Plurilingual corpora and polylanguaging, where corpus linguistics meets contact linguistics

Isabelle Léglise, Sophie Alby

▶ **To cite this version:**

Isabelle Léglise, Sophie Alby. Plurilingual corpora and polylanguaging, where corpus linguistics meets contact linguistics. Sociolinguistic Studies, 2016, 10 (3), pp.357-381. hal-01426409

HAL Id: hal-01426409
https://hal.science/hal-01426409

Submitted on 4 Jan 2017

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

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

Léglise, I., & Alby, S., 2006, "Plurilingual corpora and polylanguaging, where corpus linguistics meets contact linguistics", Sociolinguistic studies 10.3, 357-381.

Plurilingual corpora and polylanguaging, where corpus linguistics meets contact linguistics[1]

Isabelle Léglise (CNRS, UMR 8202 SeDyL, France) leglise@vjf.cnrs.fr

Sophie Alby (Université de Guyane, UMR 8202 SeDyL, France) sophie.alby@espe-guyane.fr

Abstract:

Language contact and multilingualism issues are addressed by so diverse research traditions that we consider corpora and data exchange as good ways to make these traditions discuss. In this paper, we describe a methodology for detailed analysis of heterogeneous corpora which can be used to take into account both synchronic phenomena (linguistic variation and instances of polylanguaging or codeswitching) and diachronic phenomena. We point out the epistemological questions that arise in the analysis of plurilingual data and discuss the choices made with respect to the current norms and standards followed in corpus linguistics with a view to providing multi-factorial explanations in the field of language contact and multilingualism. We rely here on the methodology developed in the research project CLAPOTY.

Keywords: plurilingual corpora, contact linguistics, polylanguaging, corpus linguistics, codeswitching

1-INTRODUCTION

The field of contact linguistics has been expanding rapidly over the past fifteen years, but is traditionally split up into several different lines of research. The diachronic tradition tends to focus on the linguistic consequences of contact through the study of *contact-induced language change*. The synchronic tradition describes the effects of multilingualism and the social meanings assigned by speakers to *codeswitching*. Few studies attempt to take account of advances in both these lines of research at once. Language contact and multilingualism issues are addressed by so diverse research traditions that we consider corpora and data exchange as good ways to make these traditions discuss. Contrary to, for example child language acquisition, there is in our field no widely available database that scholars around the world can access to test particular hypotheses nor a common methodology to annotate and tag corpora. Instead, every project has to first collect new data and devise more or less new methodology.

We would like to describe here a specific methodology for detailed analysis of corpora which can be used to take into account both synchronic phenomena (linguistic variation and instances of polylanguaging through codeswitching, codemixing or crossing) and diachronic phenomena (change over time). We will also point out the epistemological questions that arise in the analysis of data. To do this, we rely on the methodology developed in the CLAPOTY[2]

project. After briefly describing the field of contact linguistics and the perspective of corpus linguistics, we present our plurilingual corpus and discuss the choices made with respect to the current norms and standards followed in corpus linguistics. We then present our chosen methods of identification and analysis with a view to providing multifactorial explanations in the field of language contact.

2-THE FIELD OF LANGUAGE CONTACT

It has become essential for linguists to acknowledge the presence of multiple languages during fieldwork; societal multilingualism is the norm and monolingualism is the exception, a special case (Wurm, 1996). Consequently, communication ordinarily takes place not in monolingual linguistic communities perceived as homogeneous, but in a multilingual "contact zone." Developed as a branch of modern historical linguistics, contact linguistics (Goebl, Nelde, Starý, and Wölck, 1996) makes contact phenomena its focus of interest and considers that "all languages are mixed in a weak sense" (Thomason, 2003: 21) and emphasises the importance of social factors (Winford, 2003) in the results of contact. A great deal of research has been devoted to contact-induced language change, with an interest in the types of phenomena that can be observed depending on the typological characteristics of the languages in question (cf. Heine and Kuteva, 2005, Thomason, 2001b, Ross, 1999). By focusing on morphosyntactic or typological features, however, this research has neglected the social and contextual aspects of language contact. From a synchronic perspective, the study of codeswitching and bilingual speech has developed within a separate tradition, itself divided into two approaches, one focusing on grammar, the other on pragmatics. The first approach seeks to determine the linguistic structure of codeswitching (Poplack, 1980, Muysken, 1995 Backus, 2003). Various models have been proposed for correctly predicting the structure of mixed forms and the linguistic constraints imposed on them (see, for example, the model of the matrix language proposed by Myers-Scotton, 1993b). The second approach focuses on the social role and meaning of codeswitching (Auer 1995, 1999, Myers-Scotton, 1993a), its communicative function as well as its social function as an identity marker for differentiating among social groups. Contrary to the diachronic perspective, this research with a social focus is not usually interested in explaining the linguistic consequences of language contact, and concentrates instead on context and usage.

