A Multilayered and Data-Driven Method for Exploring Arabic in Multilingual Settings

Stefano Manfredi, Suat Istanbullu

A Multilayered and Data-Driven Method for Exploring Arabic in Multilingual Settings

Stefano Manfredi and Suat Istanbullu

Abstract

Scholars in Arabic dialectology are widely concerned with the linguistic effects of societal bi- and multilingualism. The present chapter intends to illustrate a non-aprioristic and computer-assisted method for the study of Arabic in multilingual settings. Taking examples from two different sociolinguistic situations, we will illustrate new solutions for annotating and analyzing plurilingual corpora by means of a multilayered annotation system based on JAXE.

1. Introduction

Let us suppose that you just got back from your fieldwork on Moroccan Arabic and that you came across the following utterances while annotating your corpus:

(1) *ġādi* *nə-bdā-w* *əl-xədma* *inšaʾḷḷāh* *composition* *kullši*
FUT 1-start-PL DEF-work God_willing composition everything
en même temps *dər-na* *inšaʾḷḷāh* *ğādi* *n-səğğil-u*
at_once do-1PL God_willing FUT 1-record-PL
deux *titres* *chaque semaine* *ʿənd=na* *deux* *titres*
two titles each week at=1PL two titles
xəss-hum *y-tkətb-u* *w* *y-trépéta-w*
have_to-3PL 3M-be_written-PL and 3M-be_repeated-PL
'We'll start working, God willing, composition, everything at once. We arranged to record, God willing. Two tracks every week. We already have two tracks. They still have to be rewritten and rehearsed'. (Caubet 2014)

(2) *ma-kən-t-š* *ʿārəf* *ṛās-i* *ġādi* *nə-bda*
NEG-be-1SG-NEG know\ACT.PTCP head-POSS.1SGFUT 1-start
r-rock *parce que* *kən-t* *ġēr* *ka-n-sməʿ*
DEF=rock because be-1SG only REAL-1-listen
musīqa fḥāl *gāʿ* *ən-nās*
music like every DEF-people
'I didn't know I was going to start playing rock because I only listened music like everybody else'. (Caubet 2014)

If you were an old-fashioned descriptive linguist looking for "pure" Moroccan Arabic, you would probably ignore the previous chunks of spontaneous speech operating an ideological erasure that falls outside the scientific scopes of linguistic analysis. If that is not the case, you will then be wondering about the occurrence of "foreign" lexical and grammatical items in your corpus. For example, you could adopt the diachronic perspective of "contact-induced change" (Weinreich 1953; Thomason and Kaufmann 1998; Heine and Kuteva 2005) and ask yourself whether nouns like *composition* 'composition' in (1) and *rock* 'rock' in (2) should be equally treated as integrated loanwords. If not, what distinguishes *composition* from *rock*? Is this the degree of morpho-phonological integration or the frequency of occurrence in your corpus? Beyond that, how should these items be transcribed? Moreover, is the derived verb *trépét(a)* 'be repeated' in (1) an integrated loanword too? On the contrary, if you embrace the synchronic standpoint of "codeswitching" and/or "codemixing" (Myers-Scotton 1993; Muysken 2000; Eirlys et al. 2013), you will obviously note that adjacent lexical items drawn from French tend to occur sentence-finally in (1). Furthermore, if you are particularly interested in the analysis of clausal structure in Moroccan Arabic, you will have to explain the occurrence of the French subordinator *parce que* 'because' in (2). More generally, when analyzing your corpus, you will be facing the longstanding problem of how contact phenomena should be labelled and conceptualized and you will perhaps reach the conclusion that there are not strong linguistic criteria for distinguishing "borrowing" from "codeswitching" (Clyne 2003; Winford 2005) and that you need new operational categories for analyzing your heterogeneous data.

Indeed, the study of language contact emerged from its traditional historical perspective and it is now undergoing a process of conceptual renewal (Nicolai 2007). Bearing in mind that contact phenomena are always affected by the social circumstances of language contact (Thomason and Kaufman 1998; Winford 2003), it is now widely accepted that multilingualism (i.e. the use of two or more linguistic varieties in the same place at the same time) represents the norm of human communication, rather than an exception (Léglise and Alby 2016). It is in this context that variants and innovations spread within linguistic communities and gradually produce language changes. Against this background, corpus-driven studies taking into account both linguistic and extra-linguistic factors for conceptualizing ongoing contact phenomena remain a major desideratum of contact linguistics. It should be also remarked that the uncritical adoption of a given terminology for labelling contact phenomena (e.g. "borrowing", "codeswitching", "codemixing", "source language", "matrix language", "interference", "calquing", etc.) always implies the adherence to an established theoretical framework and that this could lead to misleading interpretations. This has obviously to do with the limits of a top-down approach structuring empirical data around pre-established categories. In the attempt of eluding this epistemological cul-de-sac, the present chapter intends to illustrate a non-aprioristic and computer-assisted method for annotating and analyzing plurilingual corpora,[1] with a particular focus on Arabic.

Scholars in Arabic (socio)linguistics are widely concerned with the linguistic effects of societal bi- and multilingualism. First, it is commonly agreed that speakers of modern Arabic dialects are involved in a situation of diglossic bilingualism with Modern Standard Arabic and/or with regional standards (Ferguson 1959; Boussofara-Omar 2006; Mejdell 2012). At the beginning of this paragraph, we have already shown typical outputs of Arabic-French bilingualism in a post-colonial North African country. It is also well known that many ethnolinguistic minorities in Arabic-dominant countries in Africa and in the Middle East are gradually shifting to Arabic (Manfredi 2017b; Manfredi and Tosco 2018). Contrariwise, Arabic may represent the ancestral language of minority bilingual communities that are shifting to local dominant languages (Owens 2000). This is the case of the Arabic varieties of central Asia (Ratcliffe 2005) and southeastern Turkey (Arnold 2000). More recently, following the massive migration waves over the last century, Arabic started to be spoken as a heritage language by second-generation migrants all around the world (Bale 2010; Barontini 2013; Istanbullu and

[1] Following Léglise and Alby (2016), we distinguish between multilingual and plurilingual corpora. Multilingual corpora usually include monolingual sub-corpora in different languages, whereas plurilingual corpora give evidence of heterogeneous language practices implying the use of different linguistic resources in multilingual settings.

Léglise 2014; Istanbullu 2017). In such a context, Arabic is involved in an unbalanced contact situation resulting in a language shift towards local dominant languages, while being in contact with other outsider languages. Migration from Eastern Asia to Middle Eastern countries is instead the cause of the emergence of new vehicular means of communication that are generally referred to as Pidgin Arabic (Bizri 2010; Tosco and Manfredi 2013; Avram 2014). Finally yet importantly, we could mention the case of Juba Arabic (*árabi júba*), the Arabic-based pidgincreole spoken in South Sudan, which is involved in a creole-lexifier contact situation with Sudanese Arabic (Versteegh 1993; Manfredi 2017a).

Each of the previous sociolinguistic situations entails different degrees of linguistic heterogeneity (nearly monolingual production, occasional codeswitching, intensive codeswitching, high bilingual proficiency, language mixing, pidginization, creolization and decreolization, cf. Auer 1999) raising important questions about the analysis and annotation of specific contact phenomena. In this chapter, we will illustrate some new solutions for annotating and analyzing corpora of spontaneous speech in contact situations by means of a multilayered annotation system based on JAXE. The chapter is organized as follows. Section 2 presents the main functionalities of the JAXE-based annotation schema. Sections 3 and 4 respectively show the qualitative and the quantitative advantages of this multilayer annotation taking examples from two plurilingual corpora encompassing different varieties of Arabic. Section 5 finally summarizes the assets of adopting this new methodology for the study of Arabic in multilingual settings.

2. A new computational tool for the study of plurilingual corpora

During the last decades, we have assisted to the flourishing of studies in corpus linguistics thanks to the development of a number of purpose-built software. Most of these computational tools[2] are intended to transcribe, tokenize, gloss, and translate texts, possibly indexed with sound. Despite recent efforts for adapting these tools to the analysis of linguistic variation (Nagy and Meyerhoff 2015), they still unfit to provide a unique method for annotating and analyzing synchronic variation in multilingual settings. Here, we present a new JAXE-based annotation system that enables a multilayered analysis of contact phenomena at morphosyntactic, discursive and interactional levels.

JAXE[3] is an open-source text editor in Java configurable with XML schemas.[4] The XML schema we would like to introduce here is named "Corpus-Contact" and it has been developed within the ANR project *CLAPOTY: Towards a multi-model, typological and computer-assisted analysis of contact-induced language change*[5] (; Vaillant and Léglise 2014; Vaillant 2015; Léglise

[2] See the SIL Toolbox/Shoebox software https://www-01.sil.org/computing/toolbox/ (accessed 1 January 2015), the Interlinear Text Editor developed by Michel Jacobson https://michel.jacobson.free.fr/ITE/index_en.html (accessed 1 January 2015), and the ELAN software created by the Max Planck Institute https://tla.mpi.nl/tools/tla-tools/elan/ (accessed 1 January 2015), to name but a few.

[3] The JAXE software https://jaxe.sourceforge.net/fr/ (accessed 1 January 2015) has been created by Damien Guillaume, Soufiane Ayadi, Bodo Tasche, Olivier Kykal, Cyril Dedieu, Léa Guillon, Bertrand Delacretaz, and Sven Kitschke.

