatic formant analyses are proposed, however perceptual tests go beyond the scope of this thesis.

5.1.2 Computing production variation

As a preamble, we want to clarify that each considered parallel variant configuration requires a specific forced alignment run. For example, if we want to test parallel variants for 5 different target vowels, we ran 5 different specific forced alignment jobs and measured the production variation specifically for each outcome.

In the following we describe how the outcome of the automatic ABX-like categorization process is obtained from the results of the targeted forced alignments. Production variation is quantified using the output of the forced alignment (with variants) for each target category by measuring the rates of the corresponding competing categories that are selected during the automatic alignment phase. For instance, if all occurrences of a given target category were unchanged (that is the original phone was always selected without making use of any of the competing variants) the variation rate would be 0%. If in all instances a competing variant was selected, the variation rate would be 100%. In the result section, for each experiment, the variation will be described with the help of such variant rates. The figures display for each tested category the stacked variant rates which always sum up to 100%, thereby clearly featuring the rates for all (target and competing) categories. By convention, we put the percentage associated with the target category first (which is in fact the "not-a-variant" rate) and then add the "true" variant rates indicating the proportion of occurrences of a target category replaced by a competing category. As many of our experiments include more than one competing category, our experimental design is rather an ABB'X (or ABB'B"X...) design than a mere binary ABX choice.

5.1.3 Design of experiments

This section summarizes the set of production variation experiments that we will carry out using our parallel variants (automated ABX-like) paradigm. As announced earlier, a first set of experiments deals with vowel production variation and makes use of French acoustic models as the French language has a richer vowel inventory than Arabic. The second set of experiments aims at studying consonantal variation with the help of Arabic acoustic models as the consonant inventory of is richer in Arabic than in French.

With respect to vowels, three series of experiments with targeted contrasts are performed: AA vs FR vowel production variation in code-switched speech, French vowel production variation in bilingual CS speech and in monolingual speech using FACST corpus and NCCFr corpus and finally a vowel centralization study in CS speech involving both languages.

Table 5.1 summarizes the consonant variation experiments with the AA and FR consonants sets. The geminate simplification experiment is reported in separate table (Table 5.2).

5.1.4 Acoustic measurements

Beyond studying variation using forced alignment with variants, we investigate the acoustic dimension of both vowel and consonant variation experiments. These acoustic analysis are carried out in order to compete and reinforce the alignment experiments. It should be noted that the acoustic analysis that complete alignment experiments are not considered as the main part of the segmental variation in code-switched speech study. However, we consider combining the alignment results to their parallel acoustic results is a method to check the automatic alignment results in the acoustic signal data.

Concerning vowel variation in CS speech, we investigate the formant values of F1 and

Consonant target	Gemination	Pharyngalisation	+- voisement
b	bb		p
p	pp		b
d	dd	d^{ʕ}	t
t	tt	t^{ʕ}	d
g	gg		k
k	kk		g
ʒ	ʒʒ		ʃ
ʃ	ʃʃ		ʒ
m	mm		
n	nn		
f	ff		v
v	vv		f
s	ss	s^{ʕ}	z
z	zz		s
l	ll		
r	rr		
w	ww		
j	jj		
d^{ʕ}		d	
t^{ʕ}		t	
s^{ʕ}		s	

Table 5.1: Summary of AA and FR shared consonants target and the variation experiments using AA acoustic models

Geminate	P	T	K	B	D	G	F	S
Var	p	t	k	b	d	g	f	s
Geminate	ʃʃ	V	Z	ʒʒ	M	N	L	R
Var	ʃ	v	z	ʒ	m	n	l	r

Table 5.2: Summary of AA geminate consonants variation experiments

F2 in order to spot the vowel variation in the vocalic space and in order to compare the variant alignment results with the acoustic values of of the vowels. To this aim, a vocalic diagram of the F1 and F2 vowel values with variation spaces is calculated for each part of the experiment. The formants measurements are extracted from the vowel signal of the speech. The acoustic measurements are realized with the help of Praat scripts which are based on Burg algorithm (Childers, 1978). The steps of vowel variation acoustic measurements and the results are described in Section 5.2

With respect of the consonants, three methods of acoustic measurements are used following the studies. Concerning the geminates and gemination as a variation in CS speech, we referred to the duration criterion of the consonants to identify and evaluate the gemination of the simple consonants and the simplification of the geminates. To this aim, we calculated the consonant duration of both simple and geminate consonants with the help of automatic phone segmentation in the signal. We recall that, like the alignment experiment variation, the acoustic measurements of gemination variation of simple consonant concerns both AA and FR languages in CS speech and the consonant simplification variation concerns only the AA consonants. The details of the alignment experiment and its parallel acoustic measurements are described in Section 5.3.1

The emphatics variation and the consonant pharyngealization alignment results are compared to the F2 values of the adjacent vowel in CV context of the consonant targets. In this method, we referred to identify the consonant variation in the signal to the F2 onset values of theses vowels. In theory, the pharyngealization of a simple consonant shows a higher F2 onset values of the followed vowel and the simplification of an emphatic consonant. As the vowel experiment, we used Praat script program to calculate the formants values. Section 5.3.2 details the acoustic measurement steps, the obtained acoustic results and the comparison with the parallel alignment results.

In order to investigate acoustically the voicing variation in obstruents consonants in CS, we measured the voicing rate of the target consonants which is obtained from F0 voicing values calculated in the signal with Praat scripts. Section 5.3.3 describes in details the acoustic measurements methods to calculate the voicing rate (v-ratio).

5.2 Vowel variation in code-switched speech

French and Arabic Algerian (AA) languages share the three most peripheral vowels, namely /i/, /a/ and /u/ in their phonemic systems. In French, there are many additional vowels, whereas the Arabic inventory remains limited. These differences in vowel inventories in both languages may translate into notable differences in the realisations of these vowels in French as compared to their realisations in AA. How are these differences expressed in productions?

Different hypotheses seem viable depending on speakers' profiles or preferences: /i,a,u/ productions in French are expected to remain more canonical (close to the vertices of a vocalic F1-F2 triangle) than the corresponding productions in AA which may be allowed to cover a larger acoustic space as there are fewer phonemic competitors. We thus may expect larger variant rates in AA speech than in FR speech. We also may wonder whether production variation favors a given region of the acoustic space, whether a specific neighboring vowel category tends to be favored, whether the production is rather a more centralized one or a more open/closed realization and so forth.

As we have two types of speech, namely spontaneous CS speech and read speech, we may also investigate whether variant rates remain similar across these different production styles. Reading may entail a higher awareness of the vowels to be produced and hence reduce variant rates, however the reading exercise, which is far from natural to many speakers, may be a perturbing factor and hence decrease the speakers' attention in speech production, as their cognitive load might be directed towards FR/AA script decoding.

Finally, we also compare the French productions of our bilingual FACST speakers to native French speakers from the NCCFr corpus, which includes a population of mostly young Northern metropolitan French speakers. This experiment shows us whether interesting differences can be measured using our experimental setup.

The proposed investigations of vowel variation with the parallel variants paradigm then aims at addressing the following research questions: Given our bilingual French-Arabic Algerian population of speakers, do their /i,a,u/ vowel productions change significantly between their two languages. We investigate the bilingual speakers' production variation with respect to vowel timber in both their languages. Our study aims to automatically identify (with the help of an automatic alignment) vowel variants frequently produced in bilinguals. To that end, the speech corpus FACST, containing French and Algerian Arabic code-switched speech, was analyzed. A second corpus with native French speakers NCCFr was used as control group to provide a reference baseline and to compare vowel variants across the two French speech groups.

5.2.1 Vowel variation in code-switching speakers

Hereafter, we make use of the parallel variants paradigm to study production variation in bilingual speakers using both spontaneous CS speech and read speech in the two languages. We tempt to answer the following questions: how does the production of vowels vary in code-switched speech as a function of the chosen language (French or AA)? To what extent does the speech style (CS and read speech) influence the production of vowels?

5.2.1.1 Experiment

We focus on vowel production variation in bilingual speakers as a function of the spoken language, FR or AA. We limit the investigations to the three cardinal vowels [i, a, u] which are shared by the two phonological vowel systems. For each of the three target vowels, two competing vowels are introduced in order to investigate the vowel quality variation as shown in Table 5.3. The competing variants correspond to nearest neighbour vowels in a vocalic triangle representation. For the most open /a/ vowel we define two nearest neighbour configurations depending on whether variants tend to become more fronted or more posterior. Table 5.3 thus shows two lines of competing variants, the more fronted neighbours and the more posterior neighbours, the latter configuration will be termed [a]2 in the figures displaying the results.

Target vowel	Competing variants
[i]	[e, y]
[a]	[ɛ, œ] [ɔ, œ]
[u]	[o, ø]

Table 5.3: Competing vowels as used in the forced alignment with variants experiments for both AA and FR.

CS and read speech from FACST corpus are used. Results are examined at two levels, first we compare the variation due to language change in CS speech and secondly we carry out the same analyses on the read speech recordings in both languages.

5.2.1.2 Results

We first present the CS speech results in FR and AA before comparing these with the read speech results in both languages. Figure 5.2 shows the stacked variant rates for the three cardinal vowels, including two different configurations for the [a] target with (anterior or fronted [a] and posterior [a]2 competitors). For each tested configuration the first bar of stacked results shows variant rates as measured for French, the second bar is for AA. Overall, the variant rate results show that for FR, the target vowels are most frequently aligned using their canonical pronunciation with >75%. However, when the same speakers switch to AA, their AA vowels seem to vary more than their FR counterparts, as on average, less than 40% of aligned vowels correspond to their target labels.

Our variant rate results thus suggest rather stable target vowels for French productions by bilingual speakers and less stable target vowels in AA for the same vowel group (with the same speakers). Looking in more detail at figure 5.2 for each target vowel configuration, in FR, the target vowels are aligned as such as follows: [i]: 76%; [a] (ante): 75%; [a] (post): 77%; [u]: 74%). Results are very similar across vowels. In the AA vowels, the target vowel is aligned as such in less than 40%, the tendency is to prefer one of the competing variants to the target vowel. Likewise, the most frequently aligned vowels for [i] and [u] are respectively the competitor vowels [e] with 40% and [o] with 37%. The vowel [a] remains more stable in CS AA for both conditions (ante/post), as it is aligned as [a] (ante: 44%; post:39%) in a relative majority of cases. In general, the favoured variants for [a] in both conditions (ante/post) and for both languages (FR/AA) is the central vowel [œ] (ante-FR: 17%; AA: 31%; post-FR: 10%; AA: 48%). The much higher [œ] variant rate for the [a]2 (post) condition suggests that part of the vowel segments that were aligned with [ɛ] in the [a] (ante) condition are finally better taken into account by the [œ] acoustic model rather than the target [a] model when the fronted [ɛ] option is no longer available. Taking this acoustic phone model perspective (we need to remember that the alignments were carried out using native French monophone models which reflect the average production of native French speakers) these rates tell us that the produced AA [a] segments are close to the central [œ] acoustic model which achieves the highest variant rate. This model itself tends to overlap with the fronted and posterior mid-open vowels [ɛ] and [ɔ]. The fact that variant rates are higher for the "post" ([a]2) condition, with a total of 58% as compared to 48% for the ante ([a]) condition, may suggest that our code-switching speakers produced [a] segments that are globally rather backed than fronted. For the cardinal [i] and [u] vowels, however, the central variants appeared less frequently in the alignment than the mid-vowels [e] (FR: 14%; AA: 40%) and [o] (FR: 16%; AA: 37%) respectively. Globally, in CS speech, the [u] vowel has the lowest rates of segments aligned canonically (73% for French and 32% for AA), the most important competing variant being [o] with 17% in FR and 37% in AA.

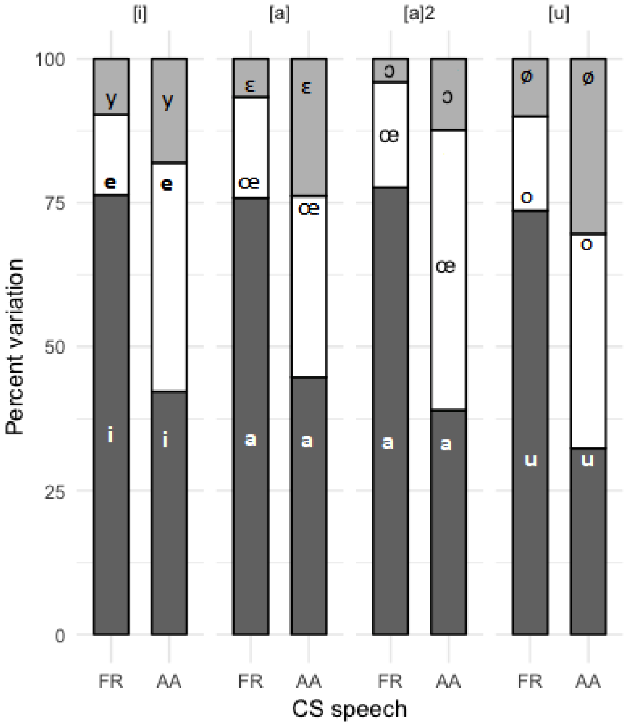


Figure 5.2: Vowel variant rates in FR and AA in CS speech. The target vowel rate is given first (bottom). All variant rates are stacked to sum up to 100%. For the [a] target, the first pair of bars corresponds to the anterior condition, the second pair to the posterior variants condition (labeled [a]2).

We now turn to the read speech data of our FACST corpus. Variant rates are shown in figure 5.3. On average, the investigated vowels are aligned as their target in the majority of cases. And as previously observed in CS speech, the tested variants suggest that vowels are globally less stable in AA than in FR in the read speech condition. The AA vowels /i/

and /a/ are aligned respectively in 52% and 33% (ante), and 44% (post) of the cases as the target vowel, compared to FR with 84% for [i], 87% for [a] (ante), and 89% for [a] (post) aligned as target vowels. The vowel /u/ is aligned with respectively 29% and 85% as the target vowel in AA and FR. The most frequent variant for [u] in read AA speech was [o] with 55% of the cases. The variant rates of /u/ are highest for AA, the competing vowels totaling 75% of the aligned [a] segments ([o] with 55% and [ø] with 20% respectively).

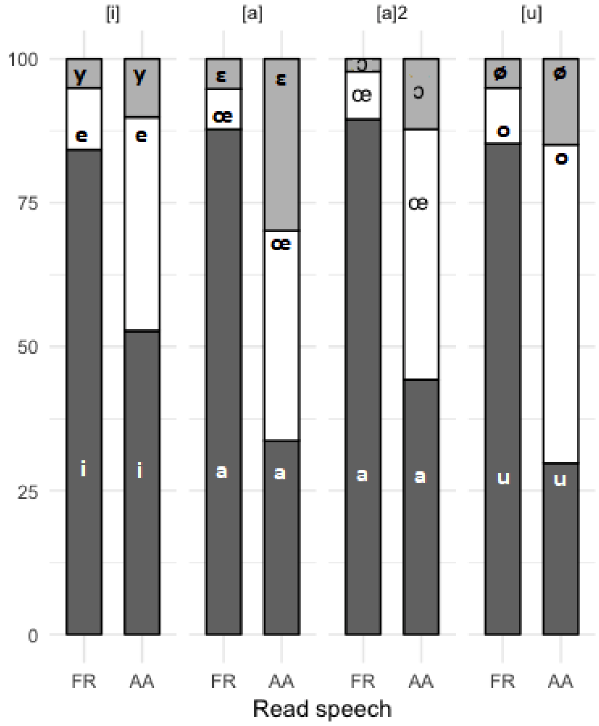


Figure 5.3: Vowel variants in FR and AA in read speech. The target vowel rate is given first (bottom). All variant rates are stacked to sum up to 100%. For the [a] target, the first pair of bars corresponds to the anterior condition, the second pair to the posterior variants condition (labeled [a]2).

The more frequent variant for /i/ is the mid-open vowel [e] (FR 11%; AA 37%), whereas in both conditions of /a/ (ante/post), the more central vowel [œ] (ante-FR: 7%, AA: 36% ; post-FR: 8%, AA: 43%) was selected in both languages.

According to our methodology, the CS vs read speech comparison globally reveals that vowel variation is higher in CS speech than in read speech for both languages. As a matter of fact, in code-switched speech, all target vowels are more often aligned with their competing variants than in the read speech data. We may note some more detailed differences between the two speech styles, especially for the AA part of the data. Read speech shows most variation in the AA /u/ vowel With respect to the /u/ vowel, the proportion of segments which are aligned as [o] rather than [ø] is higher in read speech than in CS for both languages. In FR, we observe that the three cardinal vowels are aligned as their targets [i, a, u] in most of the cases in both read and code-switched speech. When looking for the most stable vowel overall (with respect to our methodological protocol), similar tendencies can be reported for read and code-switched data and also for both languages: the vowel [i] is most stable. The variant [y] instead of the target vowel [i] appears only in 10% in FR and 18% in AA in CS speech (where its rates are highest as compared to read speech), however the variant [e] remains the most frequent (FR: 13%; AA: 40%). We may relate the front vowel [i] shift towards [y] in code-switched speech to a decrease in F3.

Before turning to acoustic formant measurements, we undertake some statistical analyses on our variant rate results. The statistical analysis shows that vowel variant rates in CS speech are vowel dependent ($\chi^2(6) = 25.52$, $p < 0.001$). Compared to the other target vowels, /i/ allows the first variant ([e] in our case) more often (27 %) for a total in AA and FR. Inversely, for the target /a/ with posterior variants, the first vowel variant [ɔ] is produced less frequently (AA:12%, FR:7%) compared to the first variant of the other vowels.

Moreover, language also has an impact on the vowel variant rates ($\chi^2(2) = 12.96$, $p < 0.01$). Overall, CS speech produces the target vowel more often in French (75 %) than in Algerian Arabic (41 %). Detailed analysis of each target vowel and its variants revealed that CS significantly modifies speech production according to the language (/i/: ($\chi^2(2) = 10.18$, $p < 0.01$); /a/ with anterior variants: ($\chi^2(2) = 21.04$, $p < 0.001$); /a/ with posterior variants: ($\chi^2(2) = 12.92$, $p < 0.01$); [u]: ($\chi^2(2) = 13.97$, $p < 0.001$)). CS produced significantly more target vowels in FR for [i] (76%), [a] with anterior variants (75%) and [u] (73%)than in AA ([i]: 42%, [a] with anterior variants: 45% and [u]: 32%). Regarding the target vowel [a] with posterior variants, CS substituted [a] more often by [œ] while speaking AA (31%) than while speaking FR (10%).

In read speech the variant rates differ according to the vowel ($\chi^2(6) = 23.37$, $p <$

$0.001)$. For the target vowel /i/ the first variant [e] was chosen more often (23%) than other target vowels. Inversely, the first variant ([ɔ]) of the target vowel /a/ (post) was chosen significantly less often (7%) compared to the first variant of the other target vowels. Furthermore, variation rates depend on the language $(\chi^2(2) = 19.19,\ p < 0.001)$. In FR read speech, the target vowel is produced more often (86 %) than while read AA (40 %), leading to higher variation rates in AA speech.

Finally, it is interesting to note that for the 3 vowels [i,a,u], CS speech has much larger variant rates in AA than in FR. The following subsection sheds more light on variations between FR and AA speech.

5.2.1.3 Formant analysis

For this part devoted to a more acoustic analysis, formants are extracted with the help of Praat. Mean values of formants F1 and F2 are computed. Beforehand, the formant values were filtered to get rid of potential major formant detection errors, especially for cardinal vowels which tend to have two very close formants. The adopted tolerance ranges of the filters were ± 400 Hz around the reference values as reported in (Gendrot and Adda-Decker, 2005) for French and the values cited in (Barkat, 2000) for Arabic.

Figure 5.4 shows the vocalic space of AA productions (left) and FR (right) in read speech. F1 and F2 means of the target vowel segments are measured using Praat formant extraction program. Formant analysis shows that AA vowels share common spaces to each other. A part of the vowel /i/ is realized with an increase of F1, which produces a more opened vowel. Vowel /a/ production is realized within a space that belongs to open and mid-open vowels (with a lowering of F1 values).Concerning the vowel /u/ it is mainly realized within a more central space and with F2 increasing values which corresponds to vowels /o/ and /œ/.

French vowels tend to be more distinctivelu separated, especially for vowel /i/ which is very stable, vowel /a/ shares a minimal vocalic space with vowel /u/ with a lowering of F2 values, which corresponds to the vocalic space of /o/. Concerning vowel /u/, it also shows an F2 increase that brings it closer to the vocalic space of /o/. Hence, this formant analysis corroborates the alignment results that showed that read AA has the higher rate of variations, and that vowel /a/ varies towards /o/ and /œ/ and that vowel /u/ is produced in a vocalic space close to vowel /o/.

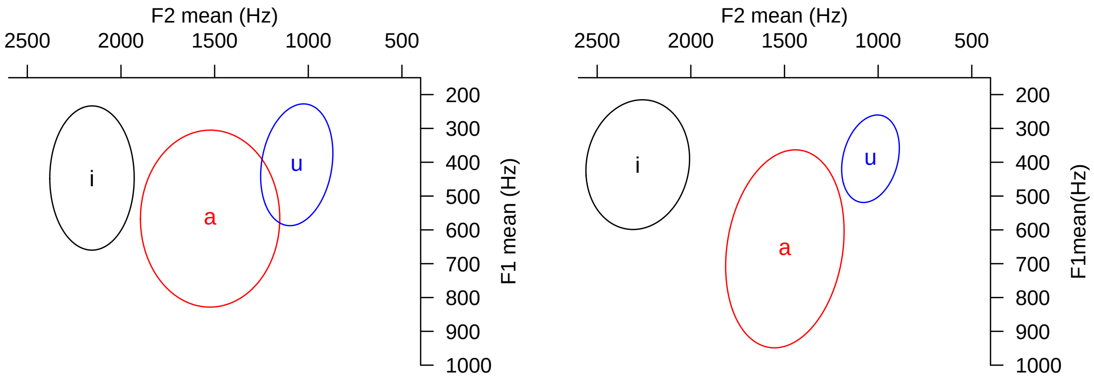


Figure 5.4: /i, a, u/ F1 and F2 in AA (left) and FR (right) read speech.

The CS /i, a, u/ vowels in Figures 5.5 show that the variation rate is higher in CS speech compared to read speech, especially for vowel /i/. The AA vowel /i/ has the highest spectrum values in F2 while F1 tends to increase. This may be related to the high variation rates achieved in the alignment results (/i/ which varies in /e/ and /y/). Vowel /a/ expands to a production with lower F1 and F2 values, which may explain its variation towards posterior /o/ in alignment results. Vowel /u/ shows increased F1 and F2 values, that correlate with production variation in /o/ and in /ø/.

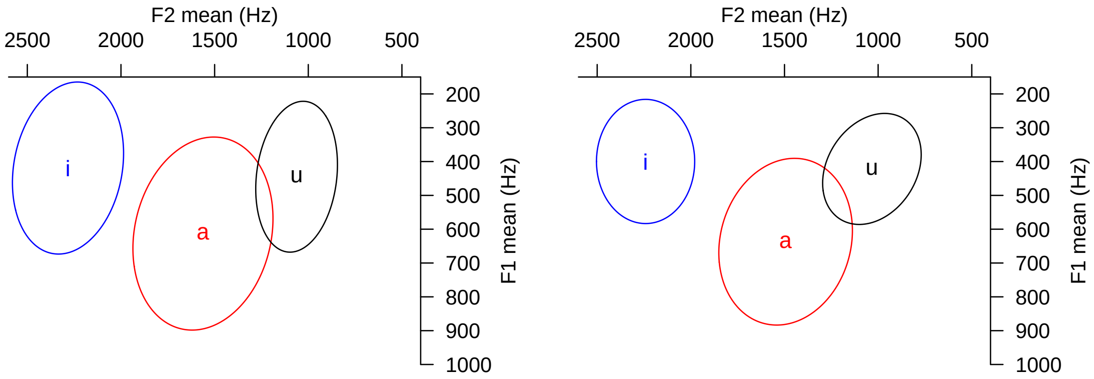


Figure 5.5: /i, a, u/ formants in AA (left) and FR (right) CS speech

5.2.2 French vowel variation in bilingual and monolingual speakers

The following experiment focuses on French vowels variation in CS bilinguals speech as compared to native French speakers. We tempt to answer the following questions: how do vowels vary in French produced by bilingual speakers in comparison to native French speakers? In contrast to the previous section we will include here also mid vowels /e/ and /o/. We may raise the following questions: are bilingual speakers producing similar vowel variation as compared to native speakers? Or are bilingual speakersless consistent and produce more variation due to a potentially interfering vowel system from the other language? Or on the contrary, do they realize the French vowels rather canonically thus well separating French vowels from Algerian Arabic vowels?

5.2.2.1 Experiment

In the following, we start measuring vowel production variation using the proposed method with spontaneous speech of native French speakers (NCCFr corpus) to establish a reference. Then, we apply exactly the same method to the FR speech parts of our CS speech data (FACST corpus). Our analyses focus on the French productions. (As opposed to the previous section, the speaker populations are of course different here). Production variation is measured for five target vowels, where each target vowel is put in parallel with two competing variants. A specific pronunciation variant lexicon is built for each condition. The defined target vowels are the peripheral vowels [i, e, a, o, u] with competing variants as shown in Table 5.4.