These two traditions do acknowledge the impact of social factors on language change and on specific contact phenomena: this is a major step forward, given that historical linguistics has long been content to study internal motivations for change (Thomason and Kaufman, 1988: 1). But gaining an understanding of contact phenomena through the detailed analysis of contact situations, either at the macrosocial or microsocial level, which Weinreich (1953) had already called for, is far from a reality. Our aim is to take into account synchronic and diachronic phenomena in all their complexity by linking together the social factors usually treated in the context of linguistic anthropology, pragmatics, and sociolinguistics/sociology of language, while at the same time getting a more precise knowledge of the linguistic factors traditionally treated by descriptive and typological linguistics. The CLAPOTY project seeks to analyse, report on, and explain contact phenomena by establishing a method of analysis that takes into account the knowledge derived from these different subdisciplines, combining it with powerful research software tools developed specifically for this programme. In this way, the researchers hope to create a multilevel and multifactorial explanatory framework for contact phenomena, drawing on typologically diverse languages and diverse sociolinguistic situations.

In this article, we present the approach adopted and the methods and procedures used to analyse the linguistic consequences of language contact on five levels of analysis: the morphosyntactic, interactional, sociolinguistic, pragmatic, and typological.

3-CORPUS LINGUISTICS AND MULTILINGUAL OR PLURINGUAL CORPORA

Over the past fifteen years, corpus linguistics has seen the development of research on multilingual corpora – that is to say, corpora which include texts in different languages, but where the expectation is that each individual text is monolingual. Wherever possible, corpus linguistics that draws on multilingual corpora makes use of comparable corpora, that is, where there is a comparable number and type (or genre) of texts in each language (see especially McEnery and Wilson, 2000, Déjean, Gaussier and Sadat, 2002); sometimes parallel corpora are used, that is to say texts and their translations, and sometimes multilingual parallel aligned corpora, that is, parallel corpora for which there are relations of equivalence of translation between the items that make up the texts.

To differentiate our materials from these multilingual corpora, here we use the term "plurilingual corpora" to designate corpora that illustrate not only instances of codeswitching and codemixing but also *languaging* through the use of various linguistic resources referred also as *translanguaging* (Garcia 2009) or *polylanguaging* (Jørgensen, Karrebæk, Madsen and Møller 2011) and that we call heterogeneous language practices (Léglise 2013a). These plurilingual corpora, unlike the multilingual ones just mentioned, are still few in number, not readily available to the linguists' community, and rarely "processed" by the available computer software. We might mention the LIPPS/LIDES[3] project, whose purpose was to develop standards for mixed-language and codeswitching transcription, or the BilingBank database accessible via the TalkBank site.[4]

Plurilingual corpora are particularly interesting because the problems of variation and non-standard forms, often neglected by the large corpora, or controlled for by general parameters (such as the type of text or speech collected), are central to them. They often illustrate not only what is generally seen as internal variation in languages – for example, morphosyntactic or lexical variations (which can sometimes be connected to stylistic or dialectal varieties), but also forms that are harder to classify. As we shall see in due course, corpora that include *transidiomatic practices* (Jacquemet 2005), *crossing* (Rampton 2005) or *trans-* or *poly-languaging* performed by plurilingual speakers with varied skills pose formidable problems, not only for the identification of forms but also of their transcription and annotation. Similarly, the very definition of plurilingual corpora and the choice of speech situations to document are questions more difficult to assess than those within the field of language documentation (Migge and Léglise, 2013) or for large standard corpora.

The CLAPOTY project uses a type of corpus linguistics which is sensitive to heterogeneous corpora and has developed tools to work on them. All these questions – of identification, notation, annotation – have been the subject of lengthy discussion among our project members. In what follows we present some of the solutions adopted, and wherever possible we describe the methodological and epistemological choices that we faced.

4-THE CLAPOTY CORPUS

4.1-Establishing a common corpus: the need for harmonisation

To make up for the lack of plurilingual corpora available in the existing literature and to create a shared basis for the research group's work, a common corpus was compiled. It consists of items of spontaneous speech, originally transcribed in traditionally separate subfields and for quite different research purposes, as shown in the three examples below.