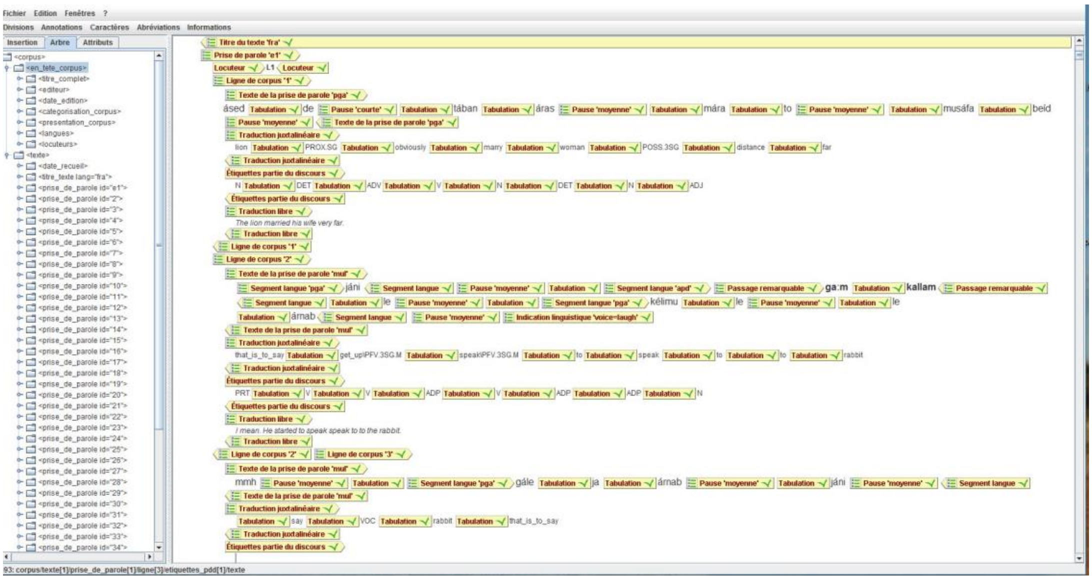
[4] A XML schema describes the building blocks of a XML document. It divides the XML document into a hierarchy of sections, each serving a specific purpose. It enables to separate a document into multiple sections so that they can be rendered differently, or used by a search engine. The elements of a XML schema can be containers, with a combination of text and other elements.

[5] *CLAPOTY: Towards a multi-model, typological and computer-assisted analysis of contact-induced language change*. https://clapoty.vjf.cnrs.fr/index.html (accessed 30 March 2014).

and Alby 2016). Its aim is to provide a common structure for normalizing the annotation and the visualization of plurilingual corpora based on a non-aprioristic approach. This means that the “Corpus-Contact” schema does not align against any pre-established operational category of contact linguistics, but it rather gives the possibility of developing a truly empirical analysis of contact phenomena that can be evaluated a posteriori in the light of different theoretical frameworks. The JAXE-based “Corpus-Contact” schema has been already adopted for annotating and analysing a number of heterogeneous corpora collected in a variety of multilingual settings, as in the case of Thrace Romani in contact with Greek (Adamou 2016), Spanish dialect contact in Colombia (Sánchez Moreano 2015; Léglise and Sánchez Moreano 2017) and trilingual language contact between Casamance Creole, French and Wolof in Senegal (Nunez 2015; Nunez and Léglise 2017). In the following paragraphs, we will introduce the main segmentation and annotation conventions, the description of “remarkable phenomena” as well as their retrievability by means of JAXE.

2.1 Segmentation and annotation

The “Corpus-Contact” schema makes use of specific configuration file for annotating texts with the open-source JAXE editor, as we can see in the following figure.


Fichier Edition Fenêtres ?
Divisions Annotations Caractères Abréviations Informations
Insertion Arbre Attributs

```
<corpus>
  <en_tete_corpus>
    <titre_complet>
    <editeur>
    <date_edition>
    <categorisation_corpus>
    <presentation_corpus>
    <langues>
    <locuteurs>
  <texte>
    <date_recueil>
    <titre_texte lang="fra">
    <prise_de_parole id="e1">
    <prise_de_parole id="2">
    <prise_de_parole id="3">
    <prise_de_parole id="4">
    <prise_de_parole id="5">
    <prise_de_parole id="6">
    <prise_de_parole id="7">
    <prise_de_parole id="8">
    <prise_de_parole id="9">
    <prise_de_parole id="10">
    <prise_de_parole id="11">
    <prise_de_parole id="12">
    <prise_de_parole id="13">
    <prise_de_parole id="14">
    <prise_de_parole id="15">
    <prise_de_parole id="16">
    <prise_de_parole id="17">
    <prise_de_parole id="18">
    <prise_de_parole id="19">
    <prise_de_parole id="20">
    <prise_de_parole id="21">
    <prise_de_parole id="22">
    <prise_de_parole id="23">
    <prise_de_parole id="24">
    <prise_de_parole id="25">
    <prise_de_parole id="26">
    <prise_de_parole id="27">
    <prise_de_parole id="28">
    <prise_de_parole id="29">
    <prise_de_parole id="30">
    <prise_de_parole id="31">
    <prise_de_parole id="32">
    <prise_de_parole id="33">
    <prise_de_parole id="34">
```

```
Titre du texte 'fra'
Prise de parole 'e1'
 Locuteur L1 Locuteur
 Ligne de corpus '1'
  Texte de la prise de parole 'pga'
   ásed Tabulation de Pause 'courte' Tabulation tában Tabulation áras Pause 'moyenne' Tabulation mára Tabulation to Pause 'moyenne' Tabulation musáfa Tabulation beid
  Pause 'moyenne' Texte de la prise de parole 'pga'
  Traduction juxtalinéaire
   lion Tabulation PROX.SG Tabulation obviously Tabulation marry Tabulation woman Tabulation POSS.3SG Tabulation distance Tabulation far
  Traduction juxtalinéaire
  Étiquettes partie du discours
   N Tabulation DET Tabulation ADV Tabulation V Tabulation N Tabulation DET Tabulation N Tabulation ADJ
  Étiquettes partie du discours
  Traduction libre
   The lion married his wife very far.
  Traduction libre
 Ligne de corpus '1'
 Ligne de corpus '2'
  Texte de la prise de parole 'mul'
   Segment langue 'pga' jáni Segment langue Pause 'moyenne' Tabulation Segment langue 'apd' Passage remarquable ga:m Tabulation kallam Passage remarquable
   Segment langue Tabulation le Pause 'moyenne' Tabulation Segment langue 'pga' kélimu Tabulation le Pause 'moyenne' Tabulation le
   Tabulation árnab Segment langue Pause 'moyenne' Indication linguistique 'voice=laugh'
  Texte de la prise de parole 'mul'
  Traduction juxtalinéaire
   that_is_to_say Tabulation get_up\PFV.3SG.M Tabulation speak\PFV.3SG.M Tabulation to Tabulation speak Tabulation to Tabulation to Tabulation rabbit
  Traduction juxtalinéaire
  Étiquettes partie du discours
   PRT Tabulation V Tabulation V Tabulation ADP Tabulation V Tabulation ADP Tabulation ADP Tabulation N
  Étiquettes partie du discours
  Traduction libre
   I mean. He started to speak speak to to the rabbit.
  Traduction libre
 Ligne de corpus '2' Ligne de corpus '3'
  Texte de la prise de parole 'mul'
   mmh Pause 'moyenne' Tabulation Segment langue 'pga' gále Tabulation ja Tabulation árnab Pause 'moyenne' Tabulation jáni Pause 'moyenne' Segment langue
  Texte de la prise de parole 'mul'
  Traduction juxtalinéaire
   Tabulation say Tabulation VOC Tabulation rabbit Tabulation that_is_to_say
  Traduction juxtalinéaire
  Étiquettes partie du discours
```

93: corpus/texte[1]/prise_de_parole[1]/ligne[3]/etiquettes_pdd[1]/texte

Figure 1. Configuration file in JAXE

In order to be interoperable with other annotated corpora, the “Corpus-Contact” annotation schema is largely inspired by the TEI[6] standards. It adopts the Unicode encoding for transcribing texts, ISO-639 codes for identifying languages, and a XML document markup for exporting documents. Besides, it also offers a number of technical innovations for annotating plurilingual corpora.

One of the foremost problems in annotating oral corpora concerns the definition of the units for segmenting the speech flow. Most oral corpora display a traditional syntactic segmentation into sentences. Other corpora propose a segmentation based on prosodic

[6] *Text Encoding Initiative*. https://www.tei-c.org (accessed 1 January 2015).

boundaries, which do not always correspond to syntactic boundaries (Mettouchi et al. 2015). The "Corpus-Contact" schema, on its part, segments the speech flow into *prises de parole* (Eng. 'speech-turns')*,* which represent the basic unit for an interactional analysis of spontaneous speech (Vaillant and Léglise 2014: 91)*.* Each speech-turn is attributable to one speaker and is subdivided into four annotation tiers: transcription,[7] morpheme-by-morpheme gloss, parts of speech and free translation.