Target vowel	Competing variants	Example
[i]	[e, y]	lit (bed) : li, le, ly
[e]	[ɛ, œ]	nez (nose) : ne, nɛ, nœ
[a]	[ɛ, œ] [ɔ, œ]	chat (cat) : ʃa, ʃɛ, ʃœ (anterior) (cat) : ʃa, ʃɔ, ʃœ (posterior)
[o]	[ɔ, ø]	chaud (hot) : ʃo, ʃɔ, ʃø
[u]	[o, ø]	loup (wolve) : lu, lo, lø

Table 5.4: Competing vowel variants for each target vowel.The last column exemplifies the effect on the pronunciation lexicon. Note that for the most open [a] vowel, two sets of variants (ante, post) are proposed.

For each target vowel, competing variants were chosen as the two nearest neighbors towards the center of the vocalic space (triangle) with respect to the target vowel's location. For the open vowel [a], we tested two different configurations, termed as anterior and posterior. We also included /i/ as a variant for /e/, and /y/ as a variant for /u/. Our initial analyses indicate that there is no significant change to the results without these extra variants.

The different pronunciation variant lexicons were tested on both corpora (French monolingual speech NCCFr and the French CS speech FACST corpus) using the LIMSI automatic alignment system with native French acoustic models.

5.2.2.2 Results

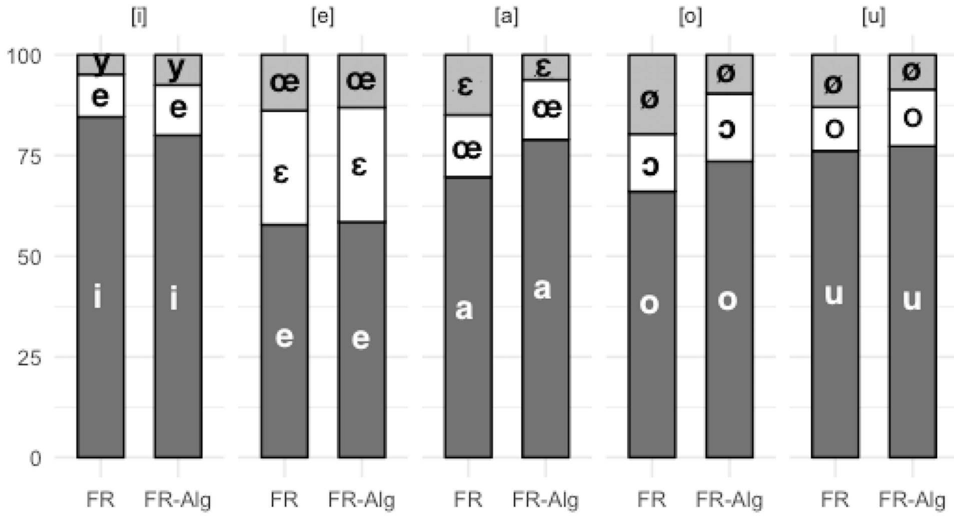


Figure 5.6: Vowel variants in monolingual FR and CS FR (FR-alg). For each target vowel on the x-axis, variant rates of monolingual FR speakers are compared to those of Algerian French (FR-Alg). /a/ corresponds to anterior variants. Variant rates are stacked and sum up to 100%.

Figure 5.6 shows the vowel variant results as a function of the five vowel targets [i, e, a, o, u]. Variant rates are stacked starting from the bottom with the target vowel to which are added the two competing vowels. For each target vowel, the first stack of results corresponds to the French natives (as a reference) before the results of the French CS speech of the bilinguals are displayed. For the /a/ target vowel, only the anterior variants are displayed as they turned out to be the most productive in monolingual French.

Overall, the results of French vowel variation in bilingual speakers' CS and monolingual speakers' speech show similar rates with slightly less variation in CS French for most vowels. In general the cardinal vowels have lowest variant rates for both native and CS French. Mid vowels /e/ and /o/ show higher variant rates as illustrated in the figure 5.6.

With respect to CS speech French, the overall variant rate measured globally is rather similar to the French natives' variant rate. However, the dynamic described above changes. The most stable French vowel in CS speech is [a] with 76% produced as target [a] for the anterior variants (82% in the posterior case). The same score of 76% is achieved by the [i] vowel, followed in decreasing order by [o] (73%), [u] (72%), [e] (57%). It is interesting to note the rather important difference in competing vowels' variant rates for the two mid vowels /e/ (around 45% when summing the two competitors) and /o/ (close to 25% when summing the two) produced in quite comparable rates for both populations (natives and bilinguals).

Although overall competitor variant rates are quite similar between the two speaker populations, we may note that French natives and Algerian Arabic-French bilinguals may prefer different variants. The most frequent variant next to the canonical [o] in French natives is the more central vowel [œ] which was chosen in 23.5% of the cases. In Algerian Arabic-French bilinguals, the most frequent variant after the canonical pronunciation is the mid open vowel [ɔ], which appeared in 18.2% of the occurrences.

For the target [a], a preference for the more central variant in both variant conditions (ante/post) can be observed in the two speaker groups. However, in the displayed "ante" condition, French natives almost equally share competitor variant rates among central [œ] and the more fronted [ɛ] vowel. Whereas a clear preference for the "ante" variants was observed in French natives here, no major difference can be noticed for our bilinguals when comparing ante/post conditions. With respect to [i] and [u], both speaker groups avoid the more central vowel and favour respectively the variants [e] and [o]. The vowel [e] is the only target vowel for which the aligned variants are almost equally distributed between [ɛ] and [œ] for both speaker groups.

Our statistical analysis doesn't show a significant difference of variation across vowels ($\chi^2(10) = 14.07$, $p = 0.17$). All vowel targets are equally likely to allow variants. Furthermore, both groups produce a comparable amount of vowel variants ($\chi^2(2) = 0.64$, $p = 0.73$). In FR CS speech, is as likely as monolingual FR to produce vowel variants. With respect to differences in vowel variant rates by group for each vowel, we observe significant differences only for the [a] target allowing anterior variants ($\chi^2(2) = 9.76$, $p < 0.01$). In this vowel condition, CS speech produces the [a] target significantly more often (74.3 %) than in monolingual FR (54.2 %).

5.2.2.3 Formant analysis

In the following acoustic analysis, the same method to calculate and filter the formants was used as in the former section of vowel variation in CS speech. However, since only the French language is studied in this analysis, we limit the filter values to the French language values cited in (Gendrot and Adda-Decker, 2005).

Formant analysis is illustrated in figure 5.7 for French natives (left) and bilingual speakers (right).

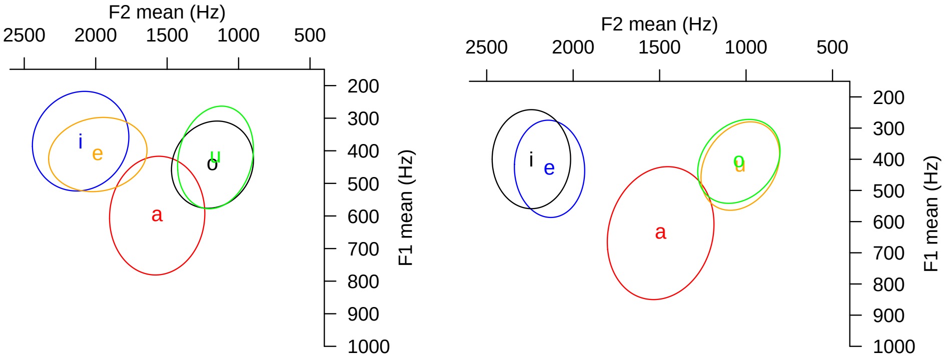


Figure 5.7: /i, e, a, o, u/ formants in FR natives (left) and FR bilingual CS speech (right)

The overall picture suggests that the French vowel quality is somewhat different in monolingual and CS speech productions. Whereas the covered vocalic space is rather continuously occupied in French native speech with different ellipses touching or overlapping each other, we can notice a more extended vocalic space with a clear separation of the high front vowels /i/ and /e/ from the /a/ in CS French from bilingual speakers. The same is almost true for the high back vowels /u/ and /o/, the ellipses of which are just touching the /a/

ellipse with almost no overlap.

Looking in detail at the French CS vocalic representations, the vowel /i/ is realized in a more peripheral and larger vocalic space with higher F1 and lower F2 on average as compared to monolingual /i/ vowels. Also, we notice that the /e/ productions vary with respect to the shape of the ellipses: the ellipse's major axis is horizontal (more F2 variation) in our data of French native speakers, it is vertical (more F1 variation) for our bilingual speakers' CS speech. In both monolingual and bilingual productions, /u/ and /o/ ellipses globally occupy the same locations in their respective vocalic triangles.

Formant analysis globally shows that vowel variation is important both in monolingual and in bilingual speech with strongly overlapping ellipses for high (and mid-high) vowels. Finally, as a general observation, it is interesting to note that the displayed vocalic triangle of the bilingual CS speakers somewhat reminds the Arabic vowel system as it allows to separate three vocalic regions for the five examined vowels.

5.2.3 Vowel centralization in code-switched speech

It has often been observed that fluent, and in particular spontaneous and casual speech gives rise to reduced vowel productions. Such reductions may be highlighted using our ABX-like variant paradigm where each target vowel is given the schwa vowel as competing option.

5.2.3.1 Experiment

In this experiment, we aim at quantifying whether the acoustic realizations of the peripheral vowels tend to move towards the center by simply testing one single competing vowel, the [ə] schwa. In this experiment, we consider all peripheral oral French vowels and also add the three nasal vowels: [i, e, ɛ, a, ɔ, o, u, ɛ̃, ɑ̃, ɔ̃] and the following AA vowels [i, a, u, iː, aː ,uː]. Considering [ə] schwa as the competing variant informs us about vowel reduction, which is also linked to vowel centralization (Delattre, 1969). This experiment informs us whether the schwa variant is particularly productive in CS context (Khattab, 2009) and whether vowels tend to become more centralized.

5.2.3.2 Results

The results in Table 5.5 show that in French monolingual speech, the oral vowel [i] as well as the three nasal vowels [ɛ̃, ɑ̃, ɔ̃] were the least affected by vowel centralization. Less than 20% of the occurrences of these vowels were aligned as [ə]. The vowel [e] was replaced by [ə] in only 20.9% of the cases and can thus be considered as a rather stable vowel as well, whereas [u] was aligned as [ə] in 25.0% of the cases. Over a third of all occurrences of the vowels /ɛ, a, ɔ, o/ were replaced by [ə]. These mid-open vowels were the most affected by vowel centralization in monolingual FR.

For the CS FR, the most stable French vowels were the three nasal vowels [ɛ̃, ɑ̃, ɔ̃] with a centralization rate below 9%. Furthermore, CS speech produced three stable French oral vowels ([i, a, u]) with less than 17% of [ə] substitutions. The most variable vowels in this speaker population were the mid-vowels [e, ɛ, ɔ, o].

Statistical analysis confirmed that vowel centralization in French is vowel dependent ($\chi^2(9) = 28.24$, $p < 0.001$). The vowel [ɔ] is most affected by vowel centralization (29.8 %) whereas the vowel [ɛ̃] is the least affected by centralization (10.7 %). With respect to group (FR, FR CS), no significant differences regarding vowel centralization are found ($\chi^2(9) = 9.50$, $p = 0.39$), although the measured figures show higher variant rates for monolingual than for bilingual speech on average.

Vowel	FR monolingual	FR CS	Δ
i	14.1	12.8	1.3
e	20.9	24.4	-3.5
ɛ	34.1	15.9	14.2
a	34.0	15.9	14.1
ɔ	39.4	20.2	19.2
o	33.5	21.6	11.9
u	25.0	16.2	8.8
ɛ̃	13.6	7.7	5.9
ɑ̃	17.5	8.7	8.8
ɔ̃	17.7	6.5	11.2

Table 5.5: Vowel centralization measured as schwa variant rates (%) CS FR speech and FR monolingual speech. The last column shows the difference between CS FR and monolingual FR

For the AA part of the study, vowel centralization has been studied in a slightly different way because the vowels [e] and [o] do not have a grapheme correspondence in Arabic. Thus we decided to include the following short and long Arabic vowels in the analysis [i, iː, a, aː, u, uː]. The alignment was still carried out with the French acoustic model used for the other experiments. These results, summarized in Table 5.6, show that, globally, long vowels are less often centralized than short vowels. This tendency was confirmed in both reading and CS speech styles.

Vowel	CS	reading
i	37.9	56.5
i:	19.7	15.0
a	49.0	42.4
a:	36.4	26.8
u	41.1	44.7
u:	33.0	24.0

Table 5.6: Vowel centralization (%) for AA in read and conversational CS speech

Globally, centralization of AA vowels is more noticeable in short vowels in read speech and the CS speech. The largest gap in schwa variant rates is measured for read speech between the front vowel [i] with 56.5% of aligned schwa variants and its long counterpart with 15.0%. We may conclude that long vowels achieve more stable realizations. Read speech produced at medium and fast rates tends to boost centralization and schwa production. However, this tendency was not confirmed by the statistical analyses ($\chi^2(1) = 1.01$, $p = 0.32$).

Statistical analyses indicated that vowel centralization is also vowel dependent in Algerian Arabic ($\chi^2(5) = 30.00$, $p < 0.001$). The vowel [iː] is less often centralized (17.4 %) than the other vowels i.e. [i, a, aː, u, uː] (39.2 %). However, speech style i.e. reading and CS do not have a significant impact on vowel centralization in Algerian Arabic ($\chi^2(5) = 7.67$, $p = 0.18$).

5.2.4 Discussion

The methodology we adopted to explore production variation relies on automatic forced alignments permitting specific variants in the pronunciation dictionary. This approach can be seen as an automated approximation of an ABX-like design widely used in perceptual studies to measure the discriminability between categories. We first examined the vowels which are shared in French and AA, namely the three cardinal [i, a, u] vowels in CS speech and in read speech. In this study all the data were produced by the same bilingual speaker set of the FACST corpus. We also examined the variation in FR CS as compared to monolingual French speech in [i, e, a, o, u], as well as vowel centralization in French (both bilingual and monolingual native French speakers) and Algerian Arabic (our bilingual FACST speakers).

When comparing production variation across languages, CS speech showed more variation in AA than in French. Indeed, variant rates are much higher in AA than in FR with about 60% of the tokens aligned as their competing variants ($\sim$ 40% as the target vowel) as shown in Figure 5.2. This higher variability is also corroborated by the acoustic analyses, as the AA vocalic triangle is larger in its global shape and displays also larger ellipses for each /i,a,u/ vowel than for French. This result suggests that the bilingual CS speakers have different production strategies according to the chosen language. In our study, CS speakers adapt to the language and vary their vowels accordingly. When comparing results across speaking styles (spontaneous CS speech vs read speech in either French or AA), we observed, as expected, less variation in read speech as compared to CS speech. However, this is particularly true for French, and less so for AA (See Figure 5.3).

Our results also shows that, overall, in FR CS speech the vowel variability is globally similar to that of the monolingual native speech (See Figure 5.6). However, it may be interesting to note that in French, code-switchers show less variation for [a] and [o] than do native French speakers in monolingual speech. On the other hand, the vowels [i, u] which are most stable in French spoken by French natives are the least stable in CS speakers. In Algerian Arabic, as presented above, only three phonological vowels exist /i, a, u/, which each present a short/long length difference. However, phonetically Algerian Arabic also has the mid front vowel [eː]. The variability observed for [i] in French and Arabic in our data suggests that this mid front vowel is rather a variant of /i/ than /a/ for these CS speakers.

In French monolingual speech, the low variability observed for high vowels, especially [i], might be explained by the density of the vowel triangle, notably for the high frontal vowels [i, y, e], which share a restricted place, hence, low variability of these vowels facilitates intelligibility.

Finally, with respect to the centralization study using parallel variants with [ə] in French, our data showed that the vowel [ɔ] is more often centralized compared to the other target vowels. Similar findings were reported in Boula de Mareüil et al. (2008) where [ɔ] was found to be close to [œ] which is also a central vowel. Our data in French showed, that bilingual CS speakers centralize their French vowels similarly to native French speakers. However, descriptively we find that in French natives, whose vowel productions are generally more variable, the most stable vowels are [i, e] and the three nasal vowels, whereas in CS speech the most stable vowels, next to the nasals, were [i, a, u]. The mid vowels [e, ɛ, ɔ, o] were frequently aligned with [ə]. This might be explained by the Algerian Arabic phonemic vowel system where mid vowels are absent and thus not very important for successful communication. The vowels [i, a, u], on the other hand are important on a phonemic level for Arabic speakers because from a structural point of view. The vowel identity determines the function of the words in Arabic (word classes and verb tenses). Hence, vowel variation replacing a peripheral vowels by central vowels could be fatal for communication and sentence structure.

With respect to Arabic, our results showed that the target [iː] is less often centralized than the other vowels. A possible reason for this result might lie in the extreme position of [iː] in the vowel triangle. In order to investigate this hypothesis, further analyses are needed on more data and more speakers.

In conclusion, our corpus based analysis suggests that in the CS speech and CS speakers are able to vary their vowel productions according to the language they speak, thereby adjusting their productions to the respective vowel systems.

5.3 Consonant variation in code-switched speech

This section deals with consonant variation in CS speech, by focusing on particularities of the consonantal systems in AA and FR languages. As Arabic has a richer consonantal system than French we mainly consider Arabic-specific consonant types and study their production variation.

The investigated consonant variation focuses on three studies. First we investigate consonant gemination in both languages as a speech variation issue in Section 5.3.1. We also study the emphatization (pharyngealization) of the coronal consonants /t,d,s/ and the non-emphatization of AA pharyngealized consonants /$t^ʕ$, $d^ʕ$, $s^ʕ$ / in Section 5.3.2. Finally, we investigate consonant voicing in obstruents that are shared by both FR and AA languages in Section 5.3.3. The experiments are realized with the help of the automatic speech alignment allowing for pronunciation variants. The alignment experiments are accompanied by verification at the acoustic level.

5.3.1 Geminates and gemination in code-switched speech

The following study focuses on production variation of geminate consonants in Arabic and French for which Arabic has a phonological opposition with simple consonants. This may influence bilinguals' production in French where this opposition does not exist.

Experiments are realized with the help of automatic speech alignment authorizing simple and geminate pronunciation variants. The alignment system makes use of Arabic acoustic models which also cover all consonants of French, permitting investigation of simple/geminate variation in both languages.

As for the previous studies on vowel variation, this work relies on two corpora, a French Algerian-Arabic CS corpus of speech from bilingual speakers, and a native French corpus of casual speech by monolinguals. The latter provides a reference baseline for consonant variation in French spontaneous speech. The corpus contains about 31h of conversational French of 46 speakers (24 females) raised in Central/Northern France (Torreira et al., 2010).

In the following, we explore three questions in geminates and gemination variation. First, we study the gemination consonant in CS speech for both French and Algerian Arabic consonants. We study also the gemination in French consonants by comparing code-switched and monolingual speech. Finally, we study the Algerian Arabic geminate simplification in the code-switched speech. The following sections present the alignment experiment, the accompanying acoustic measurements and the results for each part of the study.

5.3.1.1 Alignment experiments

To study consonant variation in CS speech, we used the forced alignment paradigm as described in section 5.1.1. Each word gets one or several pronunciations to handle hypothesized production variants across repetitions and speakers. The acoustic models used to process the bilingual CS data consist in Arabic position-independent monophone acoustic models similar to those described in (Gauvain et al., 2002; Lamel et al., 2004; Gelly et al., 2016; Lamel et al., 2009). The forced alignment system locates word and phone boundaries using the orthographic transcriptions and the best matching pronunciations chosen among the pronunciation variants that are included in its dictionary. Hence, to study geminates and gemination variation, the alignment system is used with a pronunciation dictionary which allows each geminate consonant to be replaced its corresponding simple counterpart and vice versa. In the following, the term gemination, which is generally used to designate a consonant lengthening, is also used to designate the fact that a geminate variant is preferred to its simple consonant counterpart during the acoustic alignment process.

Before describing the planned alignment experiments, Table 5.7 shows the frequency counts of the most frequent geminates in our AA data for which a simple consonant exists both in AA and French. For comparison, we add frequency counts of their simple counterparts in AA and French of the FACST CS speech as well as those of the monolingual French NCCFr corpus. Although the coronal /r/ has its corresponding geminate among the top five most frequent geminates, we did not include it in this study, as it is not shared phonetically with the French consonants. Also, the shared FR-AA uvular consonant /ʁ(χ)/ which presents also high frequencies in both languages is not included in this study. Indeed, the allophonic variation of FR /χ/ implies complex phonological issues and the phonological adaptation of FR words containing this consonant in the Arabic lexicon shifts to the rhotic /r/ (Lahrouchi, 2018). Examples: FR: "garage" /gaʁaʒe/, AA: /garaʒe/.

It is interesting to note, that the most frequent geminates in Table 5.7 are all coronals /ll, dd, nn, ss/.

We start with a quantitative overview of geminates in the AA part of the FACST corpus in order to get more information about the geminates in the AA dialect. Table 5.7 displays all consonants for which the corresponding geminate has more than 120 tokens each. We limit our investigations on these consonants, which are the most representative in our data.

Cons	Number of occurrences			
	FACST			Nccfr
	AA	AA	FR	FR
	geminate	simple	simple	simple
/l/	334	2820	7268	59227
/d/	141	950	4609	43497
/n/	138	2118	2867	27269
/s/	126	456	7072	77680

Table 5.7: The most frequent AA geminates (75% of geminate tokens) in the FACST corpus. Occurrence counts for their simple counterparts, in AA/French CS data and in French NCCFr corpus.

The first experiment will investigate gemination in CS French-AA speech, measuring consonant production variation using the Arabic acoustic models. Table 5.8 describes the target consonants and parallel variants, i.e., the simple and geminate form for each consonant, and provides examples for both languages. In the second experiment, the same protocol is applied to the French language for two speaker populations, in order to compare the production of French consonants in CS speech with monolingual speech. Hypothesizing geminate variants in French which is a language with no phonological gemination, aims at questioning whether our French data include consonants that are produced with acoustic features which make them look like geminates. The third experiment attempts to answer question whether, and if yes, how often geminates simplified in CS spontaneous speech. In this case the target consonants are the AA geminates /ll, dd, nn, ss/ and their simplification are permitted as variants. This experiment locates differences between the geminates transcribed orthographically in the FACST corpus and parallel variants chosen during automatic forced alignment.

Target	Competing variant	Examples
[l]	[ll]	لِ /li/ (for) : [li], [lli] (AA) lu /lu/ (read) : [ly], [lly] (Fr)
[d]	[dd]	دَار /da:r/ (house) : [da:r],[dda:r](AA) dent /dã/ (tooth) : [dã],[ddã](Fr)
[n]	[nn]	نُور /nu:r/ (light) : [nu:r],[nnu:r](AA) ne /ne/ (light) : [nu:r],[nnu:r] (Fr)
[s]	[ss]	سَار /sa:r/ (walked)[sa:r],[ssa:r] (AA) sept /set/ (seven)[set],[sset] (Fr)

Table 5.8: Competing geminate variant for each simple target consonant and example lexical entries.

5.3.1.2 Acoustic measurements

A major acoustic cue of gemination, also termed consonant lengthening, is an increased acoustic duration, resulting from the consonant's articulation using a longer timespan than usual for a simple consonant. For this reason, we also provide average consonant duration results. The durations are measured using the automatic segmentations produced during the phone segmentation in the automatic alignment. We thus may hypothesize longer durations in our spontaneous speech data for consonants labelled as geminates. Different gemination variant results may be expected in the two languages. One might hypothesize lower gemination variant rates in AA as gemination is phonologically distinctive here, but not in French.

5.3.1.3 Results: gemination in French-AA CS speech

Figure 5.8 displays the gemination variant rates for the coronal consonants /d,n,s,l/ for both AA (left) and FR (right). The results show that globally quite similar gemination variant rates are achieved for both languages: 22.4% for AA and 22.2% for Fr. However, we may note small differences across consonants. For AA, the consonants /s,l/ have the highest variant rates of about 25%, whereas only /s/ has a similar value for FR. The two highest gemination rates in French are measured for /s,d/ with 25% and 21% of the variation rate. The consonant with the lowest geminate variant rates is the nasal /n/ for both languages, with 19% and 16% of occurrences for AA and French respectively.