(1) CLAPOTY_Léglise (Nengee – **variation of French** spoken as second language)

1-J Ken san i e suku e fuufeli a ini **maman chambre** anda?
Ken, what are you looking for? You are making a mess in mummy's room
2-M a na faansi i mu taki a djuka
you must not speak French but Ndyuka
Ken san i meki a sikoo tide?
Ken, what did you do at school today?
3-K **ce que je faire à l'école?** … tide mi meki **bonhomme** a sikoo anga **plus**
What did I do at school? Today I drew little guys and I did adding up
4-M pikin man i mu taki
Little men you should say

(2) CLAPOTY_Alby (French-**Kali'na**)

1-E aino' yemamɨ kapɨlɨ ɨwa man' yemamɨ molo man[5] ++ ayalanatoko loten[6]
Wait! I have to do my work. This is my work. You just have to talk.
2-D **kosi'**
Bother!
3-Y non' sérieux' **otɨ poko awu wekatuya' +** c'est pas que **akinupewa to'**[7]
No, seriously! Why do I run? I am not lazy!
4-E [**caf**] **man mei'**
You were with a girl?

(3) CLAPOTY_Chamoreau (purepecha-**Spanish**)

inte acha **mas** khéri-e-s-ti **ke** **de** xo anap yamintu
DEM homme plus grand-PRED-AOR-ass3 que de ici origine tout
This man is older than everyone else here (Lit. This man is older than of everyone)

Example (1) shows the switch between an English-based Creole, Nengee, and items in French. What the author (Léglise, 2007) was originally interested in was how to visualise the alternation between these different languages, hence the choice of different graphic markers (boldface to mark items in French and normal type to mark items in Nengee). Keeping track of who is speaking was also important, in order to identify who performed which type of alternation and what competence in speech comprehension and production each of them possessed. In the extract chosen, M always speaks in Nengee, unlike K and J, who incorporate elements of French into their utterances in Nengee. But in the last line, M reformulates "bonhomme" in her own language, showing she understands French. Another interesting element is the form of the French items recorded: in the first line, "maman chambre" follows the word order in Nengee, while in the fourth line, the use of the infinitive, "ce que je faire," shows the only partial competence of the young speaker, who is acquiring French and studying it at school.

The transcription of Example 2 was originally intended to provide the information necessary for an interactional analysis of the bilingual Kali'na / French speech produced by children and adolescents in a village in north-western French Guiana (Alby, 2001). Rising intonations,

marked by the apostrophe (kosi' on line 2), pauses and so on are transcribed. These different features enable us to understand more fully the function of codeswitching such as emphasis (Y on line 3), and also to describe the characteristics of the bilingual speech used by interlocutors when they are part of a peer group. For example, on line 3 there is no pause to be heard during the utterance "c'est pas que akinupe wa to," which may indicate that the transition from one language to another should not be seen as the sign of a lack of linguistic competence but as indicating a form of bilingual variation which corresponds to a socially identified group.

Example 3, on the other hand – which illustrates the insertion of a Spanish comparative structure, "mas que ... de," into an utterance in Purepecha, an Amerindian language spoken in Mexico – was originally transcribed to show the morphosyntactic composition of the utterance (Chamoreau, 2012); it presents the necessary interlinear gloss, but makes no mention of the speakers or the dialogue being engaged in. Here the researcher was interested in the grammatical description of phenomena with no consideration of their interactional aspects. It is the comparison of this "decontextualised" example with other similar examples, such as the utterances presented in Chamoreau (1995), which gives meaning to the choice of transcription method (morpheme by morpheme with each language identified).

As we can see, the characteristics initially shared by the various members who participated in the project were minimal: (1) they all worked with first-hand data transcribed either orthographically or in IPA; (2) they translated these data (at least in the form of free translation); (3) they sought to work on the heterogeneity of their corpora by analysing both variation and alternation. The CLAPOTY project thus required lengthy discussions among the group about the harmonisation of the transcriptions and annotations.[8]

4.2-A plurilingual and heterogeneous corpus

The recordings to be added to our common corpus were selected using data from different participants, depending on their intrinsic heterogeneity and also on the diversity of the languages in general (both geographically and typologically). We work mostly on understudied languages (or language varieties), some of them in contact with languages that possess a long tradition of grammatical study. The CLAPOTY corpus currently includes 40 languages, including Amerindians, Creoles, Romance, Balkan, East Asian, Germanic etc..