As a further matter, annotated corpora usually mark words or phrases that are not in the main language of the text (i.e. codeswitching/codemixing) by means of chevrons <…> (Manfredi et al. 2015). However, in the case of plurilingual corpora, it is not rare that several languages occur in the same speech-turn. In this regard, the "Corpus-Contact" schema is different from any other annotation procedure in that it makes a basic distinction between monolingual and multilingual segments. The idea is that the language of each morphological or syntactic segment can be described by a XML attribute (called *XML.lang*, Vaillant and Léglise 2014: 92) bearing an ISO-639 code. If the language of a segment can be unequivocally identified, then we are dealing with a monolingual segment requiring only the *XML.lang* attribute*.* In contrast, if a segment may be linked to different languages, then an optional element (called *langues*) can be added for associating different languages to the segment in hand. Multilingual segments may be defined both syntagmatically and paradigmatically (Vaillant 2015). In the former case, a multilingual speech segment is morphosyntactically mixed to such an extent that it does not allow identifying a "matrix" against an "embedded" language (Myers-Scotton 1993). On a paradigmatic level, a given item is considered multilingual when it has similar phonetic forms in different languages and it does not provide enough constraints as to what language it should be assigned. For cases where two languages share a number of segmental features, such as a creole and its lexifier, a double transcription, referred to as "floating" (Ledegen 2012, cf. 3), is also possible. The choice to mark those segments as multilingual allows identifying all the available interpretations that might apply to them (Léglise 2018).

2.2 Analysis of "remarkable phenomena"

The main methodological innovation of the "Corpus-contact" annotation schema resides in its treatment of contact phenomena found in plurilingual corpora. These may be related to internally and externally induced variation or to contact-induced linguistic behaviors such as codeswitching or codemixing (Léglise and Alby 2016: 365). The methodological choice is to define, and thus annotate, those phenomena as "remarkable phenomena". The adoption of this theoretical-neutral label is functional for a non-aprioristic analysis of the outputs of language contact. In this context, the term "remarkable" refers to both phenomena that give evidence of non-standard linguistic forms and phenomena attributable to contact-induced variation. The "Corpus-contact" annotation scheme makes uses of the generic element <passage_remarquable> (Eng. 'remarkable passage') for signaling the occurrence of a remarkable phenomenon within the corpus. Every remarkable passage has an XML tag. In the relational database (cf. 2.3), several remarkable passages may be linked to a single remarkable phenomenon. The description of remarkable is data-driven and it is founded on three broad meta-categories. These are morphosyntactically remarkable phenomena (labelled as PREMS),

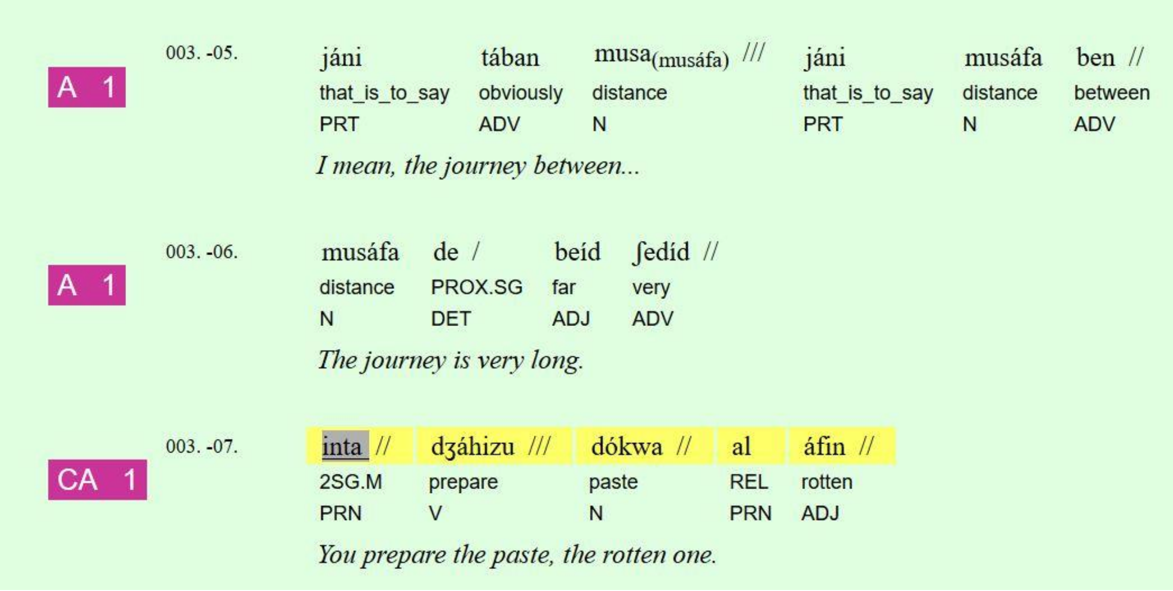
[7] The transcription tier also displays paraverbal events (incident, kinesic, vocal), linguistic indications (changes in pitch, tempo, loudness, tension, voice quality), pauses, overlaps and incomplete forms according to the TEI standards.

interactionally remarkable phenomena (labelled as PRINT), and discursively remarkable phenomena (labelled as PREDISC).

As far as PREMS are concerned, these include four subtypes of phenomena defined by their position in the chain of alternating languages (Vaillant 2015; Léglise and Alby 2016: 367). The PREMS subtypes are related to the presence of a segment of language B in language A, the sequence of two segments in languages A and B, the linking of languages A and language B, and the presence of a remarkable segment in language A (cf. 3.2).[8] PRINT, on their part, concerns the alternation of languages in interaction. In point of fact, language alternations often occur when the speech turn changes from one speaker to the next. For analyzing this kind of remarkable phenomenon, the “Corpus-Contact” schema adopts a sequentially based annotation. Each language is identified by an alphabetical label (A, B, C, ...) depending on the order in which it occurs in the corpus. The coding of PRINT is done at the speech turn level and it is automatically computed by a XSLT processor (cf. 2.3). Finally, PREDISC phenomena are related with the impact of language alternation on discourse cohesion and articulation as in the case of discourse connectors imported from one language to another.

2.3 Retrievability

After having transcribed, segmented and annotated your plurilingual corpus, a transform style sheet allows any browser compliant with XSLT to display the corpus as a sequence of aligned speech-turns as showed in Figure 2.



A 1 003. -05.

jáni	tában	musa(musáfa) ///	jáni	musáfa	ben //
that_is_to_say	obviously	distance	that_is_to_say	distance	between
PRT	ADV	N	PRT	N	ADV

I mean, the journey between...

A 1 003. -06.

musáfa	de /	beíd	ʃedíd //
distance	PROX.SG	far	very
N	DET	ADJ	ADV

The journey is very long.

CA 1 003. -07.

inta //	dʒáhizu ///	dókwa //	al	áfin //
2SG.M	prepare	paste	REL	rotten
PRN	V	N	PRN	ADJ

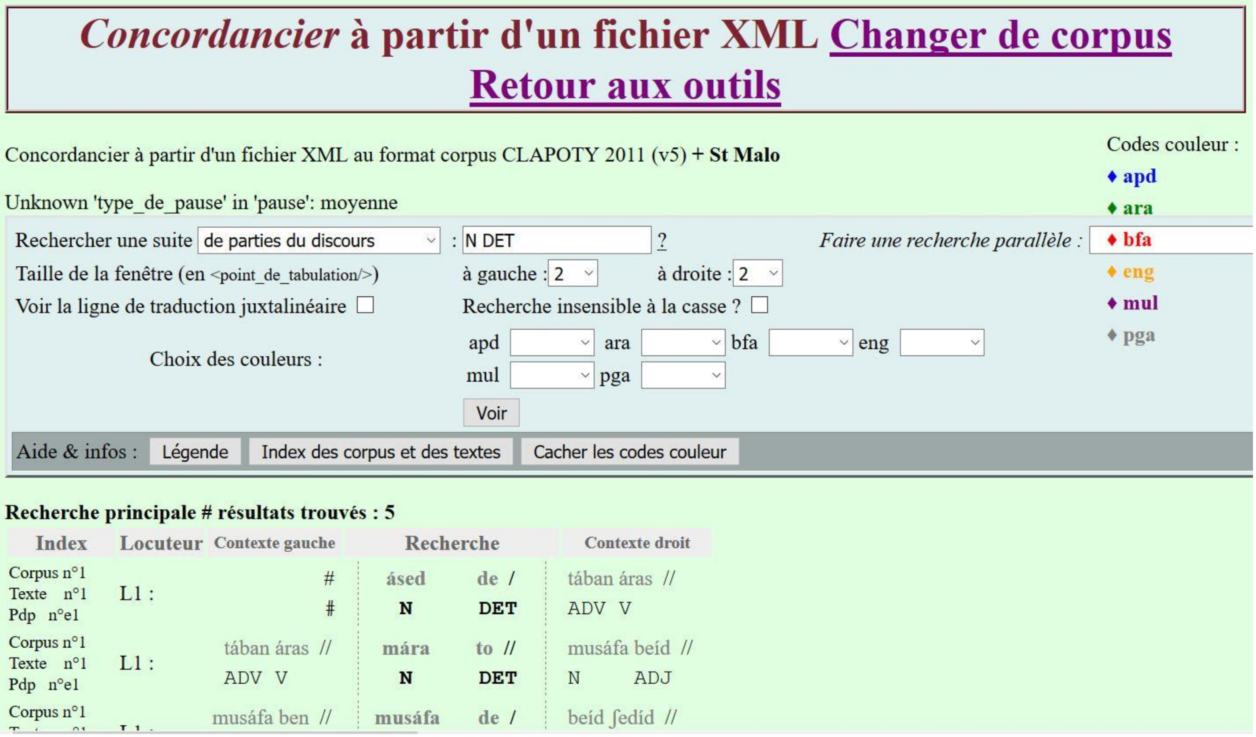
You prepare the paste, the rotten one.