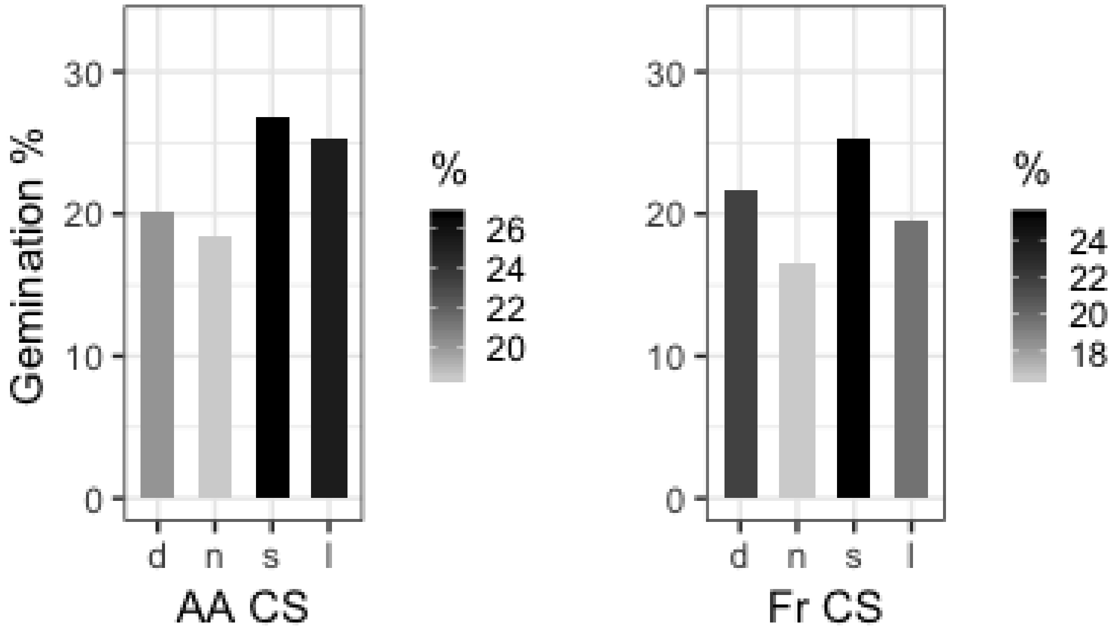


Figure 5.8: Consonant gemination rates as measured for (simple) target consonants in AA and FR CS speech.

Figure 5.9 reports average duration results of the target (simple) consonants separated into two populations according to the gemination alignment results. The tokens that were aligned unchanged as the simple target are represented by a triangle, whereas the tokens aligned as geminates are represented by a circle.

Figure 5.9 shows that, with the exception of /s/ in AA, the consonants most frequently labeled with the geminate variant have a longer duration difference between the simple and geminate forms. Trying to relate gemination variant rates with consonant durations, we may observe a correlation (*r*= +0.67) between the variant rates and the corresponding durations.

5.3.1.4 Results: gemination of French consonants in bilingual and monolingual speech

Comparing the gemination rates of the French consonants in Figure 5.10, the monolingual speech present more variation than the CS one ($\chi^2(2) = 8.01,\ p < 0.01$). This suggests that the phonological gemination contrast plays an important role in keeping the canonical

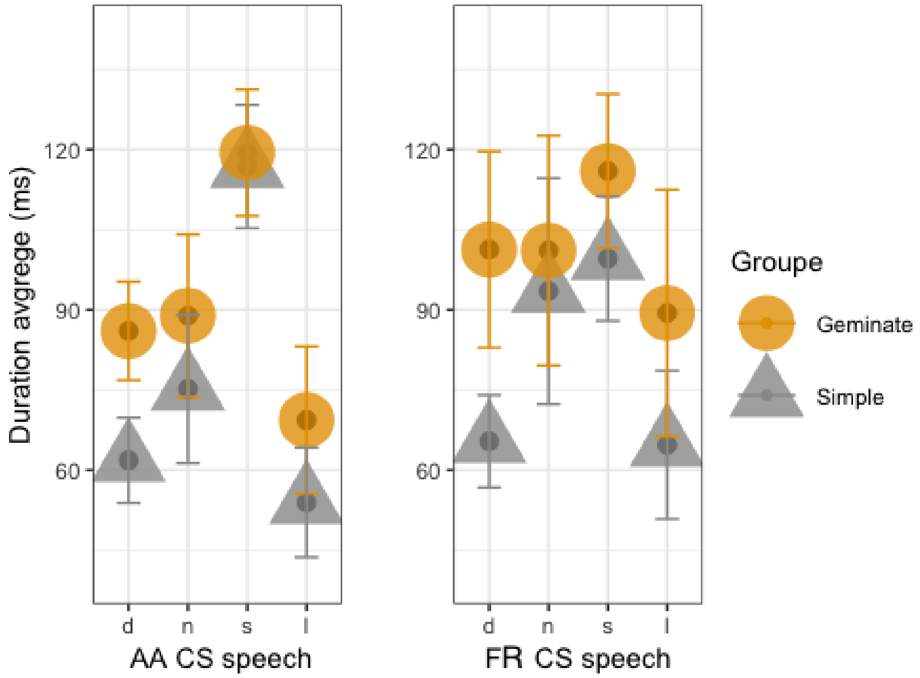


Figure 5.9: Average duration in (ms) of simple consonants AA and FR CS speech. Circles: geminate variant selected ($C \rightarrow CC$); triangles: remains simple ($C \rightarrow C$). Error bars give standard deviation.

pronunciation. In this experiment the word and the prosodic contexts of the gemination have not yet been studied. This high variation in monolingual speech suggests that these segments may have been accentuated.

5.3.1.5 Results: Algerian Arabic geminate simplification in code-switched speech

Figure 5.12 shows the percentage of segmented tokens with simplification of the original geminates of AA and the corresponding duration plots. The simplification rate is seen to be larger than the gemination rates in the previous figures which was on the order of 20%. Simplification is observed for all consonants, with the rates varying from 49% to 76%. Simplification accounts for the largest percentage of the variation, as a simplification rate of 76% for /ss/ 'S' in Figure 5.12. The duration plot shows that at the global acoustic level, the duration of tokens transcribed with their simple counterpart is significantly less than those which retained the geminated form.

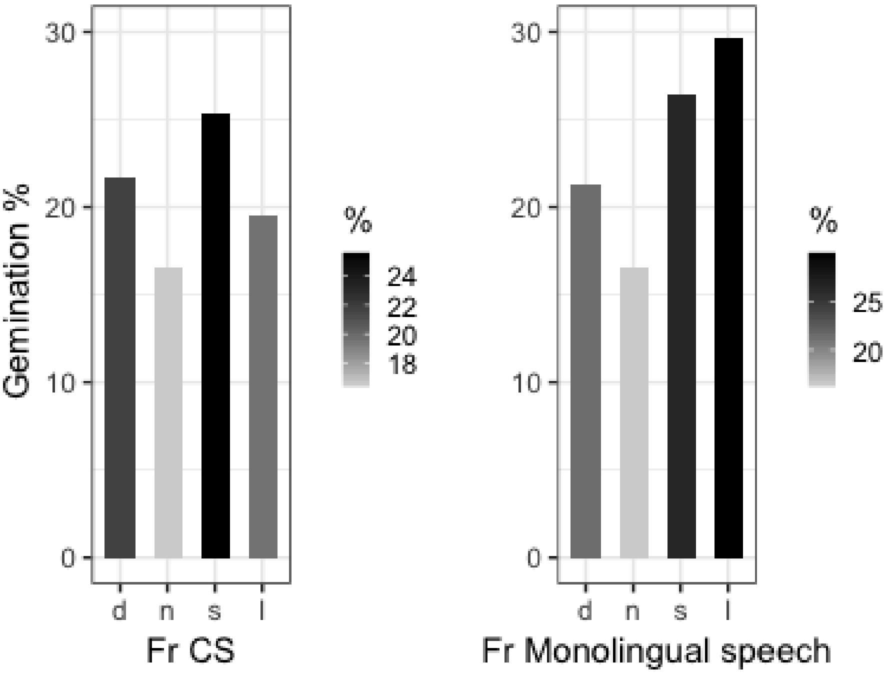


Figure 5.10: Expt 2: Gemination rates of simple consonants CS and monolingual French speech for each target consonant.

5.3.1.6 Discussion

Three points can be mentioned based on this study. First, the proposed method using automatic variant alignment can help us study the variation of simple and geminate consonants in large speech corpora. The duration analysis confirms that the aligned gemination and simplification variant labels are highly related to segment duration and that duration is a solid, although not unique, criterion to study variation in consonant gemination.

The study also shows that gemination of simple consonants, as revealed by our method, appears in both FR and AA CS speech. However, AA is the most affected by this variation despite the phonological distinction between simple consonants and geminates. In our data, the consonants most concerned by this gemination variation are /d, s, l/. By contrast, lowest amount of gemination was observed for the nasal consonant /n/ in both languages and in both corpora. The FR monolingual speech also shows high gemination variant rates, comparable to those for the FR parts in CS.

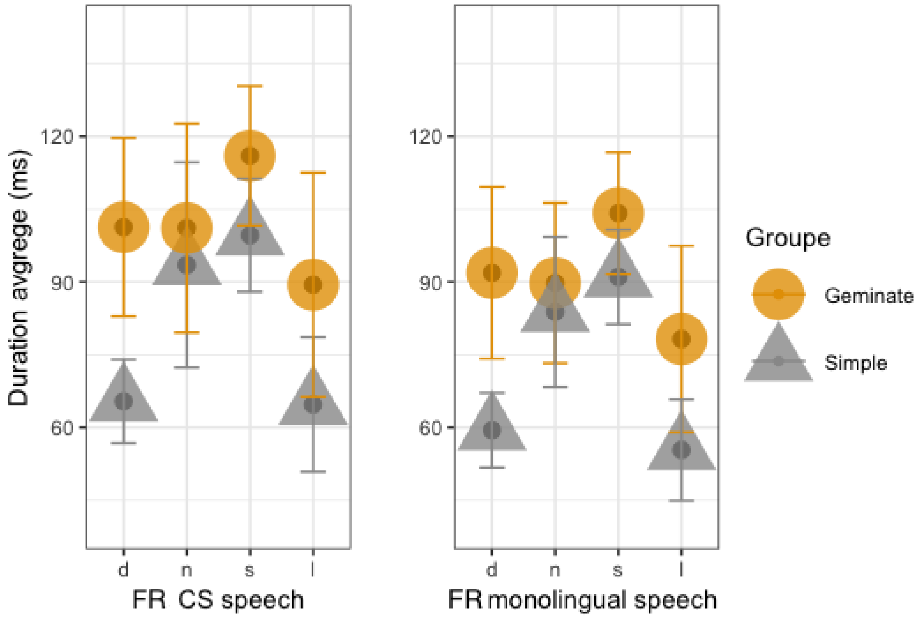


Figure 5.11: Average duration (in ms) of French simple consonants in CS and monolingual speech. Circles: geminate variant selected ($C \rightarrow CC$); triangles: target selected ($C \rightarrow C$) Error bars give standard deviation.

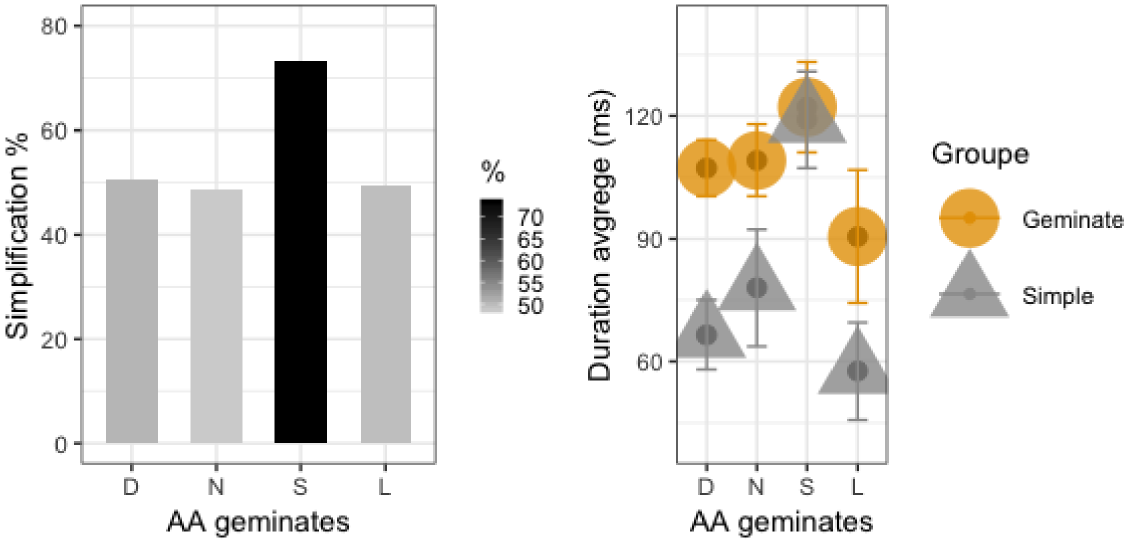


Figure 5.12: Left: Variation rate of simplified geminates. Right: Average duration (in ms) for simplified variants given by triangles ($CC \rightarrow C$) and retained geminates (circles correspond to the $CC \rightarrow CC$ variation).

Finally, the high simplification rates of geminate consonants ($> 40\%$) suggest further investigations on a methodological level: acoustic models may be biased in favor of simple consonants due to their overwhelming presence in speech. On a linguistic level, geminates may feature other correlates than duration. Further studies are required in an attempt to understand production differences by monolingual and bilingual speakers. The relatively high rate of gemination variants ($\sim$20%) in AA speech might be partly due to the post-lexical gemination which is not yet accounted for in the alignment system.

5.3.2 Emphatics and consonant pharyngealization in code-switched speech

Within the perspective of investigating on the influence of emphatic consonant production in AA on French production in CS speech, we analyze in this work three questions related to consonant variation in pharyngealization (emphatization). With the use of automatic speech alignment with allowing emphatic or non-emphatic variants, we investigate first on AA variation production of the three consonants /t ,d, s/ in their emphatic consonants counterpart, and the variation of the emphatic consonants /$t^ʕ$, $d^ʕ$, $s^ʕ$/ in their plain consonant counterpart. Secondly, we study consonant variation of /t, d, s/ in emphatic consonants of both AA and FR in CS speech. In the third part, we experiment consonant emphatization of plain consonant /t, d, s/ of FR production in CS speech and we compare the results with FR monolingual speech variation.

We also support the alignment results with acoustic analyses of the consonant variations with formant analysis of adjacent vowels in order to link the acoustic results to the alignment outputs.

This work is therefore a part of the questions of emphatic consonants influence in CS productions variation in AA and FR languages.

5.3.2.1 Alignment Experiments

Similar methods then the gemination and voicing experiments are followed to align the target consonants with the parallel variants (See sections 5.1.1, 5.3.1). Three experiments are developed according to the emphatics and emphtaization consonant variation: (a) variation of emphatics and non-emphatic consonants of AA in CS speech, (b) variation of the consonants /t, d, s/ in their pharyngealized counterparts in FR and AA CS, (c) comparison of consonant emphatisation variation of /t, d, s/ in CS FR speech and FR monolingual speech. The table 5.9 summarizes the computing consonant variants in each experiment and the corpora used.

Expts	Target	Parallel variants	Lang & speech	Corpora
(a)	/t/, /$t^ʕ$/ /d/, /$d^ʕ$/ /s/, /$s^ʕ$/	[t, $t^ʕ$] [d, $d^ʕ$] [s, $s^ʕ$]	AA CS	FACST
(b)	/t/ /d/ /s/	[t, $t^ʕ$] [d, $d^ʕ$] [s, $s^ʕ$]	AA/FR CS	FACST
(c)	/t/ /d/ /s/	[t, $t^ʕ$] [d, $d^ʕ$] [s, $s^ʕ$]	FR CS FR Monolingual	FACST NccFr

Table 5.9: Computing consonant variation of consonant emphatization variants and variation of emphatic consonants

5.3.2.2 Acoustic measurements

Formants have been calculated from the acoustic signal and the acoustic data have been aligned with phonemes segmentations (phone alignment) in order to extract the acoustic data of the concerned phonemes.

To investigate on acoustic clues of emphatic and emphatisation consonant variation, adjacent vowels in CV context of the consonants /t, d, s/ and the emphatics /$t^ʕ$, $d^ʕ$, $s^ʕ$/ has been extracted from the audio speech by computing their formant frequencies. An automatic formant extraction, with Burg algorithms implemented in Praat program (Boersma et al., 2002), has been performed on the acoustic signal and an alignment of the formant values with the transcribed vowels has been realized (Gendrot and Adda-Decker, 2005). The formant measurements were taken on three parts of the vowel segment: the beginning part, the middle part, and the end part. The formant average represents the sum of these three parts. Following the acoustic correlates of the emphatics and pharyngealized consonants variants, the transition part between the consonant and the vowel formants (F1, F2, F3) is more salient. Thus, We take into account the first part of the formants that is positioned just after the consonants, i.e. formants onset. We focus in this analysis on F2 values that correspond to the respective alignment experiments. Then, we obtain F2 onset values that follows the target consonants.

With the respect of the experiment (a), we calculated all the F2 onset values of the right-adjacent AA vowels /i, iː, a, aː, u, uː/ of /t, d, s/ and /$t^{ʕ}$, $d^{ʕ}$, $s^{ʕ}$/. Concerning the experiment (b), we calculated all the F2 onset values of the right-adjacent AA and FR vowels: followed /i, iː, a, aː, u, uː/ of AA /t, d, s/ and followed vowels of FR /t, d, s/ consonants. The experiment (c), we compare the FR F2 onset of the followed vowels of /t, d, s/ of CS with F2 onset in monolingual French.

It should be noted that we measured the acoustic analysis for all parts of the emphatic variation study. However, the method presents some difficulties to analyse the total of the target consonants and the formants values. Indeed, referring to the formants of the following vowels excludes a part of the studied consonants in VC, CCV contexts. Thus, the quantity of the data is reduced and the the consonant targets processed by the alignment experiments are not fully processed. Moreover, the extraction of the formant onset by using the automatic segmentation in phones may gives a inaccurate values due to the boundaries. The onset part of the vowel segment often include a transitional part which can impact the formal values contrary to middle part which is often more stable.

Following these acoustic details which requires more depth in analyzes that we do not address in this thesis, we present the global results of F2 onset of the followed vowels of the target consonants results by grouping all vowels. It should be remembered that, the purpose of these analyzes is to verify the alignment results and obtain acoustic information about the alignment results.

5.3.2.3 Results: emphatic and non-emphatic variation in AA

In this section, we present the alignment results of the first experiment (a) that focuses on the emphatic and their plain consonant counterpart variation in AA. Thereafter, we present the global acoustic results of the parallel experiment.

The emphatic and non emphatic variants are represented in percentage rates in two groups as shown in Figure 5.13. The first group shows the emphatic variants /$t^{ʕ}$, $d^{ʕ}$, $s^{ʕ}$/

that vary in /t, d, s/ and the second group illustrates the plain consonants that vary in emphatics. Automatic alignment results show variation in both emphatic and non emphatic AA consonants as shown in the figure. However, the alignment results indicate that the emphatic consonant are highly realized in their plain counterpart. Indeed, the simplification of the emphatic consonant /d^ʕ/ reached $\sim$60% and the fricative /s^ʕ/ vary in /s/ with 43.6%. The lowest rate of variation in this consonant group is the emphatic voiceless stop /t^ʕ/. In parallel, the emphatisation of the plain consonant /t ,d, s/ in CS AA speech represents respectively 28%, 20% and 22%. So the rates are close to each other.

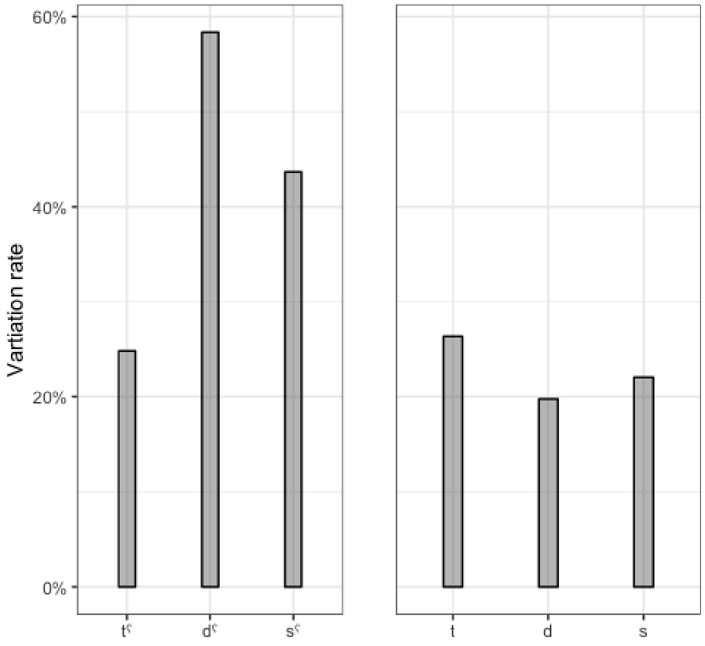


Figure 5.13: Emphatic and simple consonants variation rates in AA CS speech. The bars represent the percentage of the parallel variant.

What we learn from these alignment results is the emphatic consonants of Arabic are highly simplified in CS production and the non-emphatic consonants vary to emphatics. The consonants /d^ʕ, s^ʕ/ presents a height variant rate with 58% and 44%. The consonant that varies the least is /t^ʕ/. The plain consonants show less variation Statistical analysis shows that there is a significant difference between the emphatic variation and emphatization of the plain consonants $(\chi^2(6)\ 4.31,\ p < 0.001)$.

Concerning the acoustic analysis, Figure 5.14 presents the F2 onset of the followed vowels of the target emphatics /$t^ʕ$, $d^ʕ$, $s^ʕ$/ and simple consonants /t, d, s/ in AA CS speech. The F2 values in the box-plot includes all AA vowels.

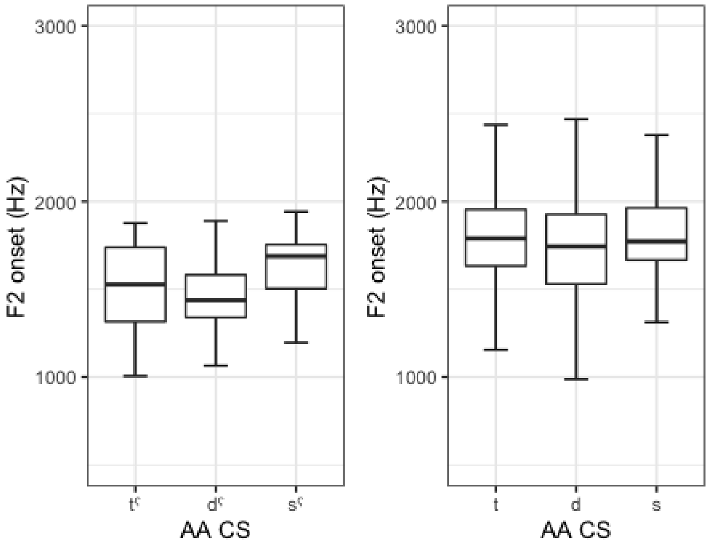


Figure 5.14: F2 onset of followed vowels in AA emphatics /$t^ʕ$, $d^ʕ$, $s^ʕ$/ and their counterpart /t, d, s/ in CS speech. The box-plot represents the quartiles and the center line corresponds to F2 median. Error bar corresponds to the standard deviation.

By comparing the emphatics consonant and their simple counterpart, we observe that the F2 is mostly lower in emphatics behind the plain consonants 1400-1650 Hz for emphatics and 1700-1800 Hz for plain consonants. However, the figure demonstrates the values are not equal in the target consonants and the variation of F2 in each consonant do not correspond to emphatization and simplification variation rates shown in the alignment experiment.

5.3.2.4 Results: consonant emphatization in CS speech

This section present the automatic alignment the results of /t, d, s/ emphatization of AA and FR in CS speech. We also compare the results with read speech in order to get more information about the impact of CS on consonant emphtaization. Thereafter, we present the global acoustic results of the parallel alignment experiment.

Figure 5.15 shows /t, d, s/ consonant variation in their emphatic counterparts, the left part represents CS speech variation and the right one represents variations in read speech that we used for comparison. In the first CS part, we observe simple consonants which vary in emphatic consonants by comparing French and AA, emphatization of /t, d, s/ is mostly superior to the French one as shown in the figure 5.15. However, the variation rate in both languages is similar in the order of each phone, /d/ is the lowest in both languages and /t/ is the highest in both languages. In first position, we observe that the/t/ varies to /t^ʕ/ and reaches a variation rate of 26% for AA and 16% in FR. Then we observe in second position the fricative /s/ which varies to /s^ʕ/, with a variation rate of 22% and 14% for AA and FR. Finally, we observe the d which has the lowest variation rate in CS with a score of $\sim$20% in AA and 11.5% in AA. Thus, consonant emphatization in CS speech is more important in AA than in French, the variation rate gap between AA and French is of 8%. This pushes

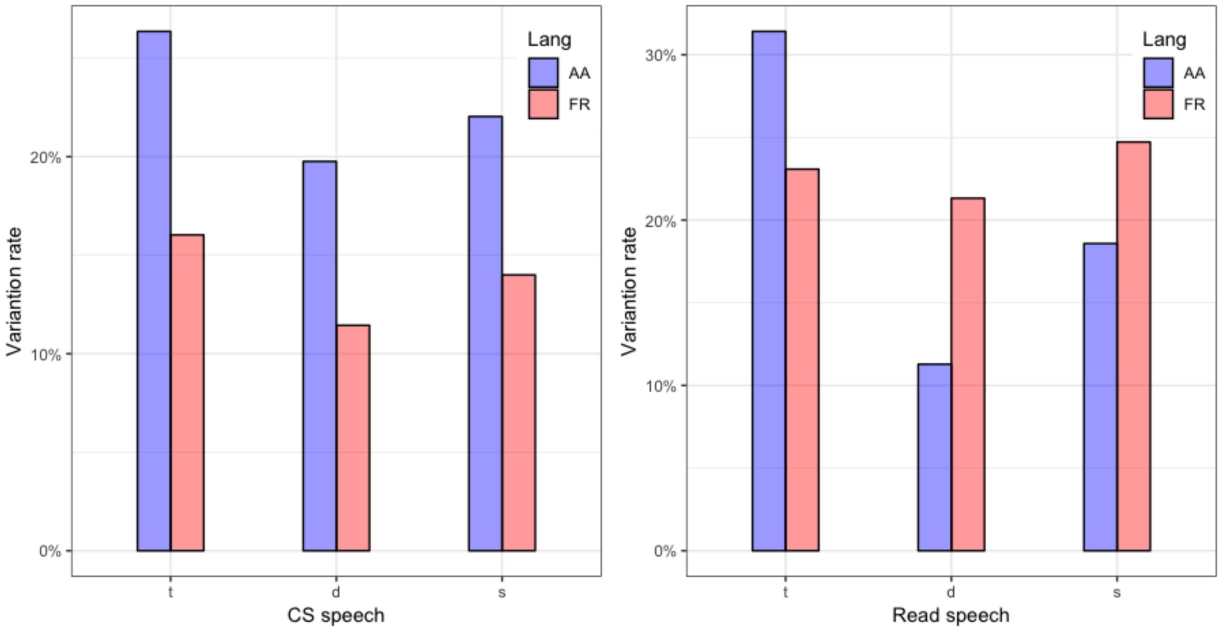


Figure 5.15: Emphatization of plain consonants /t, d, s/ in CS speech and read speech. The blue bars correspond to AA speech and the red ones corresponds to FR speech The bars represent the percentage of the parallel variant.

us to ask the question of phoneme production, in fact phoneme production of /t, d, s/ is less stable in AA than in French in CS speech $(\chi^2, df = 5,\ p = 0.30622)$. When we observe the results of variants /t, d, s/ in read speech, we notice that the variation rate in French an AA, /t/ has also the highest variation rate and d has also the lowest variation rate. However, the variation rate is higher in French compared to CS speech, in fact it reaches 23.5% for the French /t/ and 21% for the French /d/ and ∼25% for the /s/. Thus, we can affirm that these consonants emphatization is higher in French read speech compared to French CS speech. However, AA remains the language in which consonant emphatization is higher. Compared to the read speech, CS emphatization in FR is lower than in AA but also lower than in FR read speech variations. This means that CS allows a high variation for AA and a low variation frequency in FR.