This deliberate typological and geographical diversity is supplemented by a deliberate diversity with respect to the sociolinguistic situations represented in the corpora. The corpora illustrate several kinds of contact: between dialectal varieties of the same language; between stylistic varieties of the same language (for example, elements we can identify as "youth stylistic practices" integrated into "ordinary" speech); between vernacular languages; between vernaculars and lingua francas; between languages referred to as "widespread" internationally; between languages seen as minority or majority, and so forth.

Lastly, the heterogeneous nature of the corpora is manifested in the diversity of types of interaction represented. Our common corpus currently includes 170 transcribed recordings. These interactions all include at least two varieties, sometimes as many as ten languages. The majority of the interactions involve at least three people, and can go up to ten or even thirty speakers. The 13 monologues can be broken down into four categories: the political report, the broadcast monologue, the folk-tale, and the narrative.

The dialogues may be subcategorised in terms of the relationships between the interlocutors and hence in terms of the symmetry or asymmetry of their respective roles. Among the 67 symmetrical interactions are those between adolescents in a friendship context for example.

Examples of the 77 asymmetrical interactions include interactions in a school setting, crossgenerational discussions, or interactions with administrative professionals. We looked for maximum heterogeneity both external (related to the types of text and interaction in the corpus) and internal (we specifically chose recordings that showed morphosyntactic variation and plurilingual items).

4.3-Annotation and encoding of the corpora

Our corpora are encoded in xml using the editor Jaxe[9] adapted to the annotation system we have developed. We refer the reader to Vaillant and Léglise (2014) for further technical details on the annotation system established. At this point, it is important to note that the schema document Corpus-Contact that we have created is based on the standards of the TEI (Text Encoding Initiative)[10], but adapted to our needs. We mention two major adaptations here.

The TEI breaks texts up into sentences. In our case, the smallest unit we want to use for breaking up our recordings is certainly not the sentence, a unit that has no meaning in oral language, but rather turn-taking, following the practice started by Sacks, Schegloff, and Jefferson (1974). We already identified these turns in Examples 1 and 2 above, either by the speaker's initial or by the sign "-" or else by a number so as to make it easier to cite a specific passage. Turn-taking and speech turns seemed to us to be the most clearly definable units in our transcriptions. Our choice is consistent with the choices adopted in other large-scale projects on spoken language, such as TalkBank and CHILDES via CLAN (MacWhinney 2000, 2007).

The TEI recommends identifying the basic language of each sentence, and noting in angle brackets when an item from another language is introduced, for instance as <foreign item belonging to language x>. After having initially tried to identify a basic language[11] for each speech turn, we gradually abandoned the systematic allocation of one language each time. In most cases, in fact, we observed several languages in the same turn, produced by the same speaker; we decided to mark these speech turns as "multilingual" and to identify "segments" assignable to a particular language within each multilingual turn. For example, the following utterance starts in French and continues in Kali'na. Rather than selecting – often arbitrarily – one language as the matrix language (Myers-Scotton, 1993b), we propose to consider that the utterance is multilingual (we represent this visually here by a frame) and composed of several segments, in this case in two different languages, which we distinguish by boldface for French and normal for Kali'na.

(4) football match (corpus CLAPOTY_Alby)
ce n'est pas que akinupe wa to'
But I'm not lazy !

Moreover, some items may possibly belong to several languages at once: in corpora exhibiting various forms of language alternation recorded on the border between French Guiana and Suriname, the adjective [diʀɛkt] can be viewed as French, English, or Dutch, or as a borrowing from any one of these three languages, and as we were annotating them, we frequently found it difficult to decide definitely which one it was. Thus, rather than make a definite decision, we decided to label these items as themselves "multilingual" and to identify all the available options that might apply. We also sought to adopt a transcription system that would show all the different possibilities visually. For cases where two languages share a number of features, such as a creole and its lexifier, Ledegen (2012) proposed a double

transcription, referred to as “floating”, so as to see the two possible interpretations available for the analyst. We have extended this idea to all the cases where more than one transcription and more than one language was possible, even when the languages are not so “close”: for example, in the following extract, a doctor tries to say a few words in the language of his patient, uttering line 11, “a go bon” (a form we view as non-native and maybe an instance of crossing), which may correspond to the standard form “a e go bung” in Nengee (where the pronunciation of the final nasal “ung” would be a little less open than the “on” in French). We feel it is important, when making the alternative transcription, to note the proximity of what was said both to the adjective “bon” which seems to have been selected and to the adjective “bung” which is perhaps the standard form intended by the speaker.