Figure 2. The visualization of aligned speech turns

The finalized corpus is associated with a relational database storing metadata on speakers, languages and sociolinguistic settings. This allows analyzing remarkable phenomena in the light

[8] PREMS subtypes are also associated with a number of morphosyntactic subcategories: PREMS-GV (remarkable phenomenon in a Verb Phrase), PREMS-GN: in the Noun Phrase (remarkable phenomenon in a Noun Phrase), PREMS-GN-Det (remarkable phenomenon in a determined Noun Phrase), etc.

of the main sociolinguistic variables. Finally, forms and patterns can be retrieved by means of a PERL concordancer[9] allowing complex queries on the data, as showed by the following Figure.[10]



Concordancier à partir d'un fichier XML Changer de corpus Retour aux outils

Concordancier à partir d'un fichier XML au format corpus CLAPOTY 2011 (v5) + **St Malo**

Unknown 'type_de_pause' in 'pause': moyenne

Codes couleur :
- ♦ apd
- ♦ ara
- ♦ bfa
- ♦ eng
- ♦ mul
- ♦ pga

Rechercher une suite: de parties du discours : N DET ?

Faire une recherche parallèle :

Taille de la fenêtre (en <point_de_tabulation/>) à gauche : 2 à droite : 2

Voir la ligne de traduction juxtalinéaire □ Recherche insensible à la casse ? □

Choix des couleurs : apd ara bfa eng mul pga

Recherche principale # résultats trouvés : 5

Index	Locuteur	Contexte gauche	Recherche		Contexte droit
Corpus n°1 Texte n°1 Pdp n°e1	L1 :	# #	**ásed** **N**	**de /** **DET**	tában áras // ADV V
Corpus n°1 Texte n°1 Pdp n°e1	L1 :	tában áras // ADV V	**mára** **N**	**to //** **DET**	musáfa beíd // N ADJ
Corpus n°1	L1 :	musáfa ben //	**musáfa**	**de /**	beíd ʃedíd //

Figure 3. The PERL concordancer

3. Qualitative analysis of plurilingual corpora

In the following section, we will give more details about the “Corpus-Contact” annotation schema taking examples from two plurilingual corpora encompassing different varieties of Arabic.

The first one is a heterogeneous corpus of Juba Arabic, an Arabic-based pidgincreole spoken in South Sudan (Manfredi 2017). Before the South Sudan independence in 2011, Juba Arabic has been extensively exposed to contact with Sudanese Arabic, its lexifier and former dominant language. Thereafter, new contact situations between the pidgincreole and its lexifier emerged due to returnees entering South Sudan from Arabic dominant regions. This prolonged contact eventually led to different degrees of linguistic influence from the lexifier language (Manfredi 2017c). Moreover, the adoption of English as official language and main medium of formal education in South Sudan gradually leads to increasing heterogeneity in local linguistic practices. The corpus under analysis consists of recordings gathered among South Sudanese returnees who came back to Juba from Khartoum after 2011 and who provide evidence of a high degree of linguistic heterogeneity as compared to other speakers of the pidgincreole. The

[9] The concordancer has been developed by Anne Garcia-Fernandez using PERL (an Apache module) within the Labex-EFL project, strand 3, research operation *LC1 Multifactorial Analysis of Language Changes*. https://axe3.labex-efl.org/fr/LC1f (accessed 1 January 2015).

[10] In the top right corner of Figure 2, we can see the list of languages included in the corpus. These are identified by ISO-639 codes and visualized with different color codes. Multilingual segments are also identified by means of the special abbreviation *mul.*

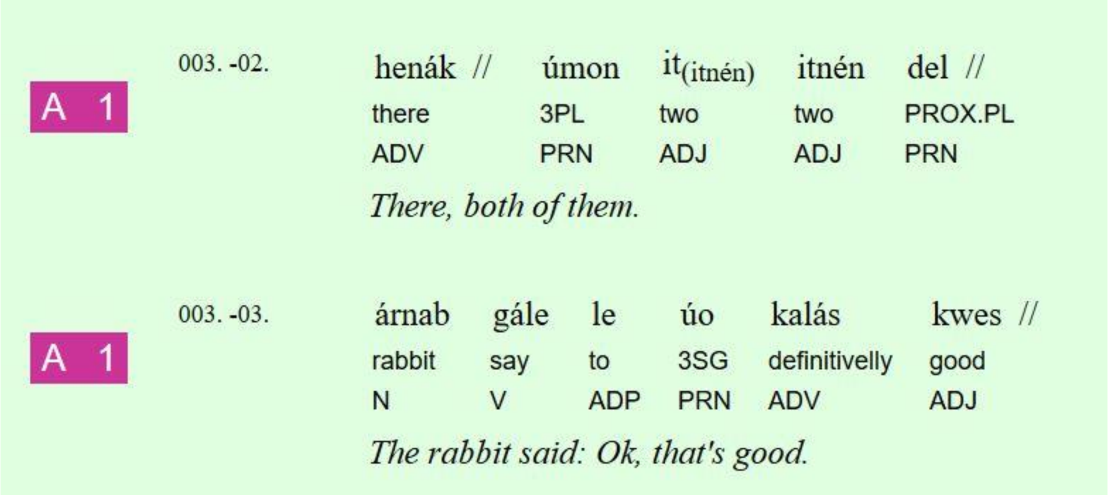
languages included in the corpus are Juba Arabic, Sudanese Arabic, Modern Standard Arabic, English and Bari (the main substrate language of Juba Arabic).

The second corpus is part of an ongoing research about heterogeneous language practices among transnational Arabic-Turkish speaking multilingual families from Antioch (Turkey). In the context of modern Turkey, Antiochian Arabic represents a minority language involved in an asymmetric contact situation resulting in a process of language shift toward Turkish, the official language of the country (Arnold 1998, 2000, 2006; Smith-Kocamahhul 2003). The corpus includes recording gathered among multilingual families that moved at different times from Antioch to Paris (France) and Berlin (Germany), where in addition to Arabic and Turkish, they speak French and German (Istanbullu 2017). The corpus is intended to provide an intergenerational analysis of ordinary language practices between members of the same family.

As we will see, both corpora are highly plurilingual and they entail different degrees of linguistic heterogeneity depending on a number of factors such as the duration and the stability of contact situations as well as the typological proximity/distance between languages in contact.

3.1 Monolingual, multilingual and floating segments

Despite the plurilingual nature of the two corpora, it obviously possible to find instances of monolingual speech turns, as we can see in Figure 4.

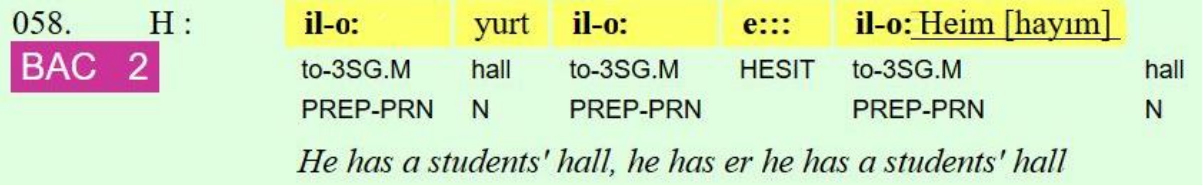


A 1	003. -02.	henák //	úmon	it(itnén)	itnén	del //	
		there	3PL	two	two	PROX.PL	
		ADV	PRN	ADJ	ADJ	PRN	
		There, both of them.					
A 1	003. -03.	árnab	gále	le	úo	kalás	kwes //
		rabbit	say	to	3SG	definitivelly	good
		N	V	ADP	PRN	ADV	ADJ
		The rabbit said: Ok, that's good.					

Figure 4. Two monolingual speech turns in Juba Arabic

The previous excerpts of corpus shows two monolingual speech turns in Juba Arabic. The alphabetical code A identifies the only language found in the speech turns (i.e. Juba Arabic, ISO-639 pga), whereas the numerical code 1 identifies the speaker. Being monolingual, the transcription tier is homogeneous and it does not contain boldface, italics or underlined segments. In this case, the annotator also chose to adopt a broad phonetic transcription including a prosodic segmentation signaled by backslashes (i.e. / minor prosodic boundary, // major prosodic boundary).

In contrast to the above, the two corpora under analysis are characterized by a high incidence of multilingual segments as showed by the following figure.

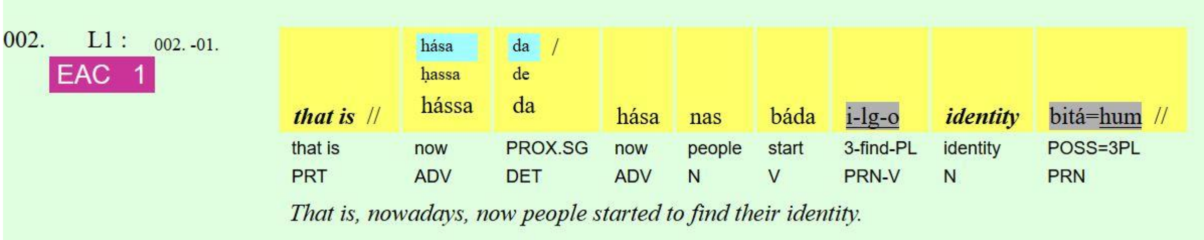


058. H :	**il-o:**	yurt	**il-o:**	e:::	**il-o:** Heim [hayım]	
BAC 2	to-3SG.M	hall	to-3SG.M	HESIT	to-3SG.M	hall
	PREP-PRN	N	PREP-PRN		PREP-PRN	N
	He has a students' hall, he has er he has a students' hall					

Figure 5. A syntagmatically Antiochian Arabic-Turkish-German multilingual speech turn

The previous excerpt of corpus shows of a syntagmatically multilingual speech turn including three different languages: Antiochian Arabic, Turkish and German. Different from monolingual speech turns, multilingual speech turns are highlighted in yellow. The three languages are respectively identified by the alphabetical labes B, A, and C, whose sequence depends on the linear order they occurs in the speech turn. Most importantly, there are no specific glosses for marking instances of codeswitching or codemixing as in the case of other oral corpora (Manfredi et al. 2015). As we can see, the speech turn begins with the Arabic possessive prepositional phrase *il-o* [ilo:] 'he has' (lit. 'to him') visualized in bond. The Turkish noun *yurt* 'hall', on its part, is visualized in an unmarked roman type, whereas its German equivalent *Heim* is underlined and coupled with a phonetic transcription [hayim] giving evidence of a non-standard realization.