Concerning the acoustic analysis, Figure 5.18 presents the F2 onset of the followed vowels in consonant targets /t, d, s/ in AA and FR CS speech. The F2 values in the box-plot includes all vowels of each language.

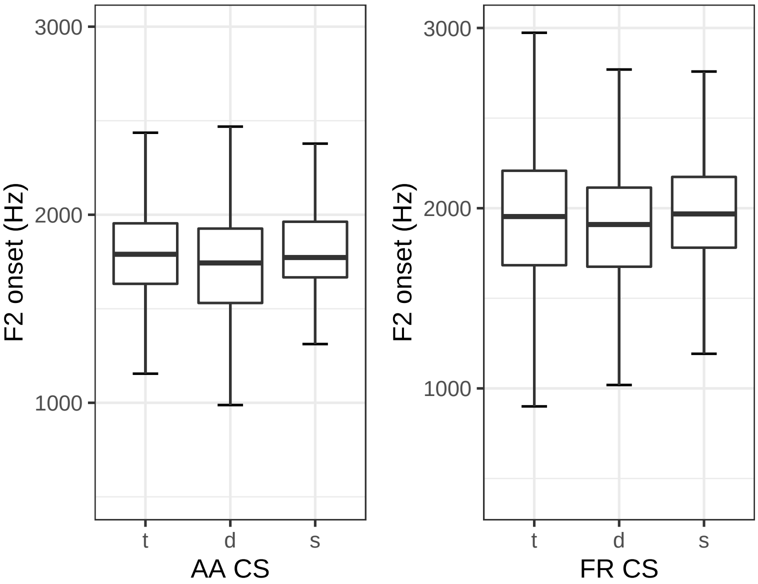


Figure 5.16: F2 onset of /t, d, s/ followed vowels in AA and FR in CS speech. The box-plot represents the quartiles and the center line corresponds to F2 median. Error bar corresponds to the standard deviation.

The figure gives global values of F2 onset. The F2 of the consonants shows that the languages have different formant onset values next to the same target consonants. Indeed, we observe that the F2 is lower in AA consonants with an average of 1750 Hz comparing to the FR consonants with 2000 Hz. It should be noted that the F2 onset results follow the automatic alignment results. Indeed, the global lowering of F2 onset in AA that may correspond to more emphatization of the consonants as a variation.

5.3.2.5 Results: French emphatisation CS and monolingual speech

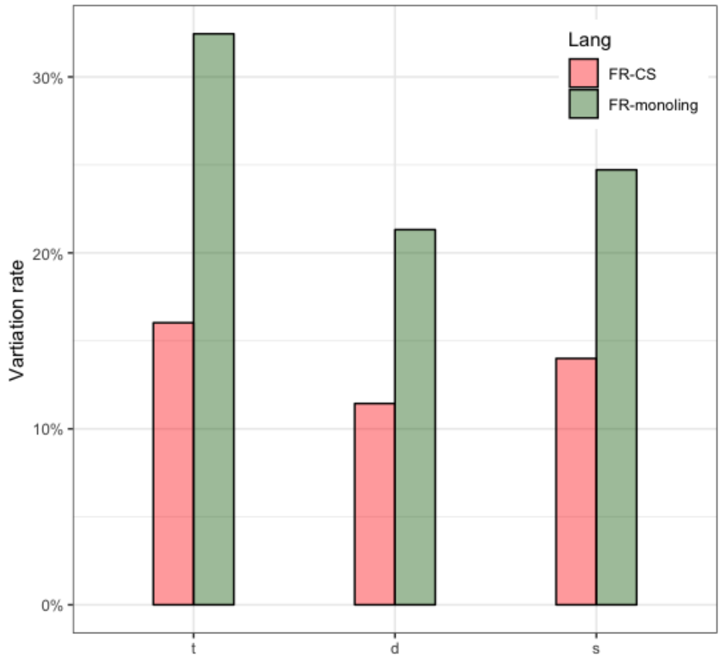


Figure 5.17: FR emphatization variant rates of /t, d, s/ in CS speech and Monolingual speech. The red bars corresponds to FR CS speech and the green ones corresponds to The FR monolingual speech. The bars represent the percentage of the parallel variant (emphatics).

This section presents the last experiment in emphatic variation. We compare the French consonants emphatisation variation with the FR monolingual speech results.

By comparing the variation rate in FR CS speech with FR monolingual speech, we notice that the variation rate is much more higher in monolingual FR speech as shown in the figure 5.17. We also notice the same order in consonants variation, /t/ remains with the highest variation rate, however monolingual FR speech has the highest emphatization rate with 32% for the /t/ which varies to /tˤ/, /s/ which varies to /sˤ/ with a variation rate of 29% and finally with the consonant /d/ in last position which reaches a variation rate of 21%. The variation rate gap in FR between monolingual and CS speech is of 12%.

The acoustic analysis are resumed in Figure 5.18. It presents the F2 onset of the followed vowels in consonant targets /t, d, s/ FR CS and FR monolingual speech.

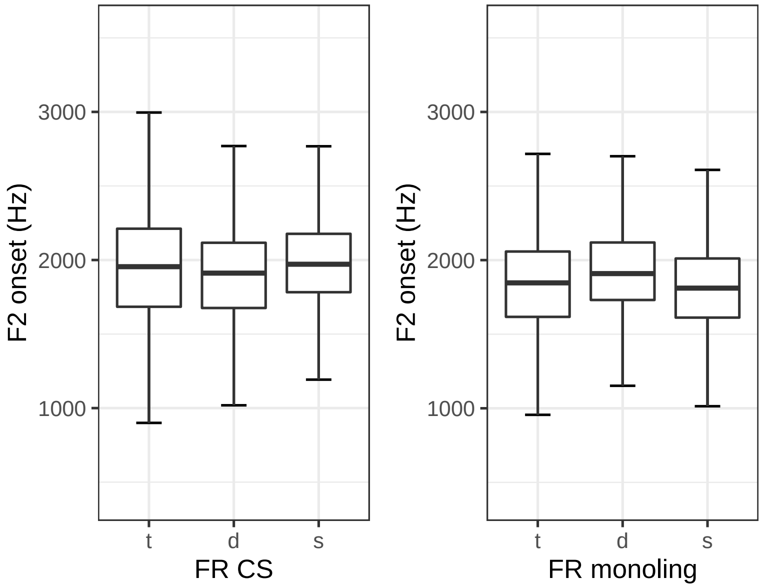


Figure 5.18: F2 onset of /t, d, s/ followed vowels in FR CS and in FR monolingual speech. The box-plot represents the quartiles and the center line corresponds to F2 median. Error bar corresponds to the standard deviation.

The figure gives global values of F2 onset. The F2 of the consonants informs that the speech types affects the formant onset values which are next to the same compared consonants. Indeed, we observe that the F2 is lower in FR monolingual consonants with an average of 1850 Hz comparing to the FR CS consonants with 2000 Hz. Then, the F2 onset results follows the automatic alignment results. The global lowering of F2 onset in FR monolingual speech may correspond to more emphatization of the consonants as a variation observed in the emphatization of FR consonants observed in the acoustic results.

5.3.2.6 Discussion

This study investigated the emphatization in AA and FR in CS speech as consonant variation in production by using automatic speech alignment with variants. From this main question, experiments that compared the consonants produced in CS speech, read speech and monolingual speech has been performed. Overall, alignment results conclude that plain consonants in FR-AA CS spontaneous speech tend to emphatization. However, by comparing FR in CS and FR in monolingual speech, the results conclude that the emphatization is highly noted in monolingual speech then the CS speech. The results show also that the AA emphatic simplification represents 42% of the variation rate in CS speech.

With the respect of the consonants /t, d, s/ emphatization, AA CS get variation at a rate of 25%, however, in CS French spontaneous consonants get a variation at a rate which is lower than the AA rate (FR CS: 16%). The alignment results allow us to conclude the French consonants emphatization has a very high rate in monolingual speech with an average of 26% of variation. Then, the CS of the bilingual speakers may have in impact to control /t, d, s/ production and allows to produce less variation then the FR monolingual speakers. The acoustic analysis showed that the lowering of F2 formant of the followed vowel corresponds globally to the alignment results.

It should be noted that French has the posterior vowels /o, ɔ, ɑ/ that could have an impact on consonants emphatization in CV context, especially with /t, s/ such as /tordre/, /sobre/ (consonants that represent a high emphatization rate in the alignment). This variation in production can be one of the factors that increases the high percentage of variation in the alignment experiments. Figure shows examples an emphatization of /t/ followed by the posterior vowel in FR monolingual speech and Fr CS speech.

Following the alignment results, the emphatic consonants in AA CS vary highly to simplification in their plain consonant counterparts. The examples shows the.

5.3.3 Obstruent voicing variation in code-switched speech

In this study, we examine the possible phonetic variation in the way voiced and voiceless obstruents may be produced by speakers in FR and AA during the CS speech. The study is divided in tow parts. First we investigate the occlusive variation and we investigate the fricative variation. The voicing and devoicing variation analysis are also investigated at the word position level. Through this study we focus on the following questions: with the use of automatic alignment, to what extent the voicing and devoicing of AA and FR obstruents in CS speech? In which word positions the voicing variations are more marked?

5.3.3.1 Alignment experiment

In this study, we focus on consonant voicing variation in CS FR-AA speech. The investigation is presented in two obstruents parts, the shared stops and fricatives of FR and AA. Variation experiments using forced alignment with variants are applied on the voicing of the voiceless stops and fricatives /p, t, k/, /f, s, ʃ/. The alignment experiment is also organized on the devoicing of the voiced stops /b, d, g/, /v, z, ʒ/. The aim of this study is first to observe the voicing and the devoicing of the consonants in the CS speech and reporting about the most and the least consonants affected by this variation in both languages. Also, the study tempts to highlight the influence of one language to another in the voicing change of the consonants. Indeed, although the opposition voicing/devoicing of the stop consonants is pertinent in both languages, the stops /p/ and /g/ in AA come from foreign languages and they are produced mostly in loan words with low occurrences (See 4.5.1). So, we ask the question about the impact of theses phonological differences between FR and AA on the voicing productions of the stop consonants during the CS speech. Also, our hypothesis expects that the voicing consonant vary in word position (initial, intern, final) and the variation is more frequent in word final position.

A forced alignment with pronunciation variants has been applied on FR and AA parts of the FACST CS speech. The acoustic model used to align the variants is the Arabic model which groups all the consonants of both languages. The system allows, from the given acoustic data of the speech signal, to the target stop consonants to give alignment with canonical pronunciation of the consonants or voicing variants. The voicing and the devoicing variation in this experiment is realized with the stop consonant opposition voicing. Example: in the French word "barrage" [baʁaʒə], the FA system displays the devoicing variant [p] of the voiced stop [b] [paʁaʒə]. Table 5.10 resumes the consonant targets and the variant experiments with word examples.

Target	Competing variants	Examples
[p]	[p, b]	پلَاصَة /plaːsa/ (place) : [plaːsa], [blaːsa] (AA) apte /apt/ (capable): [apt], [abt] (FR)
[t]	[t, d]	مَات /maːt/ (dead) : [maːt],[maːd] (AA) tour /tuʁ/ (tour): [tuʁ],[duʁ] (FR)
[k]	[k, g]	كُرسِي /kursi/ (chair): [kursiː],[gursiː] (AA) document /dokymã/ (docuemnt): [dokymã], [dogymã] (FR)
[b]	[b, p]	بَرَّا /barra/ (outside): [barra],[parra] (AA) bar /baʁ/ (bar): [baʁ], [paʁ] (FR)
[d]	[d, t]	بَرد /bard/ (cold): [bard],[bart] (AA) dormi /dormiʁə/ (sleep): [dormiʁə], [tormiʁə] (FR)
[g]	[g, k]	گَاع /gaːʕ/ (all): [gaːʕ], [kaːʕ] (AA) bague /bagə/ (ring): [bagə], [bakə] (FR)

Table 5.10: Variants experiment summary for each target stop consonant /p, t, k, b, d, g/ with examples of lexical entries.

5.3.3.2 Acoustic measurements

In order to complete the alignment experiment results, parallel acoustic investigation about the consonant voicing has been realized.

From the speech signal, we extracted from F0 the proportion of consonant voicing (*v-ratio*) following the methods proposed in) (Snoeren et al., 2006; Hallé and Adda-Decker, 2007; Kiss, 2013) with the help of Praat program (Boersma et al., 2002). This measure reports the ratio of the total voiced frames of the consonant. The v-ratio is measured on a percentage from 0% to 100% for each consonant. A v-ratio of a full voiced consonant is equal to 100 and the full voiceless consonant is equal to 0%. However, voiced and voiceless consonants, depending on the consonant production and voicing contexts, can have a partial v-ratio. Hence, we can have voiced consonants with a v-ratio more to 50% and voiceless consonants with a v-ratio less than 50%. These voicing measurements are applied on the target consonants in order to investigate consonant variation acoustics referring to the speech signal. In order to achieve this purpose, the v-ratio is calculated referring to the stop and fricative consonants position in the word (initial, intern and final) for both FR and AA in CS speech.

5.3.3.3 Results: stop consonant voicing variation

We calculated the total percentage of alignment results of voicing and devoicing of the target consonants in three word positions (initial, intern and final position). We present the results in two parts, the first part deals with CS speech, and second part with read speech which is used as a speech monolingual reference of our twenty speakers. In CS speech we present in part voicing and devoicing results of stop consonants as shown in Figure 5.19. This figure shows that consonants voicing and devoicing rate of AA stops is higher than in FR consonants with a voicing rate from 12 to 25% for all /p, t, k/ consonants. With the respect of AA the voicing percentage of /p, t, k/ consonants is less high than the devoicing stop of /p, d, g/. FR also shows more voicing and devoicing with a percentage that varies between 5% and 15% for voicing and between 15% and 25% for devoicing. When we compare voicing variants according to consonant position within the word, we notice that it is the final position that has a high score in both languages, with an average voicing rate in /p, t, k/ of 25% for AA in final position and 16% in initial and intern position. FR voicing variants reach 14% at its high variation score in final position. Concerning stop devoicing, it reaches its highest score in final position in AA with a rate of 30% for /b/ and 48% for /d/. The consonant /g/ has no final devoicing variant, but it could be related to the low occurrences of this consonant (See section 4.5.1 and number of occurrences in Appendix E.2). Final devoicing in FR has also a high score. FR stop consonant that has the highest rate of devoicing is /g/ with a score of 24%. FR stop that varies the less in all positions is /b/ with a score of 14%, /d/ is second less varying consonant with a score of 23% in final position. To sump up, in CS speech voicing and devoicing variation is higher in

AA than in FR and in both languages, devoicing in final position has the highest rate, thus consonants subject to voicing and devoicing are more often in final position. The consonant that is more subject to devoicing is /g/ in FR and /d/ in AA.

Concerning stop voicing variation, /p/ has the highest score in both languages. The phonetic realization of /p/ could explain its voicing variation in AA, /p/ tends to be produced as a [b] and since there is no minimal distinctive pair, it is considered as a free variant, example: the word پُولِسِيَّة [pu:lisiyya:] [bu:lisiyya:] *police*.

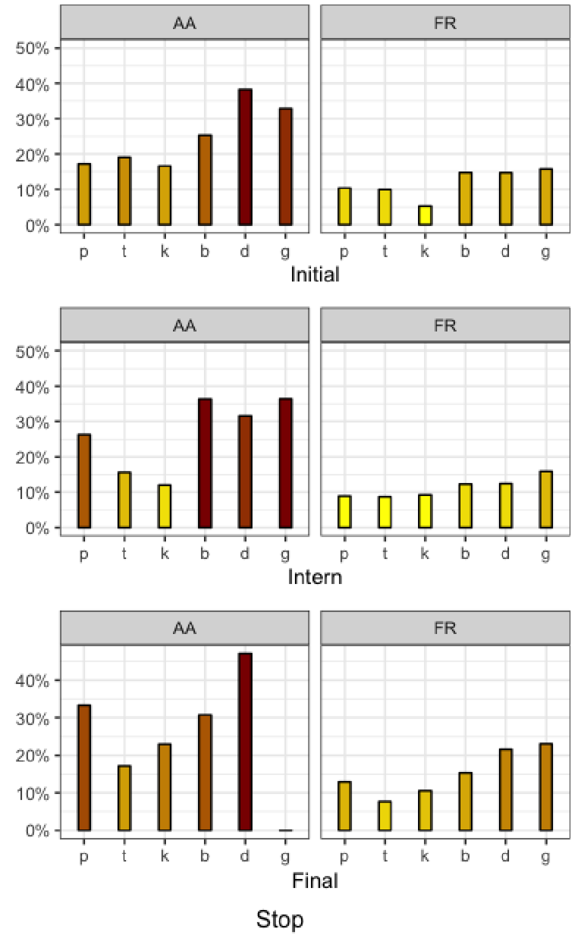


Figure 5.19: Results of stop voicing variants of FR and AA in word positions (Initial,intern, final) in CS speech. The *x* axis represents the consonants target and the *y* axis is the percentage of variant production.

Figure 5.20 shows results of AA and FR monolingual read speech, we reproduced the same experiment that we conducted on CS speech stop consonants. Read speech as explained in 3.3.2 is a three paces reading exercise (Fast, normal and slow reading). These results show that an important part of consonant target results, however some of those consonants occurrences were not realized in the word context as in AA: /d, g/ in AA initial word position, /p, g/ in intern position, /p, g/ and in final position. In FR /b, g/ in intern position, /b, g/ and in final position.

Despite the missing occurrences in some word positions, the results give us an idea about the stop consonants voicing variants. In fact, AA shows globally a higher variation rate in the three word positions. The final position variants are the highest of the three with French /p/ which is the consonant that tends the most to voicing and consonant /d/ which is the consonant that tends the most to devoicing in final position. AA also shows a high devoicing rate for stops /b, d/ in final position, and /t/ also shows a voicing rate of ∼25%. However, /k/ has a very low score. When comparing CS results to read speech results, voicing and devoicing variation in CS FR is globally lower to read speech.

Thus, we can deduce that stop voicing production is globally more stable in CS FR than in read FR, especially, final voicing for consonants /p, t/ and the final devoicing of /d/. Concerning AA, consonants voicing variant of voiceless stops in CS speech and read speech remains relatively similar concerning final position. The impact of CS on FR would be the preservation of canonical pronunciation with less voicing variants, and concerning AA, CS have a high variation rate in both speech types, but seems to have an impact only on few consonants such as final /k/ which has a higher score in CS speech compared to read speech, intern /b/.

5.3.3.4 Results: fricative consonants voicing variation

Figure 5.21 resumes the results of the shared AA and FR fricative voicing and devoicing in their three positions (initial, intern and final), as stop consonant variation we calculated the variation rate on a basis of 100%, and the stacked bars represent both parts of canonical pronunciation and the voicing and devoicing variants. As illustrates the figure, the voicing and devoicing consonant variation in both languages appears most frequent in final position. With the respect of consonant voicing of /f, s, ʃ/ AA recorded the highest variation rate voicing in first and intern position is more frequent with /ʃ/ with a rate of 28%. /s/ has the lowest variation rate in initial and intern position 13%. However, its devoicing in the final

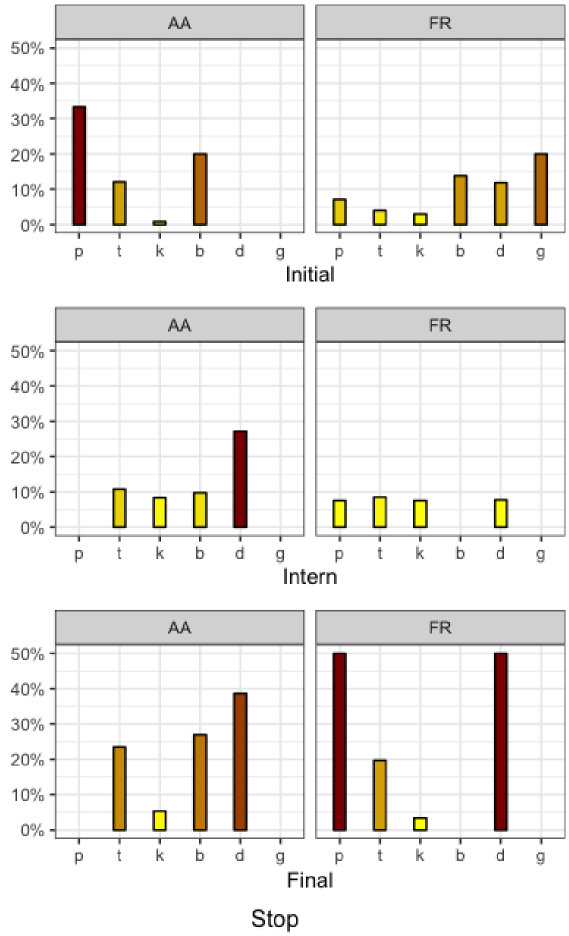


Figure 5.20: Results of stop voicing variants of FR and AA in word positions (Initial,intern, final) in read. The *x* axis represents the consonants target and the *y* axis is the percentage of variant production.

position reaches 37%. /f/ also has an growing increase in voicing variation according to the position, initial as a starting point with 13% and then 28% in intern position to finally reach its higher score in final position with 31%. Concerning FR consonants voicing, variation is very low, between 5% and 18% for all of the consonants /f, s, ʃ/ in their three positions. As AA, French final voicing is more frequent and consonant /f/ has the highest voicing rate which reaches 18%. /s/, and /f/ are following respectively in second and third position.

Concerning the devoicing variants of /f, z, ʒ/, AA consonants have higher scores in intern position rather than in initial position with a rate of 25% for both /z/ and /Z/. /v/ remains the consonant with the least devoicing rate with only 14% and 15% for initial and intern position. Concerning the final position, the figure shows no devoicing variation for /z/ and /ʒ/ and /v/ has no word occurrences in FACST corpus. FR devoicing consonants is more marked in final positions as mentioned here above, the consonant with the highest devoicing variation is /ʒ/ with a rate of 25%, it is followed by the fricative /z/ with a rate of 24% and the FR consonant with the lowest variation rate is /v/ with only 19%.

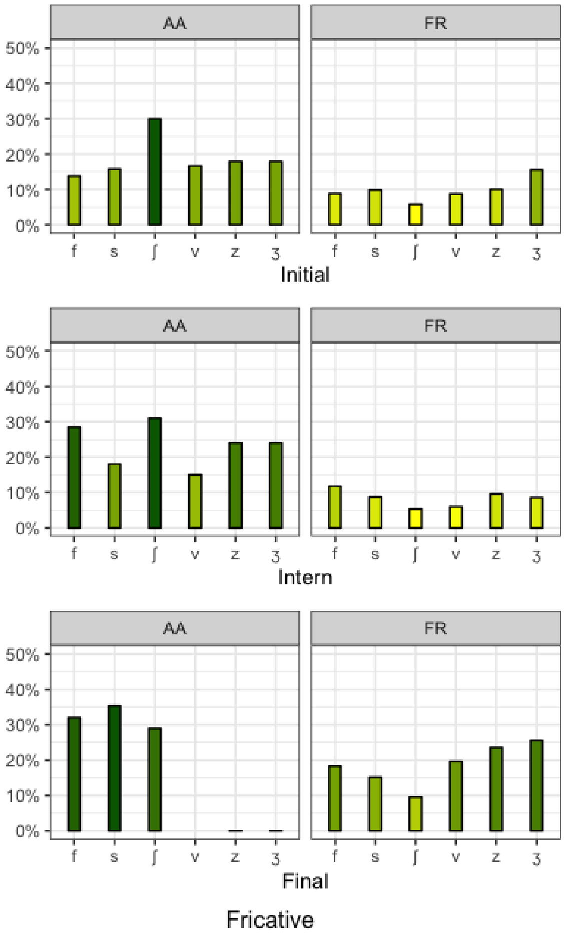


Figure 5.21: Results of fricative voicing variants of FR and AA in word positions (Initial,intern, final) in CS speech. The *x* axis represents the consonants target and the *y* axis is the percentage of the variant production.

Concerning fricative variation in read speech, the results in Figure 5.22 show that variation is globally lower in FR than in AA, concerning AA read speech, fricative /ch/ has a higher voicing rate in all word positions. Consonant /ʃ/ has an important final voicing with 34%. When comparing AA CS and read speech, the results show that the consonant /ʃ/ varies in both speech types, so CS has no impact on this variation. However, the consonants

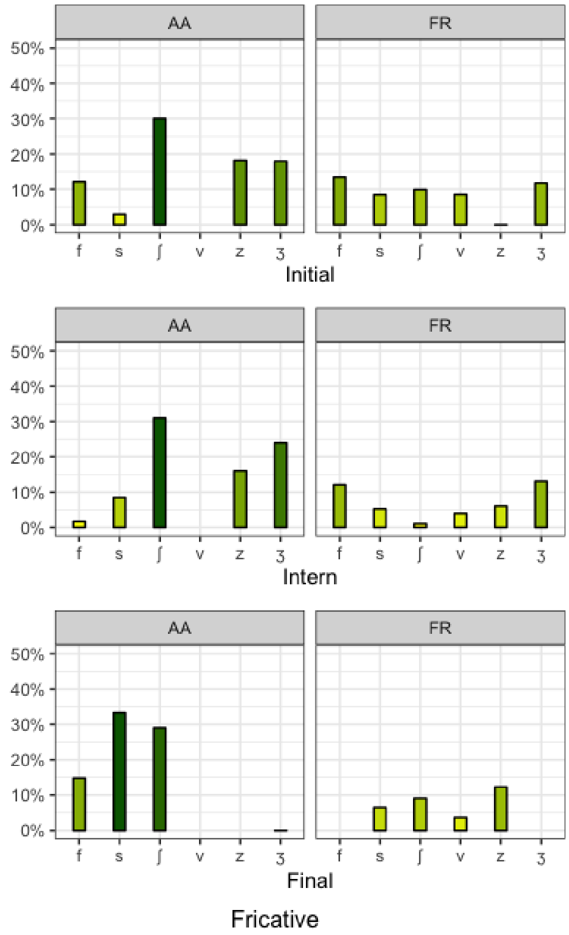


Figure 5.22: Results of fricative voicing variants of FR and AA in word positions (Initial,intern, final) in read speech. The *x* axis represents the consonants target and the *y* axis is the percentage of the variant production.

/s/ and /f/ have a higher CS variation than in read speech in their three positions with an average of 23% for CS and 12% in read speech.

Fricative voicing and devoicing in CS speech is globally higher in AA than in FR. However, the experiment shows that AA has higher voicing rate with /f, s, ʃ/ than devoicing rate variation with /v, z, ʒ/ in contrast to FR where devoicing rate is higher than voicing rate. Consonant voicing variation rate is different from a language to another, indeed, AA consonant with the highest voicing rate is /s/ and in FR it is /f/ which has the highest variation rate. Concerning /v, z, g/ devoicing, variation scale is similar in both languages. In fact, in initial and intern positions /ʒ/ has the highest devoicing rate and /v/ has the lowest devoicing rate in both languages in both initial and intern position. Devoicing in final position is clearly higher in FR than in AA. To conclude, we can say that AA in CS speech is less stable in voicing and devoicing consonants than FR. Consonants that vary the most to devoicing are similar in both languages and consonants that vary the most to voicing are different from a language to another. The final word voicing and devoicing variation is the most frequent in both FR and AA. For example, FR consonant variation (phone assimilation "je suis" [ʒəɥi] [ʒəɥi]. To sum up, we can conclude that, referring to the alignment results, assimilation phenomenon in FR and AA are different and do not include the same consonants.

5.3.3.5 Acoustic results

We addressed this acoustic analysis of target consonants in order to observe the voicing consonants through the v-ratio percentage. The results are introduced as follows, Figure 5.23 shows target stop consonants in AA and FR in CS speech, and Figure 5.24 shows fricatives v-ratio in FR and AA CS speech.

We notice that stop consonants, either in voiceless or voiced stops, present a v-ratio with balanced values in all three positions in both languages. Thus, we can conclude that there is a consonant voicing variation due to the absence of top and bottom values. AA /p, t, k/ have a voicing rate that changes with word position, the initial position has an average voicing rate of 14%, while the intern average position rate is of 12% and the final position rate is of 8%. /t/ has the highest v-ratio with a rate of 25%, and /p/ is a particular case with 25% in the beginning of the words and no variation rate in intern and final positions. FR voiceless stops have the highest values and increasing v-ratio according to word position with a /p/ that reaches 22% of voicing in word intern and 21% in word final. Concerning voiced stops, results show that there is a very high variation in all word positions, in both AA and FR, however AA has globally the lowest v-ratio values, with an average rate of 27%, when FR has a rate of 31%. The most devoiced Consonants are /b and d/ in final position with 32.5% and 30%, meanwhile /g/ in final position remains the consonant with the least devoicing variation reaching a voicing rate of 50%. So in FR, v-ratio values decrease in final position, compared to intern and initial positions. The most devoiced consonant is /d/ in final position with a rate of 29%, /g/ is in second place with 32% and /d/ comes last with 34%.

Fricatives also shows variation in their voicing rate in CS, especially in AA and specifically for the voiceless consonants /f, s, ʃ/. They show a very high variation rate in final word position, with /ʃ/ reaching 26% and /f/ reaching 22%. /s/ is the consonants that sustains the most, its devoicing rate with only 6%, 12% and 9% respectively for the three positions (initial, intern and final). FR voiceless fricatives also have very similar v-ratio shared by the three consonants with an average of 11% for initial and final word position and higher voicing rate in intern position with 19% for /ʃ/. Voiced stops has a high voicing rate in production with a v-ratio in AA final position that reaches 23% for consonant /z/ being the lowest score and 28% for /ch/. Meanwhile the FR part which has the lowest devoicing rate is the initial part which has the lowest v-ration for all of the three consonants reaching 40% of voicing with /f, z, ʒ/.

As explained here above full voiced consonant v-ratio equals to 100% and full voiceless consonant v-ratio equals to 0%.

5.3.3.6 Discussion

In this study we presented the consonant voicing and devoicing variation by using automatic alignment with variants. A parallel overview of the v-ratio the consonant targets has been presented.

We notice in CS speech obstruents voicing that stop consonants vary more than fricatives, the word position in both languages is relatively similar. In fact, most of the variation is localized in word final position. However, this experiment allows us to say that the difference between FR and AA variation is in the variation rate of each consonant which is different each time. Some consonants vary more in AA or in FR, this means that obstruents voicing and devoicing is different from a language to another. Thus, CS may have an impact on voicing variation. In fact, after a comparison with the read speech of both languages, we notice a difference in variation rates.

Acoustic analysis of v-ratio show that there is a large voicing variation in fricatives and stops. Indeed, v-ratio averages of each analyzed consonant do not correspond to top or bottom values. V-ratio rates are globally linked with the a considerable part of the variant alignment results, i.e. more variation in stops compared to fricatives. However, the analysis need to be refined at the parallel variants level of the target consonant towards the v-ratio

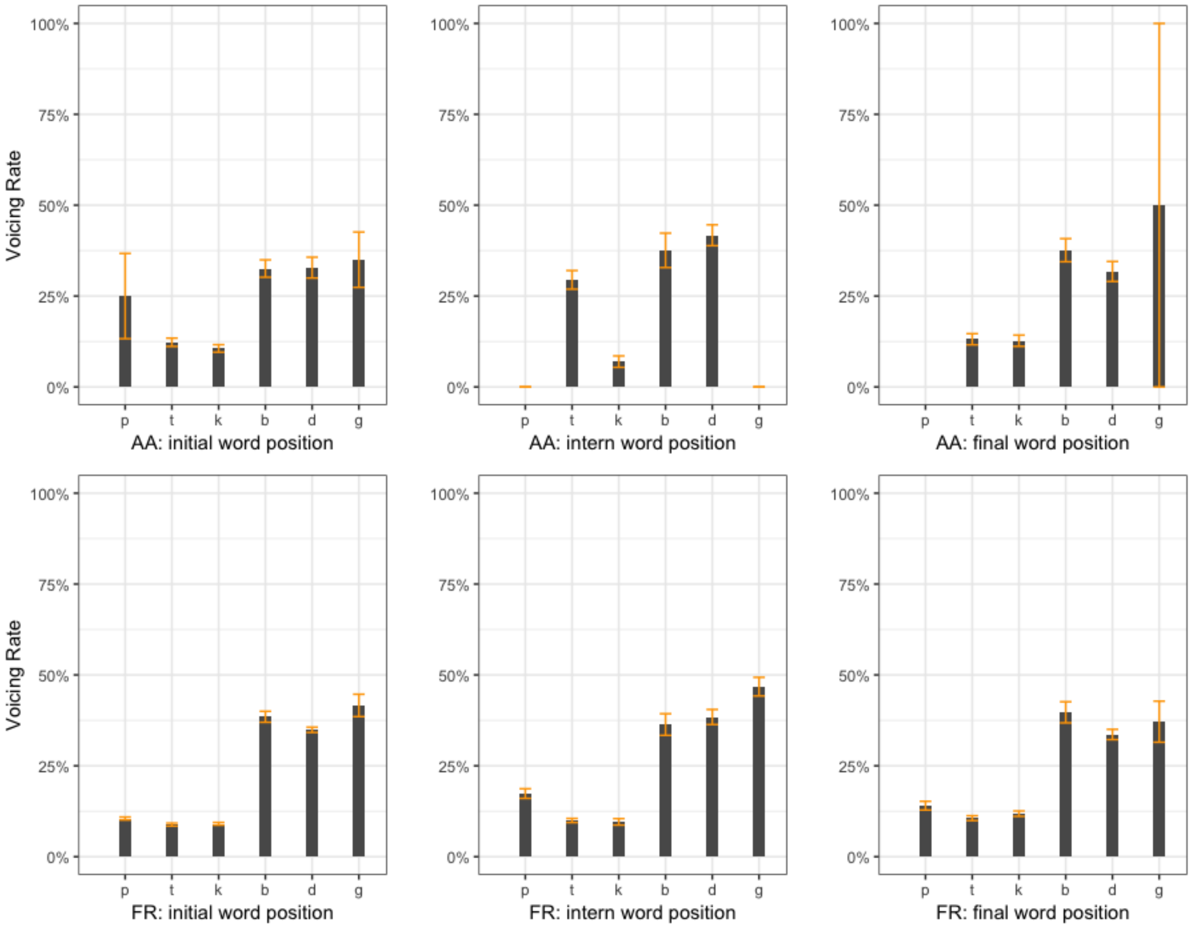


Figure 5.23: *v-ratio* in CS FR and AA stops /p, t, k, b, d, g/ in word positions: initial, medium, final (average percentage and standard deviation)

study of the target consonants realized in this work. This complementary analysis consists to calculate separately the v-ratio for each variant of the consonant target obtained from the alignment results. The aim of this analysis is to observe distinctly the acoustic clues of the obstruents variants (voicing and devoicing) and compare the acoustic results with the alignment variants.

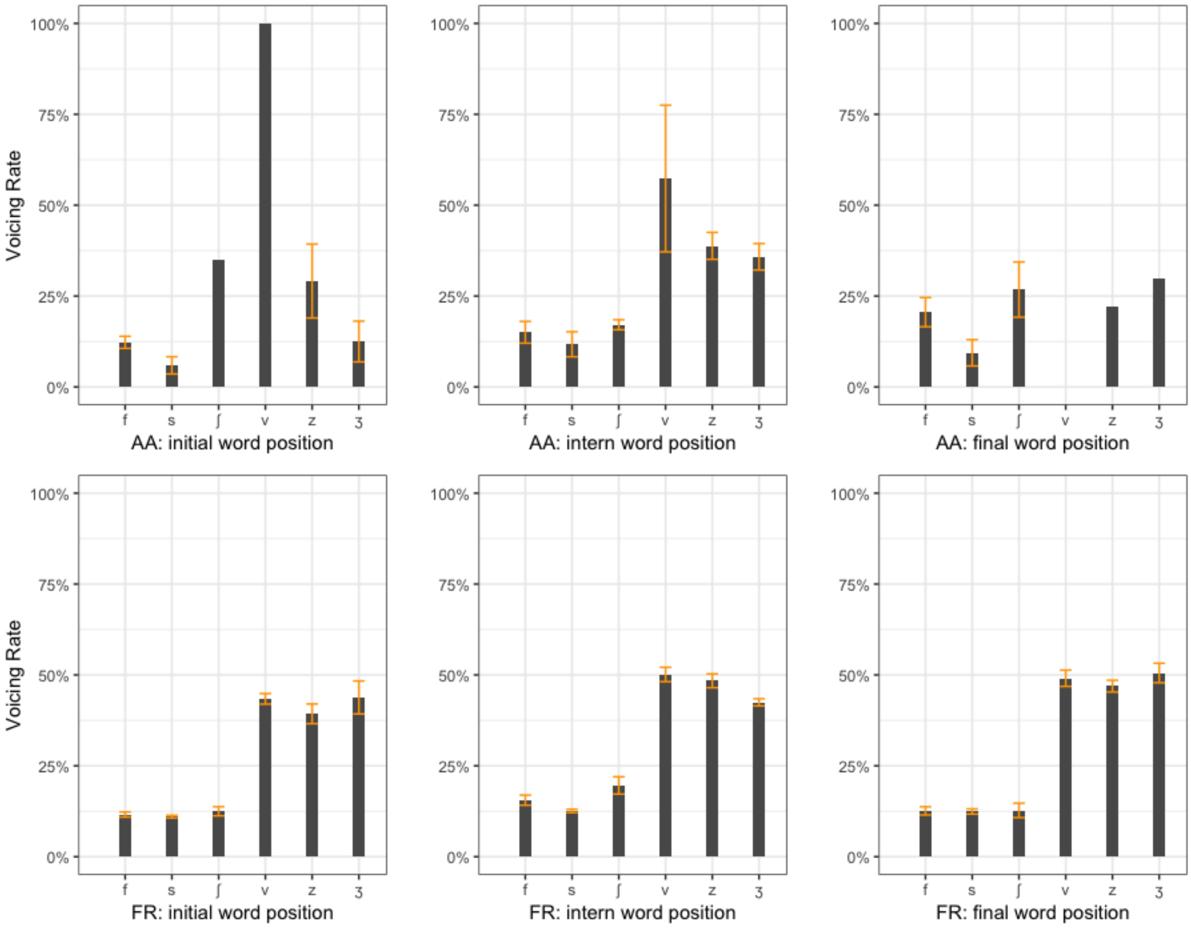


Figure 5.24: *v-ratio* of CS FR and AA fricatives /f,v,ʃ,v,z,ʒ/ in word positions: initial, medium, final (average percentage and standard deviation).

5.4 General discussion

In this chapter we presented a study of the variation in vowels and consonants in FR-AA CS speech by comparing this bilingual speech with monolingual speech and read speech. The comparison between the speech types clarifies the impact of CS on variation in vowels and consonants. The method used to study this variation is based on the automatic alignment of the speech with specific variants that allows a forced choice paradigms following the studied variations. The forced alignment with specific variants method may be considered to be globally more objective than human annotation and the forced choice guides the phonetic hypothesis to study the variation.

The vowel and consonant variation experiments show that globally CS, as a bilingual speech type, may have an impact on variation in production. In fact, speech can be linked to factors that allow more variation, like causal and journalistic speech (Adda-Decker and Lamel, 1999, 2017), and other factors, like the use of two phonological systems in one speech. So, the CS speech may help to maintain pronunciation and limit variation as observed in CS speech.

The vowel study investigated variation in the nearest neighbour vowels of a vocalic triangle representation. In the FR-AA CS study, the shared vowels /i, a, u/ are studied with the use of French vocalic system to establish the neighbouring vowel variants. The study of FR vowel variation in CS and monolingual speech, focalized on the variation of the peripheral vowels /i, e, a, o, u/. In the vowel centralization study, we investigated all FR and AA vowels.

The vowel variant studies show three major points. First, in the CS speech, AA speech is more affected by the variation than in the French speech. The acoustic analysis confirms the height variation of AA in the vocalic space. This variation could be related to the vocalic system of AA which contains only 6 vowels compared to the FR system with 15 vowels which may limit the variation. Also, by comparing the vowel variation in FR between CS and monolingual speech, the study concluded that the CS has an impact on the variation. The bilingual speakers in CS preserve more the canonical pronunciation than the monolingual speakers. The vowel centralization study in CS for FR and AA that included all vowels in each of the languages supports the above conclusion for the CS speech vowel variation. Indeed, the centralization of AA vowels is higher then the FR ones.

The three consonant variation studies addressed the main question of the consonantal influence of L*a* on L*b* in CS speech, specifically: the consonant gemination and simplification of the geminates, the emphtaization of the consonants and the consonant voicing/devoicing variation.

The gemination study focused on 4 shared consonants /d, n, l, s/. The study concluded overall that the variation in CS speech is higher in AA than in FR despite the presence of simple/geminate consonant contrast in the AA consonant system. The results also show that the simplification rate of AA geminates in CS speech is higher than the corresponding gemination rate. The gemination study also concludes that CS speech may influence the gemination variation in production. Indeed, the FR CS speech shows a lower consonant gemination rate compared to that of the monolingual speech.

Emphatization of /t, d, s/ consonants in FR and AA in CS speech revealed that AA variant rates are higher then FR ones in most of consonants. Besides, the simplification of AA emphatics in CS speech, the study shows that AA emphatics vary more then their plain consonant counterparts. It was also observed that the simplification of emphatics in AA is more frequent then the emphatization of the plain consonants.

The obstruent voicing and devoicing study shows that CS speech consonant voicing in stop consonants vary more than fricatives, the word position in both languages is relatively similar. In fact, most of the variation is localized in word final position. However, this experiment allows us to say that the difference between FR and AA variation is in the variation rate of each consonant which is different each time. Some consonants vary more in AA or in FR, this means that obstruents voicing and devoicing is different from a language to another. Thus, CS may have an impact on voicing variation. In fact, after a comparison with the read speech of both languages, we notice a difference in variation rates.

In summary, the alignment experiment with ABX choice method allows us to get an overview about variation in large scale speech data. The method also answered a number of phonetic questions about the speech production such as the difference between the CS speech and monolingual speech and the difference/similarity in phonetic production between the languages in CS speech. However, the automatic alignment needs to be validated by acoustic analyses and manual verification at the signal level in order to know the compatibility between the Hidden Markov Model (HMM) outputs, the acoustic signal data and the human perception.

Chapter 6

Conclusion and perspectives

Throughout this thesis, we addressed the code-switching phenomenon that is arousing growing interest from both linguists and speech technologists. We were particularly interested in natural, spontaneous French-Algerian Arabic code-switched speech which is highly practiced in French-Algerian bilingual community. To realize our studies, we constructed the FACST corpus, containing 8 hours of speech. The recordings are predominantly interactive dialogs with large amounts of code-switching, complemented with a minor part of read speech for control purposes.

This first focus of this thesis work was on the design of the corpus and a reflection of the best methodology to obtain code-switching speech data that could serve for linguistic, phonetic and speech technology studies. We proposed a method that consisted of first selecting code-switching speakers, and then contextualizing the conversations in order to naturally and spontaneously trigger code-switching. It was also necessary to take into account technical aspects of making the recordings in a soundproof room.

The thesis is also needed to carry out research on the topic of how to annotate and automatically process French-Arabic code-switching speech. For this, we proposed speech stretch segmentation methods that suit code-switching and that are based on language boundaries followed by segmentation based on oral phrases and speaker speech turns. We also proposed a transcription convention specifically designed for the French-Algerian Arabic language pair and that simplifies the processing of two languages that are generally written with different scripts in order to process both languages data using the same methods. This work was also an opportunity to contribute Algerian Arabic linguistic resources which are quite low compared with those for the French language.

The linguistic characteristics of the French-Algerian Arabic code-switching speech data were explored. An overview of the FACST corpus is given in terms of the words and word frequencies (types and tokens). For code-switching, sentence length in word count and duration, as well is the percentage of CS utterances. These figures show us that the code-switching if often occurs in intra-sentential position, resulting in very short speech stretches. The short switching between FR-AA CS is problematic for the automatic language identification, highlighting one of the difficult challenges of this language pair for LID and ASR in general.

We proposed to develop a forced alignment of code-switching speech by combining two monolingual alignments; FR for French data and AA for Arabic data. We also took the opportunity to improve the Arabic pronunciation dictionaries with the phonetic observations noticed during the corpus processing with spectral representations.

The proposed studies of phonetic variation of vowels and consonants reveal that the languages vary in different ways and manners in CS. The CS speech also differs from monolingual speech. Vowel production is more stable in FR in code-switching than in FR monolingual speech. Our corpus based analysis suggests that our code-switching speakers are able to vary their vowel productions according to the language they speak, thereby adjusting their productions to the respective vowel systems.

Concerning consonants, the gemination of AA plain consonants is more prevalent in AA than in FR speech. The opposition between plain consonants and geminates in a phonological system does not exclude phonetic variation tending to gemination and simplification of geminates in production especially in code-switching speech. The corpus study revealed that the consonants the most affected by gemination in FR-AA code-switching speech are /d, s, l/ whereas /n/ remained more stable both in FR and AA. Gemination variation of FR in code-switching speech is frequent in CS FR than FR monolingual speech, i.e. code-switching has a real impact on consonant variation.

With the respect of emphatisation study, code-switching also has an impact on /t, d, s/

consonant emphatization in both languages. The impact of cod-switching on French is very clear in that monolingual speech shows more variation than is found in CS speech (12% variation in FR CS and 25% in monolingual French).

In obstruent voicing variation in code-switching, AA has high variant rates comparing to the FR in stop voicing variation. The stop consonants vary more than fricatives referring to the alignment experiment. Some consonants vary more in AA or in FR, this means that obstruents voicing and devoicing is different from a language to another. Thus, CS may have an impact on voicing variation.

Perspectives and future directions

In this thesis we built and organized a CS corpus that allows to carry on our investigation on automatic speech recognition, phonetic variation studies and gives more linguistic and sociophonetic descriptions on this pair of language.

French and Algerian-Arabic is a language pair where speakers produce frequent code-switching a large community in France and Algeria uses this CS on a daily basis in communication. Developing an automatic speech recognition for FR-AA CS is worth considering in spite of the challenges that we may face (very short language switches, AA as a low resourced language, locating sufficient data to train language models for ASR). So, one of the near term future research directions would be to conduct ASR experiments on CS speech from these two languages. During this thesis we also noticed that there are regular expressions in the FR-AA corpus and words that may introduce a language switch. Future studies with the FACST corpus can include measurements of the observed language switch in order to predict CS using automatic speech recognition tools.

There have been few phonetic studies of code-switched speech and the experiments that we conducted show that there is a vowel and consonant variation that is produced in this type of speech in FR and in AA. Thus, another future research direction is to study the variation in AA CS by comparing CS speech with monolingual, spontaneous Algerian Arabic speech data.

In further investigations of the acoustic characteristics of geminate consonants, is also worth considering amplitude measures in addition to the duration measurements already explored. Other acoustic studies on FR-AA CS specificities, especially of the emphatic consonants, the format measures (F1, F2 and F3) according to the vowel type (frontal, back, opened and closed) are also interesting to consider.

The prosody of code-switched speech is another growing research field. Comparative studies of prosodic variation in code-switched speech and monolingual speech of both languages is also an interesting path to explore in future studies. Such studies may lead to the identification of relevant acoustic clues for language switch and automatic transcription of code-switched speech.

Part III

Appendix

Appendix A

FACST written outlines

A.1 Read texts

A.1.1 AA: excerpt from Algerian movie scenario "Bab El-Oued City"

(Allouache, 1994)

A.1.1.1 Arabic orthography version

بَاب الوَاد سِيتِي

اَلبَارَح زَادُو قَتلُوا ثَلَاث پُولِيسِيَة فِي الِّيل.
سَمعنَا الرَّصَاص يَضرَب مُدَّة سوَايَع طَوِيلَة.
علَاش رَانِي نَكتَبلَك؟
هَذِي ثلَاث سِنِين مَا علَى بَالِيش وِين رَاك، فِي أأمَا بلَاد؟
كُنت تقُولِي بَلِي توَلِّي وَ تَدِّينِي معَك
خَلِيتنِي نَستَنَا وَ علَاش يَا بُوعلَام وَ علَاش نسِيتنِي؟
اَلَّه يَلعَن هَذَاك النّهَار اَلسوَد هَذِيك الجُمعَة ...
اَلصِيف وَ السخَانَة كَانَت تَقتَل
كُنَّا غِيرِكِيمَا بدِينَا نَنسَاوَا اليَامَات الحَزِينَة وَ عُنف الثمَانِيَة وَ ثمَانِين.
بَاب الوَاد حُومَتنَا عَرفَت المُوت وَ الجَرحَى وَ السّجن

وَ بدَات تَرجَع لِلهُدُؤ وَ الرَّاحَة وَاش خذَاك يَا بُوعلَام هَذَاك النّهَار وَاش خذَاك.