As already remarked (cf. 2.1), it is not rare that a given item of plurilingual corpora may belong to several languages at once. Let us take into account the following example from the Juba Arabic corpus.



002. L1 : 002. -01.		hása ḥassa hássa	da / de da						
EAC 1	***that is*** //			hása	nas	báda	i-lg-o	***identity***	bitá=hum //
	that is	now	PROX.SG	now	people	start	3-find-PL	identity	POSS=3PL
	PRT	ADV	DET	ADV	N	V	PRN-V	N	PRN
	That is, nowadays, now people started to find their identity.								

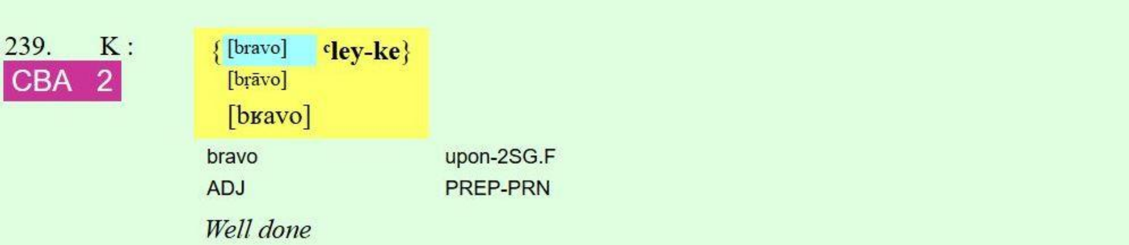
Figure 6. Floating annotation in a multilingual speech turn (Juba Arabic corpus)

Figure 6 gives evidence of a syntagmatically multilingual speech turn including English, Juba Arabic and Juba Arabic-Sudanese Arabic mixed forms (cf. 3.2). Furthermore, the speech turn includes the adverbial phrase *hássa da* 'right now' that can be related to both Juba Arabic and Sudanese Arabic. In more detail, the standard Juba Arabic form would be *hása de*, whereas the same adverbial phrase is commonly realized as *ḥassa da* in Sudanese Arabic. However, in this excerpt, the speaker geminates the voiceless alveolar sibilant /s/ alike in Sudanese Arabic, but he conflates the Arabic voiceless pharyngeal fricative /ḥ/ with a voiceless glottal fricative /h/ according to the Juba Arabic phonological rules. Besides, the Juba Arabic singular proximal demonstrative *de* is realized as [da], alike in Sudanese Arabic.[11] Of course, the question whether *da* is picked up from Sudanese Arabic or Juba Arabic does not arise for this bilingual speaker. As

[11] In the light of the common diachronic development **a > e* after dental and alveolar consonants, the Juba Arabic proximal singular demonstrative *de* 'this' most plausibly derives from the Sudanic Arabic proximal singular masculine demonstrative **da,* rather than from the proximal singular feminine demonstrative **di* (Manfredi 2017: 28).

for the annotator of the corpus, it is almost impossible to determinate whether *da* is the result of a structural integration from Sudanese Arabic due to ongoing decreolization or the output of an internal phonotactic assimilation of Juba Arabic /e/ with surrounding central vowels. In order to prevent a random decision that could erase the complexity of this particular contact situation, the segment is considered to be floating between the pidgincreole and its lexifier. Accordingly, *hássa da* is associated with two transcriptions (enlightened in light blue) reflecting the two interpretations available. When proposing alternative transcriptions, it is important to note the proximity of the segment in hand to both Sudanese Arabic form *ḥassa da* and the Juba Arabic *hása de* which is perhaps the creole form intended by the speaker.

In the same manner, the Antiochian Arabic-Turkish-French corpus encompasses items that may be found in all three languages, as in the case of *bravo* in the following figure.



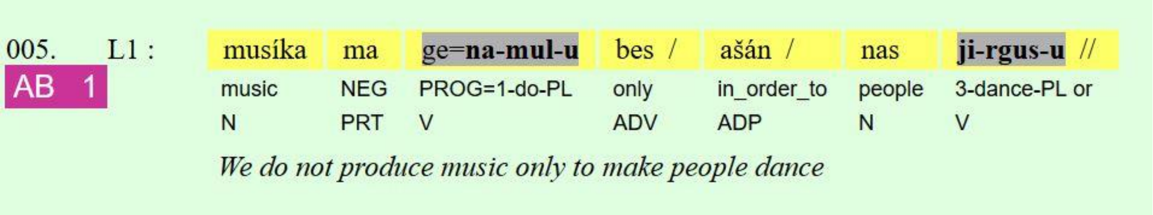
239. K :	{ [bravo]	ʿ**ley-ke**}
CBA 2	[bṛāvo] [bʁavo]	
	bravo	upon-2SG.F
	ADJ	PREP-PRN
	Well done	

Figure 7. Floating annotation in a multilingual speech turn (Antiochian Arabic corpus)

On the one hand, the realization [bravo] with a voiced alveolar trill /r/ may induce to think that we are dealing with a Turkish item followed by a prepositional phrase in Arabic. On the other hand, it is also plausible to think that the voiced alveolar trill [r] represents a non-standard realization of the French voiced uvular fricative /ʁ/. Lastly, it is not impossible that [r] is a depharyngealized variant of the Arabic /ṛ/. In absence of clearcutting phonological cues, the item is thus associated with three different transcriptions. This shows how the multiple annotation of floating segments encourage the annotator to consider different levels of analysis, beginning with the identification of a number explanatory factors, before going on to show their possible interaction in the production of a given phenomenon.

3.2 Remarkable contact phenomena

As far as the annotation of remarkable contact phenomena is concerned, this is intended to specify the layer of language processing and the type of syntagm affected by multilingualism (Vaillant 2015). The first level of analysis is that of morphosyntactic remarkable phenomena (PREMS), as exemplified by the following excerpt from the Juba Arabic corpus.

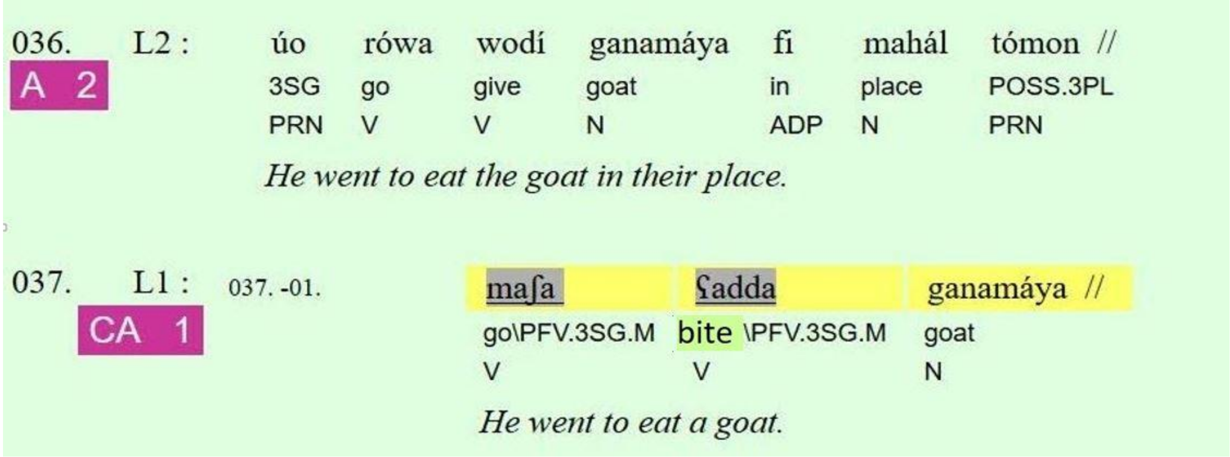


005. L1 :	musíka	ma	ge=**na-mul-u**	bes /	ašán /	nas	**ji-rgus-u** //
AB 1	music	NEG	PROG=1-do-PL	only	in_order_to	people	3-dance-PL or
	N	PRT	V	ADV	ADP	N	V
	We do not produce music only to make people dance						

Figure 8. PREMS: the integration of Sudanese Arabic pronominal affixes in Juba Arabic

One of the main typological features of Juba Arabic as compared to its lexifier is represented by the lack of personal indexes on the verb. Actually, Sudanese Arabic presents typical Arabic verbal paradigms with personal affixes indexing aspect, number and gender, whereas Juba Arabic encodes subjects by means independent pronouns preceding an invariable verbal form (Versteegh 1993; Manfredi 2017: 79). In contrast to that, in the previous excerpt of corpus we can note that the bilingual speaker makes use of personal indexes on the verb. In the case of *ge=na-mul-u,* the Juba Arabic progressive preverbal proclitic *ge=* (in roman type) precedes a verb marked for 1PL by virtue of the Sudanese Arabic prefix *na-* and suffix *-u* (in underlined roman type). This illustrates that the bilingual speaker tend to use Juba Arabic morphemes for expressing aspectual values while integrating personal affixes from Sudanese Arabic for encoding verbal agreement. However, in the case of *ji-rgus-u* the verb is marked for 3PL person and it includes the Sudanese Arabic prefix *ji-* and suffix *-u,* without any additional aspect marking. Both instances of morphological integration are highlighted in grey within a multilingual speech turn and they give evidence of very common phenomena in this particular creole-lexifier contact situation. In this case, we are dealing with two subtypes of morphosyntactic remarkable phenomena (cf. 2.2): the presence of Sudanese Arabic segments in Juba Arabic (exemplified by *ji-rgus-u*) and the sequence of segments from the two languages (exemplified by *ge=na-mul-u*).