A.1.1.2 Transliterated version

baAb al waAd siytiy

al baAraH zaAduw qatluwA FalaAF puwliysiyyaM fiy al liyl.
smacnaA al rraSaAS yaDrab muddaM swaAyac TawiylaM
claAX raAniy naktab lak ?
haViy FlaAF sniyn maA claY baAliyX wiyn raAk, fiy EmaA blaAd
kunt tquwl liy bi alliy twalliy wa taddiyniy mcaAk
xalliytniy nastannaA wa claAX yaA buwclaAm wa claAX nsiytniy
allah yalcan haVaAk al nnhaAr al aswwad haVdiyk al jumcaM...
al SSiyf wa al ssxaAnaM kaAnat taqttal
kunnaA Giyr kiymaA bdiynaA nansaAw al yaAmaAt al HaziynaM wa cunf al FFmaAniyaM wa al FFmaAniyn
baAb al waAd HuwmatnaA carfat al muwt wa al jarHaY wa al ssijn
wa bdaAt tarjac li al hudua wa al rraAHaM.
waAX xVaAk yaA buwclaAm haVaAk al nnhaAr waAX xVaAk

A.1.2 FR: excerpt from "Le Petit Price"

(De St Exupéry, 1943)

C'est alors qu'apparut le renard.
Bonjour, dit le renard!
bonjour, répondit poliment le petit prince, qui se retourna mais ne ne vit rien.
Je suis là, dit la voix, sous le pommier !
Qui es-tu? dit le petit prince. Tu es bien joli.
je suis un renard, dit le renard
Viens jouer avec moi, lui proposa le petit prince . Je suis tellement triste.

Je ne puis pas jouer avec toi, dit le renard. Je ne suis pas apprivoisé

Ah! pardon, fit le petit prince. mais, après réflexion, il ajouta: qu'est-ce que signifie apprivoiser ?

Tu n'es pas d'ici, dit le renard, que cherches-tu ?

Je cherche les hommes, dit le petit prince. qu'est ce que signifie apprivoiser ?

Les hommes, dit le renard, ils ont des fusils et ils chassent. C'est bien gênant! ils élèvent aussi des poules. c'est leur seul intérêt. Tu cherches des poules?

Non , dit le petit prince. je cherche des amis. qu'est-ce que signifie apprivoiser ?

A.2 Questions guide to elicit code-switching conversations

Conversation topics	Main questions	Side questions	Lang to solicit the most	Time devoted
About the reading session	What do you think about the read texts?	Did you read the texts before? Do you know the end of the stories? What was the end?	AA	3mn
Studies	Can you explain your university education? How do you describe student life in both countries? How did you choose these formation?	What are your studies? Where did you study? What are the differences?	AA	5mn
Work	What work(s) did you do? Would you like to continue in this work, why?	In which country? What would you have done in Algeria?	AA	4mn
Everyday Life in the countries	Could you compare life in both countries?	Which country do you prefer?	FR	4mn
Bilingualism, CS & languages	What do you think about CS practices? How can we to teach an oral language like AA?		FR	5mn
Hobbies	Do you play music? Do you like sport?	What do you think about traditional Algerian music? What can you say about popular sports in France and in Algeria?	FR	4mn

Table A.1: Question guide table of the recorded conversation and eliciting code-switching speech. The side questions are modified and adapted for each speaker answers of the the main questions. The languages to solicit the most can be changed depending to the speaker's language preferences. The duration devoted for each topic can be modified if the speaker speaks more or less about the topic. The total duration of the code-switching record is 25mn

Appendix B

FACST transcription convention

B.1 French transcription and annotations

<SP1>	Name or number of speaker
FR:	French language speech stretch
New line	New segment
,	Short pause in the speech stretch
?	Interrogative intonation
!	Exclamatory, imperative and other intonations
"orches-"	Uncompleted word, "orchestre" *orchestra*
$[euh]$	Speech disfluency in the stretch
$[hum]$	Speech disfluency in the stretch
$[ah]$	Speech disfluency in the stretch
$[hesitation]$	Hesitation at the begin, end or in the middle of the speech segment
(infor)mation	The part between brackets is not pronounced
allé/aller	Hesitation in spelling
sort/dort	Hesitation in perception to transcribe
$[pas - clair]$	Speech not clear or not audible
$[rire]$	Laugh in the speech
$[clique]$	Click with mouth

Table B.1: Convention of French transcription and annotation

NB: the speech stretches are not cut, filtered, and cleaned. They are fully transcribed, even the stretches that are intelligible.
Overlaps are managed by Transcriber. This program allows to generate two lines of transcription with a same time code.

The noise and music are absent in the records. The records were realized in soundproof room.

B.2 Algerian Arabic transcription and annotation

<SP1>	Name or number of speaker
AA:	French language speech stretch
New line	New segment
,	Short pause in the speech stretch
?	Interrogative intonation
!	Exclamatory, imperative and other intonations
"yakt-"	Uncompleted word, "yaktab" *he writes*
$[euh]$	Speech disfluency in the stretch
$[hum]$	Speech disfluency in the stretch
$[ah]$	Speech disfluency in the stretch
$[hesitation]$	Hesitation at the begin, end or in the middle of the speech segment
(muHt)aAj	The part between brackets is not pronounced
daAr/faAr	Hesitation in perception to transcribe
$[pas - clair]$	Speech not clear or not audible
$[rire]$	Laugh in the speech
$[clique]$	Click with mouth

Table B.2: Convention of Algerian Arabic transcription and annotation

NB: the speech stretches are not cut, filtered, and cleaned. They are fully transcribed, even the stretches that are intelligible.

Overlaps are managed by Transcriber. This program allows to generate two lines of transcription with a same time code.

The noise and music are absent in the records. The records were realized in soundproof room.

B.3 Algerian Arabic transcription symbols

IPA	Transliteration	Arabic letter	IPA	Transliteration	Arabic letter
p	p	پ	ʃ	X	ش
b	b	ب	x	x	خ
t	t	ت	ɣ,ʁ	G	غ
t	M	ة	ħ	H	ح
tˤ	T	ط	ʕ	c	ع
d	d	د	h	h	ه
dˤ/ðˤ	D	ض	l	l	ل
k	k	ك	r	r	ر
g	g	ڨ	w	w	و
q/g	q	ق	j	y	ي
ʔ	E	أَ	i	i	اِ
dʒ/ʒ	j	ج	iː	iy	اِي
m	m	م	u	u	اُ
n	n	ن	uː	uw	اُو
f	f	ف	a	a	اَ
v	v	■	aː	aA	اَا
θ	F	ث		Y	ى
ð	V	ذ			
s	s	س			
sˤ	S	ص			
z	z	ز			

Table B.3: Algerian Arabic transliteration symbols

B.4 Symbol sets for transcription

IPA	Ortho	Limsi symbol
aː	aA	A
a	a	a
ɑ	A	A
ʔ	E	E
i	i	i
iː	iy	I
u	u	u
uː	uw	U
w	w	w
p	p	p
b	b	b
t	t	t
t^{ʕ}	T	+
d	d	d
d^{ʕ}	D	ø
k	k	k
g	g	g
q	q	q
ʒ/ dʒ	j	ý
m	m	m
n	n	n
f	f	f
v	v	v
θ	F	þ
ð	V	ð
s	s	s
s^{ʕ}	S	ß
z	z	z
ʃ	X	c
x	x	x
ɣ,ʁ	G	ç
ħ	H	å
h	h	h
ʕ	c	æ
l	l	l
r	r	r

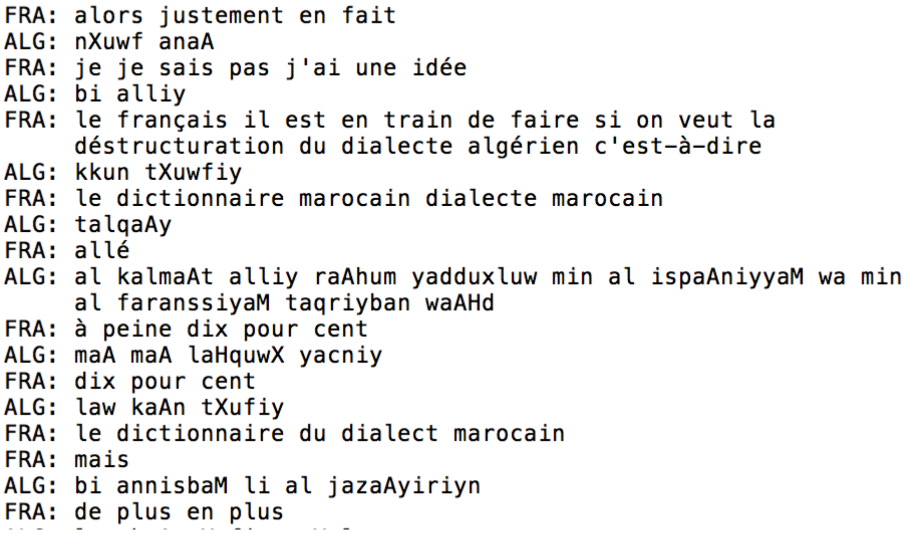
IPA	Ortho	Limsi symbol
bb	bb	B
dd	dd	D
dd ^{ʕ}	DD	Ø
gg	gg	G
pp	pp	P
tt	tt	T
tt^{ʕ}	TT	÷
kk	kk	K
qq	qq	Q
ss	ss	S
ʃʃ	XX	C
ss ^{ʕ}	SS	§
zz	zz	Z
ʒ ʒ	jj	Ý
ff	ff	F
vv	vv	V
θθ	FF	Þ
ðð	VV	¥
mm	mm	M
nn	nn	N
ll	ll	L
rr	rr	R
ɣɣ,ʁʁ	GG	Ç
ww	ww	W
jj	yy	Y
hh	hh	H
ɦ	HH	Å
xx	xx	X
ʕ	cc	Æ
ʔʔ	EE	?
	Hesitation (euh, hum, ...)	&
	Respiration	H
	Silence	.

Table B.4: Arabic set phones and symbols for for phonetic and orthographic transcription. The first part of the table shows the Arabic phones and the second part shows the phonetic and orthographic transcriptions of the geminate consonants

IPA phone	Ortho transcription	Limsi phone alignment
i	i	i
e	e	e
ɛ	e/ai/è	E
y	u	y
œ	eu	X
ə	e	x
ø	eu	@
a	a	a
ɔ	o	c
o	o/au	o
u	ou	u
ɛ̃	un/in	I
ã	an/ant	A
ɔ̃	on	O
ɥ	ui ua	h
w	w/oi	w
j	y/ill/	j
s	s	s
z	z	z
ʃ	ch	S
ʒ	g/j	Z
f	f	f
v	v	v
n	n	n
m	m	m
ɲ	gn	N
l	l	l
r	r	r
p	p	p
b	b	b
t	t	t
d	d	d
k	c/k	k
g	g	g
	Hesitation (euh, hum, ...)	&
	Respiration	H
	Silence	.

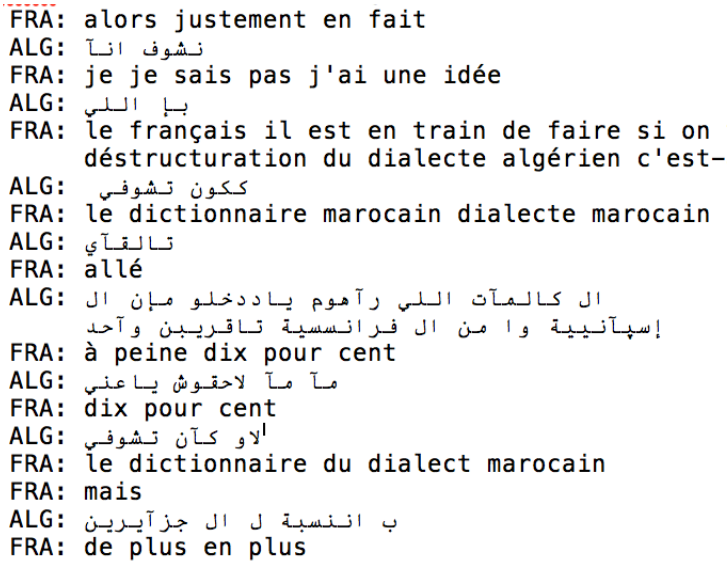
Table B.5: French set phones in orthographic transcription and phone alignment

B.5 Example of Code-switching transcriptions



```
FRA: alors justement en fait
ALG: nXuwf anaA
FRA: je je sais pas j'ai une idée
ALG: bi alliy
FRA: le français il est en train de faire si on veut la
     déstructuration du dialecte algérien c'est-à-dire
ALG: kkun tXuwfiy
FRA: le dictionnaire marocain dialecte marocain
ALG: talqaAy
FRA: allé
ALG: al kalmaAt alliy raAhum yadduxluw min al ispaAniyyaM wa min
     al faranssiyaM taqriyban waAHd
FRA: à peine dix pour cent
ALG: maA maA laHquwX yacniy
FRA: dix pour cent
ALG: law kaAn tXufiy
FRA: le dictionnaire du dialect marocain
FRA: mais
ALG: bi annisbaM li al jazaAyiriyn
FRA: de plus en plus
```

Figure B.1: Example of Code-switching transcription



```
FRA: alors justement en fait
ALG: نشوف انآ
FRA: je je sais pas j'ai une idée
ALG: بإ اللي
FRA: le français il est en train de faire si on veut la
     déstructuration du dialecte algérien c'est-à-dire
ALG:  ككون تشوفي
FRA: le dictionnaire marocain dialecte marocain
ALG: تالقآي
FRA: allé
ALG: ال كالمآت اللي رآهوم ياددخلو مإن ال
     إسپآنيية وا من ال فرانسسية تاقريبن وآحد
FRA: à peine dix pour cent
ALG: مآ مآ لاحقوش ياعني
FRA: dix pour cent
ALG: لاو كآن تشوفي
FRA: le dictionnaire du dialect marocain
FRA: mais
ALG: ب اننسبة ل ال جزآيرين
FRA: de plus en plus
```

Figure B.2: AA characters conversion of the example in the Figure B.1

Appendix C

Experience of Code-switching practice (ECSP)

C.1 Online questionnaire

Fiche personnelle et questionnaire

Ce formulaire est pour le classement et la catégorisation des données,

*Obligatoire

1. **Adresse e-mail** *

2. **Nom**

3. **Pénom** *

4. **Sexe** *

   Une seule réponse possible.

   ○ Homme
   ○ Femme

5. **Âge** *

6. **Pays et ville de naissance** *

7. **Pays et ville de résidence actuelle** *

8. **Depuis** *

   Exemple : 15 décembre 2012

9. **Niveau d’étude (exemple BAC+ 3)** *

10. **Domaine d’étude** *

11. **Profession** *

____

12. **Professions anciennes**

____

Langues et statuts des langues

13. **Quelles langues parlez-vous ?** *

____

____

____

____

____

14. **Quelle est votre langue maternelle ?** *
Plusieurs réponses possibles.

□ Arabe
□ Français
□ Autre : ______________________________

15. **Quelle est la langue de vos parents /entourage ?** *
Plusieurs réponses possibles.

□ Arabe
□ Français
□ Autre : ______________________________

16. **Comment avez-vous appris les autres langues que vous parlez ?** *

____

____

____

____

____

17. **À quel âge ?** *

____

18. **Quelles sont les langues que vous pratiquez quotidiennement ?** *

Plusieurs réponses possibles.

□ Arabe
□ Français
□ Autre : ______________________

19. **Quelle est la langue que vous pratiquer le plus ?** *

Une seule réponse possible.

○ Arabe
○ Frabçais
○ Autre : ______________________

20. **En quelle langue avez-vous fait votre éducation ?** *

Plusieurs réponses possibles.

□ Arabe
□ Français
□ Éductation bilingue
□ Autre : ______________________

21. **Y a-t-il une langue que vous pensez maîtriser le plus ?** *

Une seule réponse possible.

○ Arabe
○ Français
○ Je maîtrise les deux d'une manière égale

22. **Vous considérez-vous comme bilingue ?** *

Une seule réponse possible.

○ Oui
○ Non

23. **Préfériez-vous une langue à une autre ?** *

Une seule réponse possible.

○ Oui
○ Non

24. **Pourquoi ?** *

____

____

____

____

____

Pratiques du Code-switching

25. **Aimez-vous le mélange de langues dans vos échanges ?** *

Une seule réponse possible.

○ Oui

○ Non

○ Je ne sais pas

26. **Pourquoi mélanger-vous les langues dans vos conversations ?** *

____

____

____

____

____

27. **Avec qui vous mélangez quand vous parler ?** *

____

____

____

____

____

28. **Dans quel contexte vous mélangez-vous les langues généralement ?** *

Plusieurs réponses possibles.

□ Humour

□ Politique

□ Travail

□ Religion

□ Échanges non formels

□ Autre : ____

29. **Pourriez-vous donner quelques exemples de mélange de langues que vous utiliser fréquemment ?**

30. **Rencontreriez-vous quelquefois des difficultés lorsque vous mélangez les langues ? Exemples : blocage d'expression liée à la pensée en deux langues, difficulté de relier deux énoncés de deux langues différentes, difficultés liées au changement du système phonétique, lapsus au moment du changement des langues …** *

31. **Que pensez-vous le l'alternance de langues ? Est-ce une mode, un choix de communication ?** *

WECHRAK?
COMMENT VAS TU?
COMMENT VA LA MAISON?
J'SUIS CONTENT DE TE VOIR YA KHO…
GYPS

C.2 Participant's answers

Link to the participant's answers here. The answers are in French.

Appendix D

Phones occurrences frequency

D.1 Algerian Arabic phones

D.2 French phones

# Occ	**phone**	**# Occ**	**phone**
6471	a	153	ß
4048	A	141	D
2820	l	138	N
2542	I	133	R
2118	n	126	S
1851	t	101	’
1482	m	93	ø
1409	w	83	ç
1299	k	59	§
1261	æ	43	p
1061	r	38	W
1039	b	35	g
1003	h	31	T
950	d	28	Z
794	y	18	B
749	U	17	M
724	u	15	K
659	c	9	v
466	å	8	Å
456	s	8	÷
403	q	7	Y
389	i	7	H
353	ð	3	F
335	ý	2	Ý
334	L	2	Q
282	x	2	C
273	þ	1	Đ
233	+	1	X
211	z		

Table D.1: AA Phones occurrences frequency in CS speech in decreasing order. The phones are displayed with LIMSI convention’s symbols

#occ	phone
817	a
8671	r
7688	i
7366	l
7159	s
6338	e
6213	E
5966	x
5721	t
5412	p
4694	k
4659	d
3672	m
3348	A
2917	n
2847	Z
2381	y
2307	O
2210	v
2177	c
2073	I
2018	z
1882	j
1807	f
1641	w
1631	u
1265	b
1026	o
961	S
670	h
520	g
483	@
481	X
233	N

Table D.2: AA Phones occurrences frequency in CS speech in decreasing order. The phones are displayed with SAMPA convention's symbols

Appendix E

Segmental variation classification results

E.1 Vowel variation experiments

Vowels		AA		FR	
Target	Var	Var Rate %	#Occ	Var Rate %	#Occ
i	i	42.23		76.38	2659 / 4308
	e	39.72		13.96	1099 / 4308
	y	18.05		9.66	550 / 4308
a (anterior var)	a	44.67		75.84	5025 / 7063
	œ	31.52		17.55	1350 / 7063
	ɛ	23.81		6.61	688 / 7063
a (posterior var)	a	38.99		82.01	5093 / 7060
	œ	48.64		10.83	1306 / 7060
	ɔ	12.37		7.51	661 / 7060
u	u	32.34		73.65	667 / 1182
	o	37.26		16.35	364 / 1182
	ø	30.40		10.00	151 / 1182

Table E.1: AA and FR vowel variation in CS speech

Vowels		AA		FR	
Target	Var	Var Rate %	#Occ	Var Rate %	#Occ
i	i	52.77		84.27	1854 / 2668
	e	37.13		10.68	567 / 2668
	y	10.10		5.04	247 / 2668
a (ant var)	a	33.67		87.80	1602 / 1870
	ɛ	36.45		6.96	159 / 1870
	œ	29.88		5.24	109 / 1870
a (post var)	a	44.27		89.52	1691 / 1869
	œ	43.54		8.29	130 / 1869
	ɔ	12.19		2.19	48 / 1869
u	u	29.8		85.29	275 / 428
	o	55.30		9.66	104 / 428
	ø	14.90		5.04	49 / 428

Table E.2: AA and FR vowel variation in read speech

Vowels		FR monolingual		FR CS	
Target	Var	Var Rate %	#Occ	Var Rate %	#Occ
i	i	84.61		76.38	2659 / 4308
	ɛ	10.55		13.96	1099 / 4308
	y	4.84		9.66	550 / 4308
e	e	57.81		57.34	3180 / 5077
	ɛ	28.33		27.98	985 / 5077
	œ	13.86		14.68	912 / 5077
a (ant var)	a	69.63		75.84	5025 / 7063
	ɛ	15.43		17.55	1350 / 7063
	œ	14.94		6.61	688 / 7063
a (post var)	a	77.47		82.01	5600 / 7057
	ɔ	20.14		10.83	781 / 7057
	œ	2.40		7.51	676 / 7057
o	o	66.08		72.19	580 / 858
	ɔ	14.28		17.38	141 / 858
	ø	19.64		10.43	137 / 858
u	u	76.17		73.65	667 / 1182
	o	10.83		16.35	364 / 1182
	ø	13.00		10.00	151 / 1182

Table E.3: FR monolingual (NCCFr corpus) and FR CS (FACST) vowel variation of

E.2 Consonant variation experiments

Consonants		AA		FR	
Target	Variants	Variant Rate %	#Occ	Variant Rate %	#Occ
p	p	74.00	37/50	78.89	2852/3615
	P	26.00	13/50	21.11	763/3615
t	t	70.04	1031/1472	67.99	3247/4776
	T	29.96	441/1472	32.01	1529/4776
k	k	77.14	1083/1404	78.82	3036/3852
	K	22.86	321/1404	21.18	816/3852
b	b	76.50	739/966	71.07	710/999
	B	23.50	227/966	28.93	289/999
d	d	80.35	920/1145	78.22	2902/3710
	D	19.65	225/1145	21.78	808/3710
b	g	78.89	71/90	69.96	389/556
	G	21.11	19/90	30.04	167/556
f	f	74.32	356/479	61.83	904/1462
	F	25.68	123/479	38.17	558/1462
s	s	81.54	592/726	72.64	3812/5248
	S	18.46	134/726	27.36	1436/5248
ʃ	ʃ	88.59	761/859	86.87	344/396
	ʃʃ	11.41	98/859	13.13	52/396
v	v	88.46	23/26	75.66	1259/1664
	V	11.54	3/26	24.34	405/1664
z	z	72.65	356/490	76.72	1193/1555
	Z	27.35	134/490	23.28	362/1555
ʒ	ʒ	83.33	285/342	75.96	1523/2005
	ʒʒ	16.67	57/342	24.04	482/2005
m	m	74.02	960/1297	72.56	2115/2915
	M	25.98	337/1297	27.44	800/2915
n	n	81.31	1649/2028	80.96	2194/2710
	N	18.69	379/2028	19.04	516/2710
l	l	76.03	1970/2591	75.49	3998/5296
	L	23.97	621/2591	24.51	1298/5296
ʁ/ɣ	ʁ/ɣ	80.00	87/115	86.67	4885/5636
	ʁʁ/ɣɣ	20.00	23/115	13.33	751/5636
w	w	83.89	984/1173	88.09	991/1125
	W	16.11	189/1173	11.91	134/1125

Table E.4: FR and AA CS geminate variant rates

Consonants		AA		FR	
Target	Variants	Variant Rate %	#Occ	Variant Rate %	#Occ
p	p	96.83	61/63	88.60	1546/1745
	P	3.17	2/63	11.40	199/1745
t	t	64.46	671/1041	72.13	823/1141
	T	35.54	370/1041	27.87	318/1141
k	k	67.58	471/697	63.86	493/772
	K	32.42	226/697	36.14	279/772
b	b	77.91	529/679	93.17	232/249
	B	22.09	150/679	6.83	17/249
d	d	78.78	412/523	84.94	818/963
	D	21.22	111/523	15.06	145/963
g	g	–	–	75.00	12/16
	G	–	–	25.00	4/16
f	f	69.70	207/297	92.59	325/351
	F	30.30	90/297	7.41	26/351
s	s	66.60	674/1012	66.19	1100/1662
	S	33.40	338/1012	33.81	562/1662
ʃ	ʃ	88.17	313/355	63.21	189/299
	ʃʃ	11.83	42/355	36.79	110/299
v	v	–	–	82.56	497/602
	V	–	–	17.44	105/602
z	z	74.59	182/244	74.77	560/749
	Z	25.41	62/244	25.23	189/749
ʒ	ʒ	76.82	169/220	80.98	694/857
	ʒʒ	23.18	51/220	19.02	163/857
m	m	71.69	585/816	73.85	401/543
	M	28.31	231/816	26.15	142/543
n	n	82.23	972/1182	64.94	752/1158
	N	17.77	210/1182	35.06	406/1158
l	l	73.18	1667/2278	73.73	870/1180
	L	26.82	611/2278	26.27	310/1180
ʁ/ɣ	ʁ/ɣ	82.00	57/57	91.99	1940/2109
	ʁʁ/ɣɣ	18.OO		8.01	169/2109
w	w	91.63	963/1051	90.07	481/534
	W	8.37	88/1051	9.93	53/534

Table E.5: FR and AA read speech geminate variant rate

Consonants		NccFr	
Target	Variants	Variant Rate %	#Occurrences
p	p	90.21	42629/47255
	P	9.79	4626/47255
t	t	75.70	50084/66159
	T	24.30	16075/66159
k	k	87.98	34684/39423
	K	12.02	4739/39423
b	B	17.33	2779/16032
	b	82.66	13252/16032
d	D	21.33	2779/16032
	d	78.66	13252/16032
g	g	82.41	5398/6550
	G	17.59	1152/6550
f	f	71.22	15238/21396
	F	28.78	6158/21396
s	s	73.60	56043/75126
	S	26.40	19083/75126
s	c	84.05	4673/5560
	C	15.95	887/5560
v	v	74.90	20012/26719
	V	25.10	6707/26719
z	z	79.91	10965/13721
	Z	20.09	2756/13721
ʒ	ʒ	81.11	19300/23794
	ʒʒ	18.89	4494/23794
m	m	78.42	34813/44391
	M	21.58	9578/44391
n	n	83.42	32623/39105
	N	16.58	6482/39105
l	l	70.40	40687/50608
	L	29.60	9921/50608
ʁ	ʁ	89.92	62830/69873
	ʁʁ	10.08	7043/69873
w	w	93.48	31050/33216
	W	6.52	2166/33216

Table E.6: Monolingual FR speech geminate variant rate (NCCFr corpus)

Consonants		AA		FR	
Target	Var	Var Rate %	#Occ	Var Rate %	#Occ
t	t	73.64	1081/1468	83.97	8622/10268
	tˤ	26.36	387/1468	16.03	1646/10268
d	d	80.24	918/1144	88.56	6612/7466
	dˤ	19.76	226/1144	11.44	854/7466
s	s	77.96	283/363	86.00	9410/10942
	sˤ	22.04	80/363	14.00	1532/10942
tˤ	t ˤ	75.19	59/81	–	–
	t	24.81	10/81	–	–
dˤ	d	58.33	49/84	–	–
	dˤ	41.67	35/84	–	–
sˤ	sˤ	56.35	213/378	–	–
	s	43.65	165/378	–	–

Table E.7: AA and FR emphatization variants in CS speech

Consonants		AA		FR	
Target	Var	Var Rate %	#Occ	Var Rate %	#Occ
t	t	68.59	714/1041	76.92	870/1131
	tˤ	31.41	327/1041	23.08	261/1131
d	d	88.72	464/523	81.66	788/ 965
	ø	11.28	59/523	21.32	6283/29466
s	s	81.42	412/506	76.62	1275/1664
	ß	18.58	94/506	24.72	18578/75155
tˤ	tˤ	72.00	18/ 25	–	–
	t	28.00	7/ 25	–	–
dˤ	dˤ	61.40	35/57	–	–
	d	38.60	22/57	–	–
sˤ	sˤ	87.10	162/186	–	–
	s	12.90	24/186	–	–

Table E.8: AA and FR emphatization variants in read speech

Consonants		NccFr	
Target	Var	Var Rate %	#Occ
t	t	67.55	44774/66278
	tˤ	32.45	21504/66278
d	d	78.68	23183/29466
	dˤ	21.32	6283/29466
s	s	75.28	56577/75155
	sˤ	24.72	18578/75155

Table E.9: FR of monolingual speech emphatization variants (NccFr corpus)

Consonants		Var rate %		Word position var rate %					
Target	Var	Ovelall	#Occ	Initial	#Occ	Intern	#Occ	Final	#Occ
p	p	89.65	6566/7324	89.65	6566/7324	91.13	5016/5504	87.04	1417/1628
	b	10.35	758/7324	10.35	758/7324	8.87	488/5504	12.96	211/1628
t	t	69.27	133/192	90.09	9051/10047	91.32	2431/2662	92.29	4419/4788
	d	30.73	59/192	9.91	996/10047	8.68	231/2662	7.71	369/4788
k	k	92.71	7393/7974	94.72	4108/4337	90.79	2120/2335	89.48	1165/1302
	g	7.29	581/7974	5.28	229/4337	9.21	215/2335	10.52	137/1302
b	b	85.90	1743/2029	85.29	1061/1244	87.67	505/576	84.69	177/209
	p	14.10	286/2029	14.71	183/1244	12.33	71/576	15.31	32/209
d	d	85.28	6374/7474	85.32	5128/6010	87.53	941/1075	78.41	305/389
	t	14.72	1100/7474	14.68	882/6010	12.47	134/1075	21.59	84/389
ɡ	ɡ	82.55	918/1112	84.30	204/242	84.14	520/618	76.98	194/252
	k	17.45	194/1112	15.70	38/242	15.86	98/618	23.02	58/252
f	f	89.78	2645/2946	91.23	1707/1871	88.25	804/911	81.71	134/164
	v	10.22	301/2946	8.77	164/1871	11.75	107/911	18.29	30/164
s	s	90.22	10029/11116	90.92	5249/5773	91.30	3493/3826	84.84	1287/1517
	z	9.78	1087/11116	9.8	524/5773	8.70	333/3826	15.16	230/1517
ʃ	ʃ	93.69	802/856	94.19	405/430	94.64	265/280	90.41	132/146
	ʒ	6.31	54/856	5.81	25/430	5.36	15/280	9.59	14/146
v	v	92.03	4610/5009	91.32	2314/2534	94.03	2112/2246	80.35	184/229
	f	7.97	99/5009	8.68	220/2534	5.97	134/2246	19.65	45/229
z	z	80.81	2695/3335	90.00	18/20	90.44	927/1025	76.42	1750/2290
	s	19.19	640/3335	10.00	2/20	9.56	98/1025	23.58	540/2290
ʒ	ý	85.54	3430/4010	84.47	2345/2776	91.50	893/976	74.42	192/258
	c	14.46	580/4010	15.53	431/2776	8.50	83/976	25.58	66/258

Table E.10: FR CS voicing variants rates in three word positions (initial, intern, final)

Consonants		Var rate %		Word position var rate %					
Target	Var	Ovelall	#Occ	Initial	#Occ	Intern	#Occ	Final	#Occ
p	p	76.47	39/51	82.76	24/29	73.68	14/19	66.67	2/3
	b	23.53	12/51	17.24	5/29	26.32	5/19	33.33	1/3
t	t	82.33	1230/1494	80.95	595/735	84.42	325/385	82.89	310/374
	d	17.67	264/1494	19.05	140/735	15.58	60/385	17.11	64/374
k	k	82.32	1164/1414	83.41	573/687	87.94	248/282	77.08	343/445
	g	17.68	250/1414	16.59	114/687	12.06	34/282	22.92	102/445
b	b	69.96	750/1072	74.69	422/565	63.70	265/416	69.23	63/91
	p	30.04	322/1072	25.31	143/565	36.30	151/416	30.77	28/91
d	d	63.62	731/1149	61.74	142/230	68.37	454/664	52.94	135/255
	t	36.38	418/1149	38.26	88/230	31.63	210/664	47.06	120/255
g	g	65.56	59/90	67.16	45/67	63.64	14/22	100.00	1/1
	k	34.44	31/90	32.84	22/67	36.36	8/22	0.00	0/1
f	f	74.01	356/481	86.21	100/116	71.49	158/221	68.06	98/144
	v	25.99	125/481	13.79	16/116	28.51	63/221	31.94	46/144
s	s	79.35	584/736	84.29	118/140	81.97	382/466	64.62	84/130
	z	20.65	152/736	15.71	22/140	18.03	84/466	35.38	46/130
ʃ	ʃ	83.69	802/856	84.19	405/430	84.64	265/280	80.41	132/146
	ʒ	16.31	54/856	15.81	25/430	15.36	15/280	19.59	14/146
v	v	84.62	66/78	83.33	15/18	85.00	51/60	0.00	0/0
	f	15.38	12/78	16.67	3/18	15.00	9/60	0.00	0/0
z	z	77.55	570/735	82.05	96/117	76.00	456/600	100.00	18/18
	s	22.45	165/735	17.95	21/117	24.00	144/600	0.00	0/18
ʒ	ʒ	80.00	510/653	82.05	85/113	80.00	235/290	78.00	200/250
	ʃ	20.00	143/653	17.95	15/113	20.00	55/290	19.00	50/250

Table E.11: AA CS voicing variants rates in three word positions (initial, intern, final)

Consonants		Var rate %		Word position var rate %					
Target	Var	Ovelall	#Occ	Initial	#Occ	Intern	#Occ	Final	#Occ
p	p	66.67	42/63	66.67	42/63	—	—	—	—
	b	33.33	21/63	33.33	21/63	—	—	—	—
t	t	84.70	908/1072	87.87	268/305	89.21	372/417	76.57	268/350
	d	15.30	164/1072	12.13	37/305	10.79	45/417	23.43	82/350
k	k	95.82	711/742	99.14	230/232	91.67	55/60	94.67	426/450
	g	4.18	31/742	0.86	2/232	8.33	5/60	5.33	24/450
b	b	79.29	601/758	80.00	444/555	90.20	46/51	73.03	111/152
	p	20.71	157/758	20.00	111/555	9.80	5/51	26.97	41/152
d	d	59.85	313/523	—	—	72.84	236/324	61.31	122/199
	t	40.15	210/523	—	—	27.16	88/324	38.69	77/199
g	g	—	—	—	—	—	—	—	—
	k	—	—	—	—	—	—	—	—
f	f	88.89	264/297	87.80	108/123	98.31	58/59	85.22	98/115
	v	11.11	33/297	12.20	15/123	1.69	1/59	14.78	17/115
s	s	93.48	946/1012	97.10	402/414	91.55	542/592	66.67	4/6
	z	6.52	66/1012	2.90	12/414	8.45	50/592	33.33	2/6
ʃ	ʃ	69.73	599/859	71.00	163/230	70.00	343/490	68.00	108/159
	ʒ	29.27	260/859	30.00	67/230	31.00	148/490	29.00	17/53
v	v	—	—	—	—	—	—	—	—
	f	—	—	—	—	—	—	—	—
z	z	82.79	303/366	81.82	162/198	83.93	141/168	—	—
	s	17.21	63/366	18.18	36/198	16.07	27/168	—	—
ʒ	ʒ	77.55	570/735	82.05	96/117	76.00	456/600	100.00	18/18
	ʃ	22.45	165/735	17.95	21/117	24.00	144/600	0.00	0/0

Table E.12: AA read speech voicing variants rates in three word positions (initial, intern, final)

Consonants		Var rate %		Word position var rate %					
Target	Var	Ovelall	#Occ	Initial	#Occ	Intern	#Occ	Final	#Occ
p	p	92.74	3346/3608	92.86	2446/2634	92.49	899/972	50.00	1/2
	b	7.26	262/3608	7.14	188/2634	7.51	73/972	50.00	1/2
t	t	90.53	2332/2576	96.04	727/757	91.49	1182/1292	80.27	423/527
	d	9.47	244/2576	3.96	30/757	8.51	110/1292	19.73	104/527
k	k	96.63	1492/1544	97.05	1151/1186	92.50	111/120	96.64	230/238
	g	3.37	52/1544	2.95	35/1186	7.50	9/120	3.36	8/238
b	b	86.17	430/499	86.17	430/499	—	—	—	—
	p	13.83	69/499	13.83	69/499	—	—	—	—
d	d	88.58	1706/1926	88.11	1489/1690	92.31	216/234	50.00	1/2
	t	11.42	220/1926	11.89	201/1690	7.69	18/234	50.00	1/2
g	g	80.00	8/10	80.00	8/10	—	—	—	—
	k	20.00	2/10	20.00	2/10	—	—	—	—
f	f	87.46	614/702	86.52	199/230	87.92	415/472	—	—
	v	12.54	88/702	13.48	31/230	12.08	57/472	—	—
s	s	92.73	3085/3327	91.51	1605/1754	94.74	720/760	93.48	760/813
	z	7.27	242/3327	8.49	149/1754	5.26	40/760	6.52	53/813
ʃ	ʃ	91.91	989/1076	90.13	539/598	98.95	189/191	90.94	261/287
	ʒ	8.09	87/1076	9.87	59/598	1.05	2/191	9.06	26/287
v	v	94.61	1704/1801	91.46	514/562	96.00	1057/1101	96.38	133/138
	f	5.39	97/1801	8.54	48/562	4.00	44/1101	3.62	5/138
z	z	90.62	1392/1536	100.00	2/2	93.94	667/710	87.74	723/824
	s	9.38	144/1536	0.00	0/2	6.06	43/710	12.26	101/824
ʒ	ʒ	87.93	1508/1715	88.21	1182/1340	86.93	326/375	—	—
	ʃ	12.07	207/1715	11.79	158/1340	13.07	49/375	—	—

Table E.13: FR read speech voicing variants rates in three word positions (initial, intern, final)

Bibliography

Abainia, K. (2019). Dzdc12: a new multipurpose parallel algerian arabizi–french code-switched corpus. *Language Resources and Evaluation*, pages 1–37.

Abidi, K., Menacer, M.-A., and Smaili, K. (2017). Calyou: A comparable spoken algerian corpus harvested from youtube. In *18th Annual Conference of the International Communication Association (Interspeech).*

Abidi, K. and Smaïli, K. (2017). An empirical study of the algerian dialect of social network. In *ICNLSSP 2017-International Conference on Natural Language, Signal and Speech Processing*.

Abidi, K. and Smaïli, K. (2018). An automatic learning of an algerian dialect lexicon by using multilingual word embeddings. In *11th edition of the Language Resources and Evaluation Conference, LREC 2018*.

Adda-Decker, M. (2006). De la reconnaissance automatique de la parole à l'analyse linguistique de corpus oraux. *JEP*, pages 389–400.

Adda-Decker, M., Boula de Mareüil, P., and Lamel, L. (1999a). Pronunciation variants in french: schwa & liaison. In *Proceedings of the XIVth International Congress of Phonetic Sciences*, pages 2239–2242.

Adda-Decker, M., DELAIS-ROUSSARI, E., FOURGERON, C., Gendrot, C., and Lamel, L. (1999b). Etude sur grand corpus de la liaison dans la parole spontanée familière.

Adda-Decker, M. and Lamel, L. (1999). Pronunciation variants across system configuration, language and speaking style. *Speech Communication*, 29(2-4):83–98.

Adda-Decker, M. and Lamel, L. (2017). Discovering speech reductions across speaking styles and languages.

Adda-Decker, M. and Snoeren, N. D. (2011). Quantifying temporal speech reduction in french using forced speech alignment. *Journal of Phonetics*, 39(3):261–270.

Ahmed, B. H. and Tan, T.-P. (2012). Automatic speech recognition of code switching speech using 1-best rescoring. In *2012 International Conference on Asian Language Processing*, pages 137–140. IEEE.

Al-Ani, S. H. (1970). *Arabic phonology: An acoustical and physiological investigation*, volume 61. Walter de Gruyter.

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.

Al-Masri, M. and Jongman, A. (2004). Acoustic correlates of emphasis in jordanian arabic: Preliminary results. In *Proceedings of the 2003 Texas Linguistics Society Conference. Somerville, MA: Cascadilla Proceedings Project*, pages 96–106.

Al-Qaysi, N. and Al-Emran, M. (2017). Code-switching usage in social media: a case study from oman. *International Journal of Information Technology and Language Studies*, 1(1):25–38.

Al-Solami, M. (2013). Arabic emphatics: Phonetic and phonological remarks. *Open Journal of Modern Linguistics*, 3(04):314.

Al-Tairi, H., Watson, C., and Brown, J. (2016). Secondary tongue retraction in arabic emphatics: An acoustic study. In *16th Australasian International Conference on Speech Science and Technology (SST2016).*

Al-Tamimi, F. and Heselwood, B. (2011). Nasoendoscopic, videofluoroscopic and acoustic study of plain and emphatic coronals in jordanian arabic. *Instrumental studies in Arabic phonetics*, 319:165.

Al-Tamimi, J. and Khattab, G. (2011). Multiple cues for the singleton-geminate contrast in lebanese arabic: Acoustic investigation of stops and fricatives. In *ICPhS*, pages 212–215.

Allouache, M. (1994). Bab el-oued city. Algeria: Flash back audiovisual, France: les Matins films. Jacques Bidou, Jean-Pierre Gallèpe et Yacine Djadi.

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.

Amazouz, D., Adda-Decker, M., and Lamel, L. (2017). Addressing Code-Switching in French/Algerian Arabic Speech. In *Proc. ISCA Interspeech 2017*, pages 62–66.

Amazouz, D., Adda-Decker, M., and Lamel, L. (2018). The French-Algerian Code-Switching Triggered audio corpus (FACST). In *Proc. Eleventh International Conference on Language Resources and Evaluation LREC 2018*, pages 1468–1473.

Aref, M. and Aref, M. (2015). Un code-switching inédit en classe de langue: la déromanisation graphique et morphosyntaxique de la l2. *Canadian Modern Language Review*, 71(4):406–440.

Auer, P. (1984). *Bilingual conversation*. John Benjamins.

Auer, P. (1995). The pragmatics of code-switching : a sequential approach. *One speaker, two languages: Cross-disciplinary perspectives on code-switching*, pages 115–135.

Auer, P. (1996). Bilingual conversation, dix ans après. *Acquisition et interaction en langue étrangère*, 7:9–34.

Auer, P. (2005). A postscript: Code-switching and social identity. *Journal of pragmatics*, 37(3):403–410.

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

Auer, P. (2013). *Code-switching in conversation: Language, interaction and identity*. Routledge.

Balukas, C. and Koops, C. (2015). Spanish-english bilingual voice onset time in spontaneous code-switching. *International Journal of Bilingualism*, 19(4):423–443.

Barkat, M. (2000). *Détermination d'indices acoustiques robustes pour l'identification automatique des parlers arabes*. PhD thesis, Université Lumière Lyon-2.

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.

Basnight-Brown, D. M. and Altarriba, J. (2007). Code-switching and code-mixing in bilinguals: Cognitive, developmental, and empirical approaches. *Nova Science Publishers*.

Beattie, G., Cutler, A., and Pearson, M. (1982). Why is Mrs. Thatcher interrupted so often? *Nature*, 300:744–747.

Benamrane, A. (2013). *Etude acoustique des fricatives de l'arabe standard (locuteurs algériens)*. PhD thesis, Université de Strasbourg.

Benhattab, A. L. (June,2016). Algerian arabic and french code switching as a linguistic strategy in algerian manga. *International Journal of Language and Linguistics*, 3.3:73–78.

Bentahila, A. and Davies, E. E. (2002). Language mixing in rai music: Localisation or globalisation?. *Language & Communication*, 22(2):187–207.

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.

Blom, J.-P. (1972). Social meaning in linguistic structure: Code-switching in norway. jj gumperz, d. hymes, eds. directions in sociolinguistics: The ethnography of communication.

Blom, J.-P. and Gumperz, J. J. (2000). Social meaning in linguistic structure: Code-switching in norway. *The bilingualism reader*, pages 111–136.

Boersma, P. et al. (2002). Praat, a system for doing phonetics by computer. *Glot international*, 5.

Boersma, Paul., W. D. (2017). Praat: doing phonetics by computer. [Computer program], version 6.0.46, copyright 1995-2017.

Botero, C., Bullock, B., Davis, K., and Toribio, A. J. (2004). Perseverative phonetic effects in bilingual code-switching. In *34th Linguistic Symposium on Romance Languages, University of Utah, March*.

Boula de Mareüil, P., Vieru-Dimulescu, B., Woehrling, C., and Adda-Decker, M. (2008). Accents étrangers et régionaux en français. *Traitement Autom. Lang*, 49(3):135–163.

Boumans, L. P. C. (1998). *The syntax of code-switching: Analysing Moroccan Arabic/Dutch conversation*. Tilburg: Tilburg University Press.

Boxberger, L. (1981). Acoustic characteristics of arabic pharyngeal and pharyngealized consonants. *Kansas Working Papers in Linguistics*, 06.

Boyer, H. (2001). L'incontournable paradigme des représentations partagées dans le traitement de la compétence culturelle en français langue étrangère. *Ela. Études de linguistique appliquée*, 2(3):333–340.

Broersma, M. (2011). Triggered code-switching: Evidence from picture naming experiments. *Modeling bilingualism from structure to chaos: In honor of Kees de Bot*, pages 37–57.

Broersma, M. and De Bot, K. (2006). Triggered codeswitching: A corpus-based evaluation of the original triggering hypothesis and a new alternative. *Bilingualism: Language and cognition*, 9(1):1–13.

Brugnara, F., Falavigna, D., and Omologo, M. (1993). Automatic segmentation and labeling of speech based on hidden markov models. *Speech Communication*, 12(4):357–370.

Buckwalter, T. (2002). Arabic transliteration. www.qamus.org/transliteration.html/, [Online; accessed February 4, 2020].

Bullock, B. E. (2009). Phonetic reflexes of code-switching. In Ludmila Isurin, D. W. and de Bot, K., editors, *Multidisciplinary Approaches to Code Switching*, chapter 10, pages 163–181. Cambridge University Press, Cambridge.

Bullock, B. E. and Toribio, A. J. (2009). Themes in the study of code-switching. In Bullock, B. and Toribio, A., editors, *The Cambridge Handbook of Linguistic Code-switching*, chapter 1, pages 1–19. Cambridge University Press, Cambridge.

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

Caramazza, A., Yeni-Komshian, G., and Zurif, E. (1974). Bilingual switching: The phonological level. *Canadian Journal of Psychology/Revue canadienne de psychologie*, 28(3):310.

Caubet, D. (2002). Métissages linguistiques ici (en france) et là-bas (au maghreb). *Ville-école-intégration enjeux*, 130:117–132.

Çetinoğlu, Ö. (2017). A Code-Switching Corpus of Turkish-German Conversations. In *Proceedings of the 11th Linguistic Annotation Workshop*, pages 34–40.

Chan, J. Y., Cao, H., Ching, P., and Lee, T. (2009). Automatic recognition of cantonese-english code-mixing speech. *International Journal of Computational Linguistics & Chinese Language Processing, Volume 14, Number 3, September 2009*, 14(3).

Chan, J. Y., Ching, P., and Lee, T. (2005). Development of a cantonese-english code-mixing speech corpus. In *Ninth European Conference on Speech Communication and Technology*.

Childers, D. G. (1978). *Modern spectrum analysis*. IEEE Computer Society Press.

Clyne, M. (1987). Constraints on code switching: How universal are they? *Linguistics*, 25(1):739–64.

Clyne, M., Clyne, M. G., and Michael, C. (2003). *Dynamics of language contact: English and immigrant languages*. Cambridge University Press.

Clyne, M. G. (1980). Triggering and language processing. *Canadian Journal of Psychology/Revue canadienne de psychologie*, 34(4):400–06.

Cooper, W. E. and Paccia-Cooper, J. (1980). *Syntax and speech*. Harvard University Press, 1 edition.

Cotterell, R. and Callison-Burch, C. (2014). A multi-dialect, multi-genre corpus of informal written arabic. In *LREC*, pages 241–245.

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.

Crystal, D. (2011). *A dictionary of linguistics and phonetics*, volume 30. John Wiley & Sons.

De St Exupéry, A. (1943). *Le petit prince*. New York, NY: Reynal & Hitchcock.

Delattre, P. (1966). *Studies in French and comparative phonetics: selected papers in French and English*, volume 18. Walter de Gruyter GmbH & Co. KG.

Delattre, P. (1969). An acoustic and articulatory study of vowel reduction in four languages. *IRAL-International Review of Applied Linguistics in Language Teaching*, 7(4):295–326.

Delattre, P. (1971). Consonant gemination in four languages: an acoustic, perceptual, and radiographic study part i. *IRAL-International Review of Applied Linguistics in Language Teaching*, 9(1):31–52.

Dell, F. and Elmedlaoui, M. (2012). *Syllables in tashlhiyt berber and in Moroccan Arabic*, volume 2. Springer Science & Business Media.

Delvaux, V., Metens, T., and Soquet, A. (2002). Propriétés acoustiques et articulatoires des voyelles nasales du français. *XXIVèmes Journées d'étude sur la parole, Nancy*, 1:348–352.

Derivery, N. (1997). *La phonétique du français*. FeniXX.

Derradji, Y., Queffélec, A., and Smaali-Dekdouk, C.-B. Y. (2002). le français en algérie: lexique et dynamique des langues.

Deuchar, M., Davies, P., Herring, J., Couto, M. C. P., and Carter, D. (2014). Building bilingual corpora. *Advances in the Study of Bilingualism*, pages 93–111.

Dey, A. and Fung, P. (2014). A hindi-english code-switching corpus. In *LREC*, pages 2410–2413.

Dolmazon, J.-M., Bimbot, F., Adda, G., El-Beze, M., Caërou, J.-C., Zeiliger, J., and Adda-Decker, M. (1997). Organisation de la premiere campagne aupelf pour l'évaluation des systemes de dictée vocale. *Journées Scientifiques et Techniques du Réseau Francophone d'Ingénierie de la Langue de l'AUPELF-UREF*, pages 13–18.