The second level of analysis concerns interactional remarkable phenomena (PRINT) highlighting sequences where language switch occurs in correspondence of changes in speech turns (Auer 1995). In the following example, Juba Arabic is encoded by the alphabetical label A, whereas C encodes Sudanese Arabic. The first speech turn is associated with speaker 2, whereas the second one is uttered by speaker 1.



036.	L2 :	úo	rówa	wodí	ganamáya	fi	mahál	tómon //
A 2		3SG	go	give	goat	in	place	POSS.3PL
		PRN	V	V	N	ADP	N	PRN

He went to eat the goat in their place.

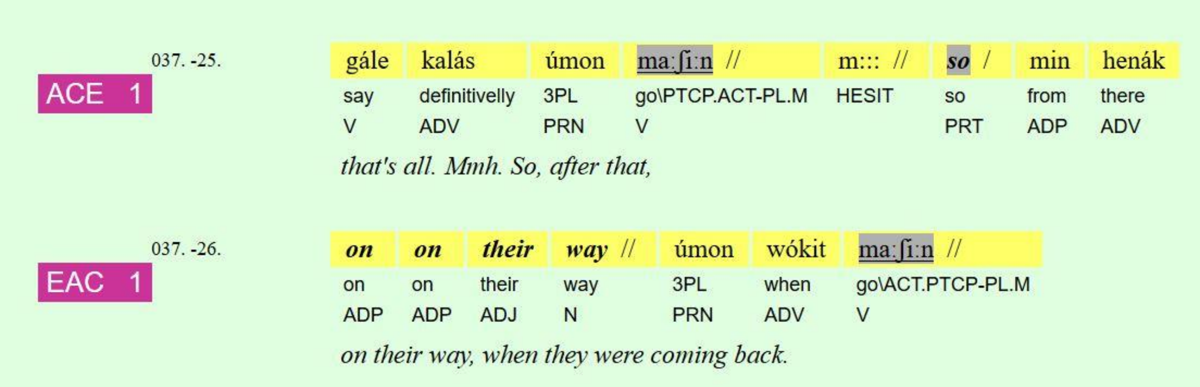
037.	L1 :	037. -01.	maʃa	ʕadda	ganamáya //
CA 1			go\PFV.3SG.M	bite \PFV.3SG.M	goat
			V	V	N

He went to eat a goat.

Figure 9. PRINT: Juba Arabic—Sudanese Arabic language change at speech turn taking

As we can see, the first speech turn is uttered in Juba Arabic (A). In contrast, the second speech turn begins with the Sudanese Arabic verbs *maʃa* ‘go’ and *ʿaḍḍa* ‘bite’, corresponding to the Juba Arabic verbs *rówa* and *ádi,* and ends with the Juba Arabic noun *ganamáya* ‘goat’. Accordingly, it is associated with the two alphabetical labels CA. The language switch at the beginning of the second multilingual speech turn is highlighted in grey. This allows analyzing all the PRINT sequences within one or more corpora in order to detect the structural organization of the interactions in diverse plurilingual contexts. In this regard, it is important to remark that the same strategy of annotation is possible when three or more varieties/languages are present in the same speech turn.

Finally, the third level of analysis concerns discursive remarkable phenomena (PREDISC) exemplified by the transfer of discourse markers illustrating the points at which language switch can occur. The following example shows the occurrence of the English discourse marker *so*, highlighted in grey, within a multilingual speech turn encompassing segments in Juba Arabic, Sudanese Arabic, and in English.



037. -25.

ACE 1

gále	kalás	úmon	maːʃiːn //	m::: //	***so*** /	min	henák
say	definitivelly	3PL	go\PTCP.ACT-PL.M	HESIT	so	from	there
V	ADV	PRN	V		PRT	ADP	ADV

that's all. Mmh. So, after that,

037. -26.

EAC 1

on	***on***	***their***	***way*** //	úmon	wókit	maːʃiːn //
on	on	their	way	3PL	when	go\ACT.PTCP-PL.M
ADP	ADP	ADJ	N	PRN	ADV	V

on their way, when they were coming back.

Figure 10. PREDISC: The occurrence of English discourse markers in Juba Arabic

The switch occurring at the level of discourse markers is a very common output of bilingual speech production. Therefore, the systematic annotation of this phenomenon facilitates a fine-graded analysis of the transfer of discourse markers and allows evaluating universalistic claims about their transferability (Matras 2008).

All things considered, by pointing out the multiplicity of relevant levels of analysis of language contact in multilingual settings, the multilayered annotation of remarkable phenomena, in combination with the information stored into the relational database, enables researchers to avoid simplistic explanations for complex linguistic dynamics and paves the road for multifactorial explanations of contact-induced language changes (Chamoreau and Léglise 2012).

4. Quantitative treatment of plurilingual corpora

Apart for the qualitative annotation and analysis of plurilingual corpora, the adoption of the JAXE-based schema of annotation gives the possibility to develop a qualitative analysis of language practices in multilingual settings. In this section, we will give some examples of possible statistical treatment of the Antiochian Arabic corpus.

Once that each segment of the corpus has been defined as monolingual or multilingual and thus assigned to one or more languages by means of *XML.lang* attribute (cf. 2.1), it is possible to extrapolate statistical figures by exporting data into Excel. This allows quantifying the incidence of monolingual and multilingual production both at the level of speech turns and at the individual level of the speakers. Let us look at the following charts.

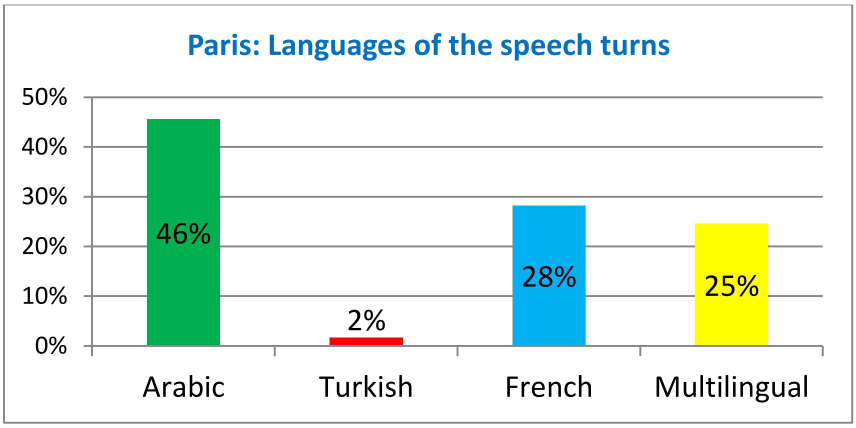


Chart 1. Monolingual and multilingual speech turns in the Paris corpus

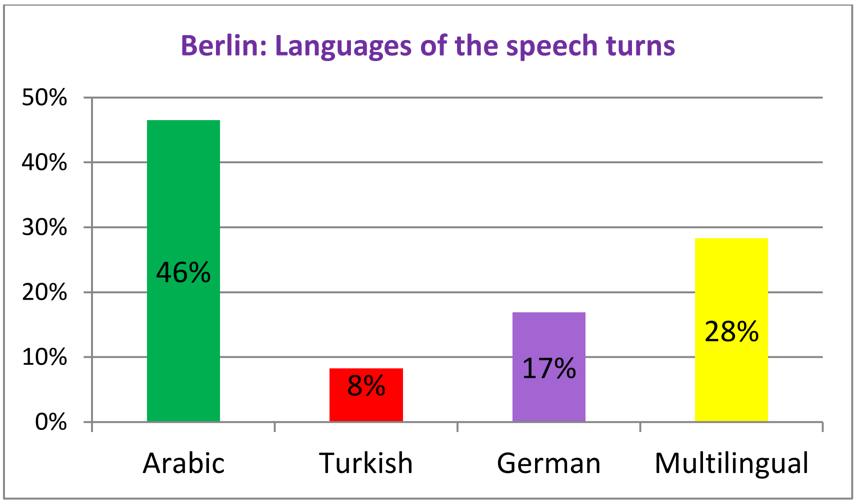


Chart 2. Monolingual and multilingual speech turns in the Berlin corpus

Charts 1 and 2 display information about the percentage of monolingual and multilingual speech turns across three generations of Antiochian migrants in Paris and in Berlin. These figures show general trends in linguistic practices. First of all, despite the fact that the use of Antiochian Arabic is regressing in Turkey because of the dominant position of Turkish, it remains the most commonly used language in ordinary interactions of diaspora communities. Secondly, we can note that the incidence of Turkish monolingual speech turns is much more important in Berlin than in Paris. This fact could be explained by the vehicular role played by Turkish for first-generation migrants in Germany (Dirim and Auer 2012). As a direct consequence of the limited use of Turkish in France, the national language of the host country is more commonly used in Paris than in Berlin. Against this background, multilingual speech production touches about one-fourth of the speech turns in both diasporic settings.