Ezeizabarrena, M.-J. and Aeby, S. (2010). Les phénomènes de code-switching dans les conversations adulte-enfant (s) en basque-espagnol: une approche syntaxique. *Corpus*, 3(9).

Fagyal, Z., Kibbee, D., and Jenkins, F. (2006). *French: A linguistic introduction*. Cambridge University Press.

Fant, G. (1960). Acoustic theory of speech production (mouton, the hague)(1970). *The closely spaced horizontal lines shown in Fig. 1A are the harmonics of the fundamental frequency of phonation, and are typically revealed in narrowband spectrograms.*

Ferrat, K. and Guerti, M. (2013). Classification of the arabic emphatic consonants using time delay neural network. *International Journal of Computer Applications*, 80(10).

Ferrat, K. and Guerti, M. (2017). An experimental study of the gemination in arabic language. *Archives of Acoustics*, 42(4):571–578.

Forney, G. D. (1973). The viterbi algorithm. *Proceedings of the IEEE*, 61(3):268–278.

Fougeron, C. and Smith, S. (1999). Illustrations of the IPA: French. *Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet*, pages 78–81.

Foulkes, P., Scobbie, J. M., and Watt, D. (2010). Sociophonetics. *The handbook of phonetic sciences*, pages 703–754.

Franco, J. C. and Solorio, T. (2007). Baby-steps towards building a spanglish language model. In *International conference on intelligent text processing and computational linguistics*, pages 75–84. Springer.

Gardner-Chloros, P. (1991). *Language selection and switching in Strasbourg*. Clarendon Press.

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

Gardner-Chloros, P. (2009b). *Sociolinguistic factors in code-switching*. Cambridge University Press.

Gardner-Chloros, P., Charles, R., and Cheshire, J. (2000). Parallel patterns? a comparison of monolingual speech and bilingual codeswitching discourse. *Journal of Pragmatics*, 32(9):1305–1341.

Gauvain, J.-L., Lamel, L., and Adda, G. (2002). The LIMSI broadcast news transcription system. *Speech communication*, 37(1-2):89–108.

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.

Gelly, G., Gauvain, J.-L., Lamel, L., Laurent, A., Le, V. B., and Messaoudi, A. (2016). Language recognition for dialects and closely related languages. *Odyssey, Bilbao, Spain*.

Gendrot, C. and Adda-Decker, M. (2005). Impact of duration on f1/f2 formant values of oral vowels: an automatic analysis of large broadcast news corpora in french and german. In *Ninth European Conference on Speech Communication and Technology*.

Ghazeli, S. (1978). *Back consonants and backing coarticulation in Arabic*. PhD thesis, University of Texas at Austin.

Giannini, A. and Pettorino, M. (1982). The emphatic consonants in arabic. *Speech Laboratory Report*.

Gravier, G., Bonastre, J.-F., Geoffrois, E., Galliano, S., McTait, K., and Choukri, K. (2004). The ester evaluation campaign for the rich transcription of french broadcast news. In *LREC*.

Grosjean, F. (1982). *Life with two languages: An introduction to bilingualism*. Harvard University Press.

Grosjean, F. (2001). The bilingual's language modes. *One mind, two languages: Bilingual language processing*, 122.

Grosjean, F. (2008). *Studying bilinguals*. Oxford University Press, USA.

Gross, S. (2006). Code switching//encyclopedia of language and linguistics.

Guella, N. (2011). Emprunts lexicaux dans des dialectes arabes algériens. *Synergies Monde arabe*, 8:81–88.

Gumperz, J. J. (1977). The sociolinguistic significance of conversational code-switching. *RELC journal*, 8(2):1–34.

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

Hadjadj, C. (2015). *An ivestigation of language use in arabiccaricature in algerian newspapers*. PhD thesis, Universite of Tlemcen.

Hallé, P. A. and Adda-Decker, M. (2007). Voicing assimilation in journalistic speech. In *16th International congress of phonetic sciences*, pages 493–496.

Ham, W. (2013). *Phonetic and phonological aspects of geminate timing*. Routledge.

Hamed, I., Elmahdy, M., and Abdennadher, S. (2018). Collection and analysis of code-switch egyptian arabic-english speech corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*.

Harrat, S., Meftouh, K., Abbas, M., and Smaili, K. (2014). Building resources for algerian arabic dialects. In *15th Annual Conference of the International Communication Association Interspeech*.

Harrat, S., Meftouh, K., Abbas, M., and Smaïli, K. (2016). An Algerian dialect: Study and resources. *International Journal of Advanced Computer Science and Applications-IJACSA*, 7(3).

Hartsuiker, R. J., Pickering, M. J., and Veltkamp, E. (2004). Is syntax separate or shared between languages? cross-linguistic syntactic priming in spanish-english bilinguals. *Psychological Science*, 15(6):409–414.

Hassan, Z. M. (1981). *An experimental study of vowel duration in Iraqi spoken Arabic*. PhD thesis, University of Leeds.

Hassan, Z. M. (2003). Temporal compensation between vowel and consonant in swedish & arabic in sequences of cv: C & cvc: and the word overall duration. *PHONUM*, 9:45–48.

Hasselmo, N. (1961). American swedish, harvard university. *Unpublished Ph. D. thesis*.

Hoffmann, C. (1991). An introduction to bilingualism.

Houdebine, A.-M. (1981). Introduction à la phonétique du français.

Hymes, D. (1962). The ethnography of speaking', in t. gladwin and wc sturtevant (eds), anthropology and human behavior. washington, dc: Anthropological society of washington.

Imseng, D., Bourlard, H., Caesar, H., Garner, P. N., Lecorvé, G., and Nanchen, A. (2012). Mediaparl: Bilingual mixed language accented speech database. In *2012 IEEE Spoken Language Technology Workshop (SLT)*, pages 263–268. IEEE.

Jake, J. L., Myers-Scotton, C., and Gross, S. (2002). Making a minimalist approach to codeswitching work: Adding the matrix language. *Bilingualism: language and cognition*, 5(1):69–91.

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

Kees de Bot, M. B. and Isurin, L. (2009). Sources of triggering in code switchings. In Isurin, L., Winford, D., and De Bot, K., editors, *Multidisciplinary approaches to code switching*, chapter 4, pages 85–102. John Benjamins Publishings, Amsterdam / Philadelphia.

Kenstowicz, M. J. (1994). *Phonology in generative grammar*, volume 7. Blackwell Cambridge, MA.

Khattab, G. (2007). A phonetic study of gemination in lebanese arabic. *ICPhS XVI Proc*, pages 153–158.

Khattab, G. (2009). Phonetic accommodation in children's code-switching. In Barbara E. Bullock, A. J. T., editor, *The Cambridge Handbook of Linguistic Code-switching*, chapter 9, pages 142–159. Cambridge University Press, Cambridge.

Kheder, S. and Kaan, E. (2016). Processing code-switching in algerian bilinguals: Effects of language use and semantic expectancy. *Frontiers in psychology*, 7.

Kiparsky, P. (2003). Syllables and moras in Arabic. *The syllable in optimality theory*, pages 147–182.

Kiss, Z. G. (2013). Measuring acoustic correlates of voicing in stops and fricatives. *VL1xx-Papers in linguistics presented to László Varga on his 70th Birthday*, pages 289–312.

Kootstra, G. J., Stell, G., and Yakpo, K. (2015). A psycholinguistic perspective on code-switching: Lexical, structural, and socio-interactive processes. *Code-switching between structural and sociolinguistic perspectives*, pages 39–64.

Kurdi, M. Z. (2016). *Natural Language Processing and Computational Linguistics: Speech, Morphology and Syntax*. Cognitive Science. Wiley-ISTE, 1 edition.

Ladefoged, P. (2003). Phonetic fieldwork. In *Proc. 15th ICPhS*, pages 203–206, Barcelona.

Ladefoged, P. and Disner, S. F. (2012). *Vowels and consonants*. John Wiley & Sons.

Ladefoged, P. and Johnson, K. (2014). *A course in phonetics*. Nelson Education.

Ladefoged, P. and Maddieson, I. (1996). *The sounds of the world's languages*, volume 1012. Blackwell Oxford.

Lahrouchi, M. (2018). Not as you r: Adapting the french r into arabic and berber.

Lamel, L., Gauvain, J.-L., Adda, G., Adda-Decker, M., Canseco, L., Chen, L., Galibert, O., Messaoudi, A., and Schwenk, H. (2004). Speech transcription in multiple languages. In *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*, volume 3, pages iii–757. IEEE.

Lamel, L., Messaoudi, A., and Gauvain, J.-L. (2008). Investigating morphological decomposition for transcription of arabic broadcast news and broadcast conversation data. In *Ninth Annual Conference of the International Speech Communication Association*.

Lamel, L., Messaoudi, A., and Gauvain, J.-L. (2009). Automatic speech-to-text transcription in Arabic. *ACM Transactions on Asian Language Information Processing (TALIP)*, 8(4):18.

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.

Li, Y., Yu, Y., and Fung, P. (2012). A mandarin-english code-switching corpus. In *LREC*, pages 2515–2519.

Ljolje, A. and Riley, M. (1991). Automatic segmentation and labeling of speech. In *Acoustics, Speech, and Signal Processing, 1991. ICASSP-91., 1991 International Conference on*, pages 473–476. IEEE.

Lüdi, G. (1998). Filecode-switching comme variété mixte. *Sociolinguistica*, 12:140–154.

Lüdi, G., Py, B., and Py, B. (1986). *Etre bilingue*. P. Lang.

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., Lyu, R.-Y., Chiang, Y.-c., and Hsu, C.-N. (2006). Speech recognition on code-switching among the chinese dialects. In *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, volume 1, pages I–I. IEEE.

Lyu, D.-C., Tan, T.-P., Chng, E. S., and Li, H. (2010). Seame: a mandarin-english code-switching speech corpus in south-east asia. In *Eleventh Annual Conference of the International Speech Communication Association*.

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.

Macnamara, J. (1967). Problems of bilingualism. *Journal of social issues*, 23(2):n2.

Malmberg, B. (1944). La coupe syllabique dans le système consonantique du français. *Acta linguistica*, 4(1):61–66.

Mareüil, P. B. d. and Adda-Decker, M. (2002). Studying pronunciation variants in French by using alignment techniques. In *Seventh International Conference on Spoken Language Processing*.

Mariani, J. (1990). Reconnaissance automatique de la parole: progrès et tendances. *TS. Traitement du signal*, 7(4):239–266.

Matras, Y. (2009). *Language contact*. Cambridge University Press.

Menacer, M. A., Mella, O., Fohr, D., Jouvet, D., Langlois, D., and Smaïli, K. (2017). Development of the arabic loria automatic speech recognition system (alasr) and its evaluation for algerian dialect. *Procedia Computer Science*, 117:81–88.

Meunier, C. (2014). *VARIATION DE LA PAROLE: CONTRAINTES LINGUISTIQUES ET MECANISMES D'ADAPTATION*. PhD thesis, Aix-Marseille Université.

Mirjam Broersma, Ludmila Isurin, S. B. (2009). Triggered code switching: Evidence from dutch – english and russian – english bilinguals. In Isurin, L., Winford, D., and De Bot, K., editors, *Multidisciplinary approaches to code switching*, chapter 5, pages 103–128. John Benjamins Publishings, Amsterdam / Philadelphia.

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*.

Mohamed, Y. (2001). Pharyngealization in arabic: Modelling, acoustic analysis, airflow and perception. *Revue de La Faculté des Lettres El Jadida*, 6:51–70.

Mokhtar, K. (2018). The linguistic friction in algeria. *Sociology International Journal*, 2(2):134–140.

Mondada, L. (2007). Le code-switching comme ressource pour l'organisation de la parole-en-interaction. *Journal of language contact*, 1(1):168–197.

Moore, D. (2002). Code-switching and learning in the classroom. *International journal of bilingual education and bilingualism*, 5(5):279–293.

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

Myers-Scotton, C. (1990). Codeswitching and borrowing: Interpersonal and macrolevel meaning. *Codeswitching as a worldwide phenomenon*, pages 85–110.

Myers-Scotton, C. (1993). *Duelling languages: Grammatical structure in codeswitching*. Oxford University Press.

Myers-Scotton, C. (1995). *Social motivations for codeswitching: Evidence from Africa*. Oxford University Press.

Myers-Scotton, C. (2001). The matrix language frame model: Development and responses. *Trends in Linguistics Studies and Monographs*, 126:23–58.

Niesler, T. et al. (2018). A first south african corpus of multilingual code-switched soap opera speech. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018).*

Ohala, J. J. (1993). The phonetics of sound change. *Historical linguistics: Problems and perspectives*, pages 237–278.

Pfaff, C. (1979). Constraints on language mixing. *Language*, 1(55):291–318.

Piccinini, P. E. (2016). *Cross-language Activation and the Phonetics of Code-switching*. PhD thesis, UC San Diego.

Piccinini, P. E. and Garellek, M. (2014). Prosodic cues to monolingual versus code-switching sentences in english and spanish. In *Proceedings of the 7th Speech Prosody Conference*, pages 885–889. Citeseer.

Poplack, S. (1980). Sometimes i'll start a sentence in spanish y termino en espanol: toward a typology of code-switching1. *Linguistics*, 18(7-8):581–618.

Poplack, S. (2001). Code-switching (linguistic). *International encyclopedia of the social and behavioral sciences*, 12:2062–2065.

Post, R. E. (2015). *The impact of social factors on the use of Arabic-French code-switching in speech and IM in Morocco*. PhD thesis, University of Texas.

Rabiner, L. R. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286.

Richer, J.-J. (2015). *La didactique des langues interrogée par les compétences: Essai sur les sciences du langage*. EME éditions.

Ridouane, R. (2010). Geminates at the junction of phonetics and phonology. *Papers in laboratory phonology*, 10:61–90.

Ritchie, W. C. and Bhatia, T. (2004). Social and psychological factors in language mixing. *The handbook of bilingualism*, 46(13):336.

Ronjat, J. (1913). *Le développement du langage observé chez un enfant bilingue*. H. Champion.

Saadane, H. (2015). *Le traitement automatique de l'arabe dialectalisé: aspects méthodologiques et algorithmiques*. PhD thesis, Université Grenoble Alpes.

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

Sankoff, D. and Poplack, S. (1981). A formal grammar for code-switching. *Research on Language & Social Interaction*, 14(1):3–45.

Sara, S. I. (2007). *Sībawayh on? imālah (inclination): Text, Translation, Notes and Analysis*. Edinburgh University Press.

Schiel, F., Draxler, C., Baumann, A., Ellbogen, T., and Steffen, A. (2012). The production of speech corpora.

Schiffrin, D. (2001). Discourse markers: Language, meaning, and context. *The handbook of discourse analysis*, 1:54–75.

Scotton, C. M. (2006). *Multiple voices: An introduction to bilingualism.* Blackwell Pub.

Scotton, C. M. and Ury, W. (1977). Bilingual strategies: The social functions of code-switching. *International Journal of the Sociology of Language*, 1977(13):5–20.

Shen, H.-P., Wu, C.-H., Yang, Y.-T., and Hsu, C.-S. (2011). Cecos: A chinese-english code-switching speech database. In *2011 International Conference on Speech Database and Assessments (Oriental COCOSDA)*, pages 120–123. IEEE.

Simonin, J. and Wharton, S. (2013). *Sociolinguistique du contact. Dictionnaire des termes et concepts*. ENS éditions.

Snoeren, N. D., Hallé, P. A., and Segui, J. (2006). A voice for the voiceless: Production and perception of assimilated stops in french. *Journal of Phonetics*, 34(2):241–268.

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

Souag, M. L. (2006). Explorations in the Syntactic Cartography of Algerian Arabic. Master's thesis, School of Oriental and African Studies (University of London).

Sreeram, G., Dhawan, K., and Sinha, R. (2018). Hindi-english code-switching speech corpus. *arXiv preprint arXiv:1810.00662*.

Sridhar, S. N. and Sridhar, K. K. (1980). The syntax and psycholinguistics of bilingual code mixing. *Canadian Journal of Psychology/Revue canadienne de psychologie*, 34(4):407.

Stoltz, J. (2011). *L'alternance codique dans l'enseignement du FLE: Étude quantitative et qualitative de la production orale d'interlocuteurs suédophones en classe de lycée*. PhD thesis, Linnaeus University Press.

Torreira, F., Adda-Decker, M., and Ernestus, M. (2010). The Nijmegen corpus of casual French. *Speech Communication*, 52(3):201–212.

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.

Uriel, W. (1953). Languages in contact. *The Hague: Mouton*, 1(1).

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

Wang, L. and Liu, H. (2013). Syntactic variations in chinese–english code-switching. *Lingua*, 123:58–73.

Watson, J. C. (1999). The directionality of emphasis spread in arabic. *Linguistic Inquiry*, 30(2):289–300.

Weinreich, U. (1953). Languages in contact: Problems and findings. *Publications of the Linguistic Circle of New York*, 1(1).

White, C. M., Khudanpur, S., and Baker, J. K. (2008). An investigation of acoustic models for multilingual code-switching. In *Ninth Annual Conference of the International Speech Communication Association*.

Wiedemann, F. (2015). *Code-Switching im algerischen und tunesischen Rap: Eine vergleichende Analyse von Lotfi Double Kanons „Klemi "und Baltis „L'album avant l'albombe "*, volume 6. University of Bamberg Press.

Woolford, E. (1983). Bilingual code-switching and syntactic theory. *Linguistic inquiry*, 14(3):520–536.

Yilmaz, E., Andringa, M., Kingma, S., Dijkstra, J., Kuip, F., Velde, H., Kampstra, F., Algra, J., Heuvel, H., and van Leeuwen, D. A. (2016). A longitudinal bilingual frisian-dutch radio broadcast database designed for code-switching research. In *Portoroz, Slovenia: European Language Resources Association.*

Yılmaz, E., van den Heuvel, H., and van Leeuwen, D. (2016). Investigating bilingual deep neural networks for automatic recognition of code-switching frisian speech. *Procedia Computer Science*, 81:159–166.

Yuan, J., Ryant, N., Liberman, M., Stolcke, A., Mitra, V., and Wang, W. (2013). Automatic phonetic segmentation using boundary models. In *Interspeech*, pages 2306–2310.

Zaboot, T. (2010). La pratique langagière de locuteur (s) bilingue (s)'. *Synergies Algérie*, 9:201–10.

Zentella, A. C. (1997). *Growing up bilingual: Puerto Rican children in New York*. Wiley-Blackwell.

Zeroual, C., Hoole, P., and Fuchs, S. (2006). Etude par transillumination des consonnes occlusives simples et géminées de l'arabe marocain. *Proceeding of the XXVIth Journées d'Etudes sur la parole*.

Zeroual, C., Hoole, P., Gafos, A. I., Sock, R., Fuchs, S., and Laprie, Y. (2008). Spatio-temporal and kinematic study of moroccan arabic coronal geminate plosives. *Proceedings of the 8th ISSP*, pages 133–136.

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

Ziamari, K. (2013). Arabe marocain et français en contact. description morphosyntaxique et comparaison typologique. *Travaux-Cercle linguistique d'Aix-en-Provence*, 24:337–348.

Zribi-Hertz, A. (1988). L'oral, la syntaxe et l'astérisque: questions méthodologiques, avec et sans réponses. *Linx*, 18(1):33–45.

Glossary

Allophone Various phonetic realizations of a phoneme in a language, which do not contribute to distinctions of meaning. Some allophones do not belong to the phonological system of the language. Example:[x] in French [kxistal] "cristal". xxiv, 39

Arabizi transliteration Arabic speech transliterated with Latin characters, it used heavily on the internet and SMS. xxiv, 31

Base language The dominant language in code-switching utterance, is called also the *matrix language* (Myers-Scotton, 1993, p.20), into which elements from the embedded language are inserted. Synonyms: host language (vs. guest language) (Sridhar and Sridhar, 1980). xxiv

Bilingual word L*b*words which are modified and integrated morphologically in the structure of L*a*. It can be bilingual substantives, bilingual verbs, bilingual adverbs or other lexical and grammatical units. xxiv, 66

Embedded language Language elements of a second language inserted into the matrix language (a first language) in code-switching utterance. Synonyms: guest language (vs. host/base language) (Sridhar and Sridhar, 1980). xxiv, 19, 54

L*b* The second language that appears in code-switching speech, it appears too as a a dominated language in the utterance.. xxiv, 10, 11, 12, 13, 14, 19, 32, 51, 58, 78, 90, 100, 162

L*a* The first language that appears in code-switching speech, it appears too as a dominant language of the utterance.. xxiv, 10, 14, 16, 32, 51, 58, 78, 101, 162

Segment A part of signal corresponds to one phone into the audio signal. xxiv

Speech stretch A part of speech that corresponds to the grouped words in one language. xvi, xxi, xxii, xxiv, 57, 58, 66, 68, 69, 73, 78, 79, 80, 81, 82, 87, 90, 91, 100, 101

Word truncation Deletion one or more last syllables of the word in oral or written production. xxiv, 60

Auteurs, Authors: Djegdjiga Amazouz

Titre: Études linguistiques et phonétiques du code-switching français-arabe: analyses de grands corpus et traitement automatique de la parole

Title: Linguistic and phonetic investigations of French-Algerian Arabic code-switching: large corpus studies using automatic speech processing

Résumé

Cette thèse traite du code-switching français-arabe algérien à l'aide d'outils de traitement automatique de la parole. Un corpus de 7h30 de parole de 20 locuteurs bilingues (5h de parole spontanée et 2h30 de parole lue) a été conçu, enregistré et annoté. L'un des premiers défis abordés a consisté à développer des méthodes de traitement des données telles que la segmentation en langues, la transcription du français et de l'arabe algérien. Les méthodes d'alignement automatique de la parole ont été adaptées pour traiter les données du code-switching en combinant deux systèmes d'alignement monolingues, produisant ainsi des transcriptions orthographiques et phonémiques avec des localisations temporelles dans les deux langues. Une expérience a été menée pour détecter automatiquement les changements de langue, mais cela reste un défi, en particulier pour les durées monolingues très courtes. Le second aspect de la recherche porte sur l'influence du système phonologique de la langue *a* sur la deuxième langue *b* dans la parole du code-switching, en l'occurrence les productions phonétiques de l'arabe et du français. Le corpus annoté a été utilisé pour effectuer des études phonétiques sur la variation des voyelles et des consonnes en utilisant un paradigme de discrimination automatique de type ABX. Avec ce paradigme, nos résultats sur la variation de la production correspondent aux hypothèses a priori: considérant les voyelles périphériques /i, a, u/, des taux de variantes plus élevés sont mesurés en arabe algérien (40%) qu'en français (27%). Une comparaison avec des locuteurs de langue maternelle française suggère que les locuteurs bilingues ont des productions de voyelles plus conservatrices que les locuteurs natifs (34%), du moins dans le code-switching. Trois études sur la variation des consonnes ont également été menées: la gemination, l'emphatisation. Globalement, les consonnes présentent des tendances similaires à celles des voyelles: 42% de taux de variantes pour l'arabe algérien et 30% pour le français en code-switching, contre 38% pour les natifs français. De futures études utilisant ce corpus novateur pourront contribuer à démêler l'interaction complexe entre la variation phonétique et les systèmes phonologiques chez les bilingues dans le code-switching.

Mots-clés
Code-switching, parole spontanée, traitement automatique des langues, variation phonétique, grands corpus.

Abstract

This thesis investigated French-Algerian Arabic code-switching using automatic speech processing tools. A corpus of 7h30 of code-switched speech from 20 French-Algerian Arabic speakers (5h of spontaneous speech and 2h30 of read speech) has been designed, recorded and annotated. One of the first challenges tackled consisted of developing data processing methods such as language segmentation, code-switching utterance segmentation as well as transcription in French and Algerian Arabic dialect. Automatic speech alignment methods were adapted to process the code-switched data by combining two monolingual alignment systems thus producing time-stamped orthographic and phonemic transcriptions in both languages. An experiment was conducted to automatically detect language switches, however this remains a challenge especially for small speech stretches. A second aspect of this thesis' research studied the influence of the phonological system of language *a* on the second language *b* in code-switched speech, in this case the phonetic productions of French and Algerian Arabic. The annotated corpus was used to carry out phonetic studies on vowel and consonant variation using an automatic ABX-like phone discrimination paradigm. With this paradigm, our results on variation in code-switched speech vowel productions are in line with a priori hypotheses: considering the peripheral /i,a,u/ vowels, higher variant rates are measured in Algerian Arabic (40%) than in French (27%). A comparison with native French control speakers suggests that the bilingual speakers have more conservative vowel productions than natives (34%), at least in code-switched speech. Three types of consonant variation were also explored: gemination, emphatization and voicing alternation. Overall, consonants show similar trends to vowels: 42% variant rates for Algerian Arabic, 30% for French in code-switched speech, compared with 38% for French natives. Future studies using this innovative corpus will contribute to disentangle the complex interplay between phonetic variation and phonological systems in bilingual code-switching speakers.

Key-words
Code-switching, spontaneous speech, Speech processing, phonetic variation, large speech corpora.

Université Sorbonne Nouvelle - Paris 3
ED 268 - Langage et langues : description, théorisation, transmission
4, rue des irlandais 75005 PARIS