The previous figures may be further detailed giving evidence of individual language practices across generations as we can see in the next two charts.

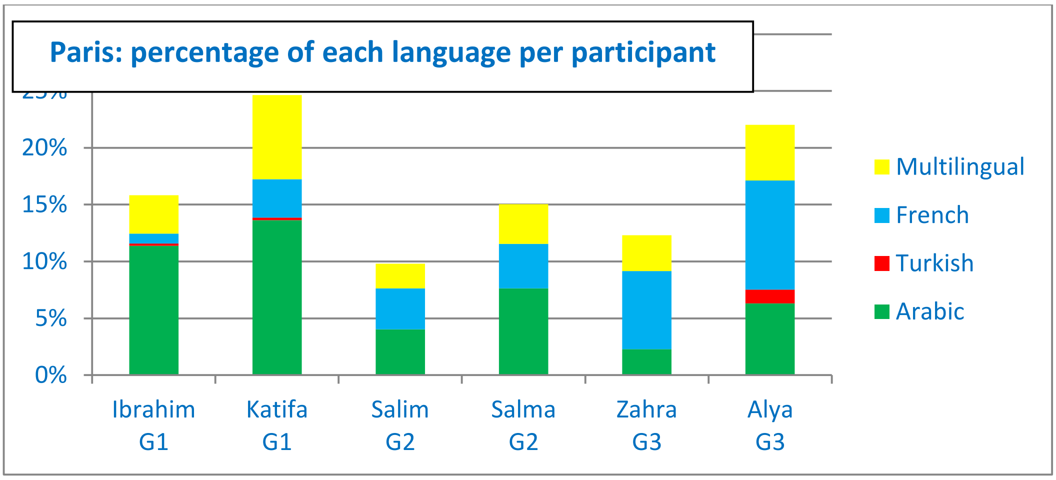


Chart 3. Individual language practices in Paris

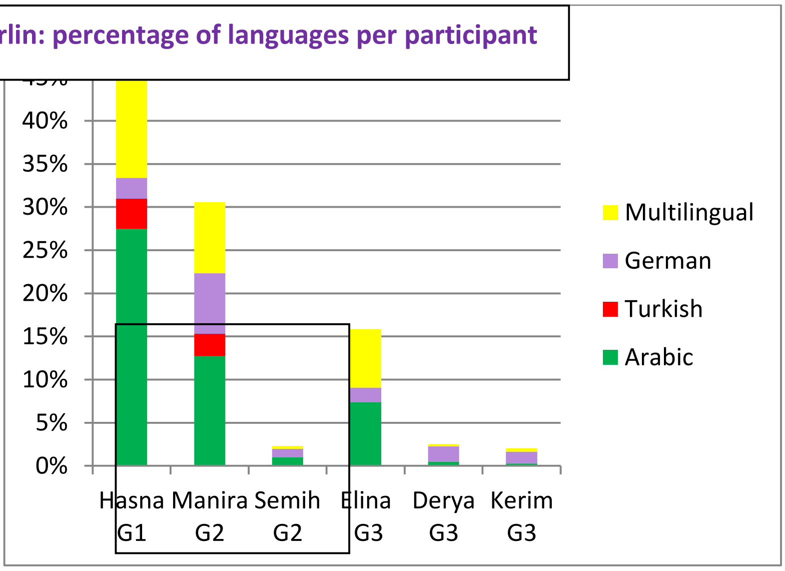


Chart 4. Individual language practices in Berlin

In comparing the previous figures, we can note that all the multilingual Antiochian migrants in Paris speak Arabic regardless of their different generational memberships (i.e. G1, G2, G3). This is quite surprising if we consider that migrants usually tend to shift towards the local dominant language within three generations (Heran et al. 2002). Looking at individual language practices in Berlin, both Arabic and Turkish are clearly losing ground to German as a consequence of lack of family support.

The quantitative treatment of plurilingual corpora by means of JAXE can enables other kinds of research. For instance, we could investigate the occurrence of a given language in relation with different parts of speech and/or semantic domains. In this regards, it has been noticed that Turkish is widely used in relation with toponyms, address terms and education-related terms, both in Paris and in Berlin (Istanbullu et Léglise 2014; Istanbullu 2017). The same holds true for ordinal numbers (Istanbullu 2017), a tendency that has already been observed among Arabic-Turkish bilingual speakers in Turkey (Procházka-Eisl and Procházka 2018) and that demonstrates an uninterrupted continuity of languages practices within transnational families. That being said, it should be stressed that this kind of quantitative queries may be applied to other plurilingual corpora in order to evaluate the functions performed by different languages in their social and geographical contexts.

5. Conclusions

In this chapter, we have briefly introduced a new computer-assisted method for annotating and analysing plurilingual corpora on the basis of two case studies involving different varieties of Arabic. However, the theoretical and methodological issues we have raised are not circumstantial but they are rather part of a broader reflection on the dynamics of multilingualism and on the language changes produced by them in the long run. The JAXE-based non-aprioristic method of annotation enables linguists to explore heterogeneous corpora regardless of whether these include different languages, dialects, registers or styles. In view of that, the adoption of this method of annotation can bring important advantages to the qualitative and quantitative study of Arabic in a wide range of sociolinguistic situations. For instance, it is not difficult to imagine its adoption for a detailed analysis of the highly variable forms of spoken “mixed Arabic” traditionally viewed as monolingual productions (Mejdell 2012). This may concern diglossic codeswitching/codemixing encompassing Modern Standard Arabic as well as different dynamics of dialect mixing or levelling. Furthermore, given that multilingualism is not limited to oral communication, we could envisage the adoption of this method for the study of written multilingual production (plurality of writing systems and encodings, multiplicity of genre-specific varieties, different levels of conformity to writing standards, etc. Vaillant 2015) in both traditional manuscripts and internet-mediated contexts. More generally, in view of the fact that Arabic is widely involved by external multilingualism in majority Arabic-speaking countries as well as in diaspora, this method can offer a fine-grained viewpoint on the multifactorial nature of language contact (Chamoreau and Léglise 2012) by combining different levels of analysis (phonological, morphosyntactic, interactional sociolinguistic, pragmatic, and typological). All in all, we hope we have provided a structured and informative introduction to a new corpus-driven method for the analysis of Arabic in multilingual settings.

References

Adamou, Evangelia. 2016. *A Corpus Driven Approach to Language Contact: Endangered Languages in a Comparative Perspective*. Berlin and Boston: De Gruyter Mouton.

Arnold, Werner. 1998. *Die arabischen Dialekte Antiochiens*. Wiesbaden: Harrassowitz Verlag.

———. 2000. ‘The Arabic Dialects in the Turkish Province of Hatay and the Aramaic Dialects in the Syrian Mountains of Qalamûn: Two Minority Languages Compared’. In Jonathan Owens (ed.), *Arabic as a Minority Language*. Berlin and New York: Mouton de Gruyter, 347–370.

———. 2006. ‘Antiochia Arabic’. In Kees Versteegh, Mushira Eid, Alaa Elgibali, Manfred Woidich and Andrzej Zaborski (eds), *Encyclopedia of Arabic Language and Linguistics*. *Vol. I A–Ed*. Leiden and Boston: Brill, 111–119.

Auer, Peter. 1995. ‘The Pragmatics of Code-Switching: A Sequential Approach’. In Lesley Milroy and Pieter Muysken (eds.), *One Speaker, Two Languages. Cross-Disciplinary Perspectives on Code-Switching.* Cambridge: Cambridge University Press, 115–135.

———. 1999. 'From Code-Switching via Language Mixing to Fused Lects: Toward a Dynamic Typology of Bilingual Speech'. *International Journal of Bilingualism* 3 (4): 309–332.

Avram, Andrei. 2014. 'Immigrant Workers and Language Formation: Gulf Pidgin Arabic'. *Lengua y Migración* 6 (2): 7–40.

Bale, Jeffrey. 2010. 'Arabic as a Heritage Language in the United States'. *International Multilingual Research Journal* 4 (2): 125–151.

Barontini, Alexandrine. 2013. *Locuteurs de l'arabe maghrébin-langue de France : Une analyse sociolinguistique des représentations, des pratiques langagières et du processus de transmission*. Doctoral dissertation. Paris: INALCO.

Bizri, Fida. 2010. *Pidgin Madame: une grammaire de la servitude.* Paris: Paul Geuthner.

Boussofara-Omar, Naima. 2006. 'Diglossia'. In Kees Versteegh, Mushira Eid, Alaa Elgibali, Manfred Woidich and Andrzej Zaborski (eds), *Encyclopedia of Arabic Language and Linguistics*. *Vol. I A–Ed*. Leiden and Boston: Brill, 629–637.

Caubet, Dominique. 2014. 'Moroccan Arabic Corpus'. In Amina Mettouchi, Martine Vanhove and Dominique Caubet (eds), *ANR, CorpAfroAs: A Corpus for Spoken Afro-Asiatic Languages*. https://corpafroas.tge-adonis.fr/ (accessed 5 February 2016).

Chamoreau, Claudine, and Isabelle Léglise. 2012. 'A Multi-Model Approach to Contact-Induced Language Change'. In Claudine Chamoreau and Isabelle Léglise (eds), *Dynamics of Contact-Induced Language Change*. Berlin and Boston: De Gruyter Mouton, 1–15.

Clyne, Michael. 2003. *Dynamics of Language Contact*. Cambridge: Cambridge University Press.

Dirim, Inci, and Peter Auer. 2012. *Türkisch sprechen nicht nur die Türken*. *Über die Unschärfebeziehung zwischen Sprache und Ethnie in Deutschland*. Berlin and New York: Walter de Gruyter.

Eirlys, Davies, Abdelali Benthalia, and Jonathan Owens. 2013. 'Codeswitching and Related Issues Involving Arabic'. In Jonathan Owens (ed.), *The Oxford Handbook of Arabic Linguistics.* Oxford: Oxford University Press, 326–348

Ferguson, Charles A. 1959. 'Diglossia'. *Word* 15: 325–340.

Heine, Bernd, and Tania Kuteva. 2005*. Language Contact and Grammatical Change*. Cambridge: Cambridge University Press.

Heran, François, Alexandra Filhon, and Christine Deprez. 2002. 'La dynamique des langues en France au fil du XXè siècle'. Population et Sociétés 376. https://www.ined.fr/fichier/s_rubrique/18724/pop_et_soc_francais_376.fr.pdf (accessed 1 January 2016).

Istanbullu, Suat. 2017. *Pratiques langagières intergénérationnelles : le cas de familles transnationales plurilingues (Antioche, Île-de-France, Berlin)*. Doctoral dissertation. Paris: Institut National des Langues et Civilisations Orientales.

Istanbullu, Suat, and Isabelle Léglise. 2014. 'Transmission de langues minoritaires dans la migration : Le cas de communautés arabo-turcophones'. Final Report handed over to the DGLFLF (The General Delegation for the French Language and in languages of France). Paris: SeDyL https://www.vjf.cnrs.fr/sedyl/recherches.php?voirlong=0&type=projet&programme=DGLFLF_Arabo_Turc (accessed 12 February 2015).

Ledegen, Gudrun. 2012. 'Prédicats 'flottants' entre le créole acrolectal et le français à la Réunion : Exploration d'une zone ambigüe'. In Claudine Chamoreau and Laurence Goury (eds), *Changements linguistiques et langues en contact : Approches plurielles du domaine prédicatif.* Paris: CNRS Éditions, 251–270.

Léglise Isabelle. 2018. 'Pratiques langagières plurilingues et frontières de langues', in Michelle Auzanneau, M. & Luca Greco (eds.), *Dessiner les frontières*, Lyon : ENS Editions, 143-169.Léglise, Isabelle, and Sophie Alby. 2016. 'Plurilingual Corpora and Polylanguaging, Where Corpus Linguistics Meets Contact Linguistics'. *Sociolinguistic Studies* 10 (3): 357–381.

Léglise, Isabelle, and Claudine Chamoreau. 2013. 'Variation and Change in Contact Settings'. In Isabelle Léglise and Claudine Chamoreau, *The Interplay of Variation and Change in Contact Settings*. Amsterdam and Philadelphia: John Benjamins, 1–20.

Léglise, Isabelle, and Santiago Sánchez Moreano. 2017. 'From Varieties in Contact to the Selection of Linguistic Resources in Multilingual Settings'. In Reem Bassiouney (ed.), *Identity and Dialect Performance.* Edinburgh: Edinburgh University Press, 143–159.

Manfredi, Stefano. 2017a. *Arabi Juba : Un pidgin-créole du Soudan du Sud.* Leuven-la-Neuve: Peeters.

———. 2017b. 'Arabic as a Contact Language'. In Reem Bassiouney and Elabbas Benmamoun (eds), *The Routledge Handbook of Arabic Linguistics.* London and New York: Routledge, 407–420.

———. 2017c. 'The Construction of Linguistic Borders and the Rise of National Identity in South Sudan: Some Insights into Juba Arabic'. In Reem Bassiouney (ed.), *Identity and Dialect Performance: A Study of Communities and Dialects.* London and New York: Routledge, 138–163.

Manfredi, Stefano, and Mauro Tosco (eds). 2018. *Arabic in Contact.* Amsterdam and Philadelphia: John Benjamins.

Manfredi, Stefano, Marie-Claude Simeone-Senelle, and Mauro Tosco. 2015. 'Language Contact, Borrowing and Codeswitching'. In Amina Mettouchi, Martine Vanhove and Dominique Caubet (eds), *Corpus-Based Studies of Lesser-Described Languages: The CorpAfroAs Corpus of Spoken Afroasiatic Languages*. Amsterdam and Philadelphia: John Benjamins, 283–308.

Mejdell, Gunvor. 2012. 'Diglossia, Codeswitching, Style Variation, and Congruence: Notions for Analyzing Mixed Arabic'. *Al-ʿArabiyya* 44–45: 29–39.

Mettouchi, Amina, Martine Vanhove, and Dominique Caubet (eds). 2015. *Corpus-Based Studies of Lesser-Described Languages: The CorpAfroAs Corpus of Spoken AfroAsiatic Languages*. Amsterdam and Philadelphia: John Benjamins.

Muysken, Peter. 2000. *Bilingual Speech: A Typology of Code-Mixing*. Cambridge: Cambridge University Press.

Myers-Scotton, Carol. 1993. *Duelling Languages: Grammatical Structure in Codeswitching*. Oxford: Clarendon Press.

Nagy, Naomi, and Miriam Meyerhoff. 2015. 'Extending ELAN into Variationist Vociolinguistics'. *Linguistic Vanguard: A Multimodal Journal for the Language Sciences* 1 (1): 271–281.

Nicolai, Robert. 2007. 'Language Contact: A Blind Spot in "Things Linguistic"'. *Journal of Language Contact* 1: 11–21.

Nunez, Joseph Jean-François. 2015. *L'alternance entre créole afro-portugais de Casamance, français et wolof au Sénégal : Une contribution trilingue à l'étude du contact de langues*. Doctoral dissertation. Paris: INALCO.

Nunez, Joseph Jean-François, and Isabelle Léglise. 2017. 'Ce que les pratiques langagières plurilingues au Sénégal dissent à la linguistique de contact'. In Michelle Auzanneau, Margaret Bento and Malory Leclère (eds), *Espaces, mobilités et éducation plurilingues : Éclairages d'Afrique ou d'ailleurs.* Paris: Editions des archives contemporaines, 99–119.

Matras, Yaron. 1998. 'Utterance Modifiers and Universals of Grammatical Borrowing'. *Linguistics* 36 (2): 281–331.

Owens, Jonathan (ed.). 2000. *Arabic as a Minority Language*. Berlin and New York: Mouton de Gruyter.

Procházka-Eisl, Gisela, and Stephan Procházka. 2018. 'The Arabic Speaking Alawis of the Çukurova: The Transformation of a Linguistic into a Purely Religious Minority'. In Christiane Bulut (ed.), *Linguistic Minorities in Turkey and Turkic-Speaking Minorities of the Peripheries*. Wiesbaden: Harrassowitz Verlag, 309–328.

Ratcliffe, Robert R. 2005. 'Bukhara Arabic: A Metatypized Dialect of Arabic in Central Asia'. In Éva Ágnes Csató, Bo Isaksson and Carina Jahani (eds), *Linguistic Convergence and Areal Diffusion: Case Studies from Iranian, Semitic and Turkic.* London and New York: Routledge, 141–159.

Sánchez Moreano, Santiago. 2015. *Conséequences linguistiques et identitaires du contact linguistique et dialectal à Cali (Colombie) : Le cas des variations de l'ordre des constituants*. Doctoral dissertation. Paris: Paris Diderot.

Smith-Kocamahhul, Joan. 2003. *Language Choice, Codeswitching and Language Shift in Antakya, Turkey*. Doctoral dissertation. Christchurch: University of Canterbury.

Tosco, Mauro, and Stefano Manfredi. 2013. 'Pidgins and Creoles'. In Jonathan Owens (ed.), *The Oxford Handbook of Arabic Linguistics.* Oxford: Oxford University Press, 495–519.

Thomason, Sarah G., and Terrence Kaufman. 1988. *Language Contact, Creolization, and Genetic Linguistics*. Berkley: University of California Press.

Vaillant, Pascal. 2015. 'Annotation of Pluringual Corpora. Experience from the Clapoty Project'. Paper presented at the *International Research Days on Social Media*, Université de Rennes 2, 24 October 2015.

Vaillant, Pascal, and Isabelle Léglise. 2014. 'A la croisée des langues. Annotation et fouille de corpus multilangue'. *Revue des Nouvelles Technologies de l'Information* 2: 81–100.

Versteegh, Kees. 1993. 'Levelling in the Southern Sudan: From Arabic Creole to Arabic Dialects'. *International Journal of the Sociology of Language* 99: 65–97.

Weinreich, Uriel. 1953. *Languages in Contact: Findings and Problems.* The Hague, Paris and New York: Mouton Publishers.

Winford, Donald. 2003. *An Introduction to Contact Linguistics*. Oxford: Blackwell Publishing.

———. 2005. 'Contact-Induced Changes. Classification and Processes'. *Diachronica* 22 (2): 373–427.