(5) kosokoso (corpus CLAPOTY_Léglise)

10-Inf1 tu parles qu’elle va mieux qu’hier !
You bet she’s doing better than yesterday!

11-Doc a go *bon?*
bung
go well
Doing well?

12-F1 **a e mama / mama fu mi e go?**
my mother / is my mother all right?

13-Doc mama ça *go bon?*
go bung
mother it’s doing well?

In this example, the alternative transcripts sometimes involve more than two languages, as is the case in line 13 which includes items attributable to French (*ça* and *bon*), items of ‘non-native’ Nengee (*mama* and *bung*), and one item, *go*, which can be classified as either Nengee or English. Using this method has made it possible to show that in some corpora almost all the turns can be attributed to one or the other language, as in Example 6.

(6) Discussion between men in a bar in Saint-Laurent (corpus CLAPOTY_Migge)

2-B a fu den man dati ya
a fu den man
FOC for the man DEM yes
It’s because of these men, yes

3-C i wani go na dape a didon
i wani go **a didon**
2 want go at there 3 lie.down
Do you want to go where he is lying?

Our deliberate choice to present all the possibilities in the transcriptions transforms the way we look at the corpora. Rather than viewing extract 6 as spoken in Nengee with some inserted items of Sranan Tongo (Migge and Léglise, 2011), we can conclude that the speakers prefer to use items common to both languages to express themselves (on language boundary and the use of common items, see Léglise, to appear). But,at times, they select a particular feature from one or the other language from their linguistic repertoire.

5-METHODS FOR A MULTILEVEL DESCRIPTION AND ANALYSIS OF “REMARKABLE PHENOMENA”

One of the major challenges posed by this type of data is to develop tools to describe and explain the linguistic phenomena observed in plurilingual corpora. Those phenomena may look like variation (traditionally viewed as internal or linked to language contact) or look like more obvious contact phenomena such as codeswitching and codemixing or broadly trans- or

polylanguaging. The methodological choice we have made is to define (and annotate) all the phenomena we wanted to work on as "remarkable phenomena." We use "remarkable" in both senses of the word: either the phenomena observed are out of the ordinary (not found in ordinary language), in which case we start with a sense of distance from the expected or standard form (Léglise, 2013b) and treat the observed form as "remarkable," deserving of particular interest, or else the phenomena observed seem to be exemplary or typical of phenomena already known and fully described in the literature on language contact, in which case we start with a sense of frequency and typicality. This choice involves minimal assumptions with respect to the terminology used, and thus avoids all the multiplicity of terms that bedevil the field of language contact, often used inconsistently by different authors and frames. Making this choice avoids involvement in endless terminological debates such as that on the distinction between borrowing and codeswitching for example, or between calque, interference, and transfer (Mackey, 1976, Zentella, 1997).

The position we have adopted is to use the most "neutral" terms possible, by sorting interesting phenomena into three meta-categories based on their behaviour or characteristics: morphosyntactically remarkable phenomena (PREMS), interactionally remarkable phenomena (PRINT), and discursively remarkable phenomena (PREDISC). Once the remarkable phenomena have been annotated (they appear highlighted in grey as in (7)), they are first described using a multilevel approach, then analysed (see below).

(7) Municipal council (corpus CLAPOTY_Lescure _Alby)[12]

ot~~i~~	nok~~i~~	molo	**CCOG**	**compt-~~i~~-l~~i~~**	**rappeler** poko	s-~~i~~-yan o'wainen
euh	euh	DEM	CCOG	account-GEN	remind busy.with	1-put-PRS 2.to

I remind you about the account of the CCOG

5.1-PREMS

The first level on which to understand remarkable phenomena is that of the linguistic material produced, at the level of speech production and the sequence of its usual or unusual morphosyntactic features. To describe these remarkable items, we propose to use a notation system based on our categories for annotating languages. The first descriptor concerns the place where the remarkable phenomenon occurs in the speech production: below this the remarkable phenomenon is noted in brackets [], and may concern the consecutive occurrence of an item of a language A and an item of a language B, or the unusual form of an item from language A, or the insertion of an item from language A into a language B, and so on. We present an example of each type below.

a) the sequence language A + language B is remarkable.

(8) President Cardenas in Tanaco (corpus CLAPOTY_Chamoreau)

para	ampe=i	wé-ka-sïn-i	t'u	ima-ni	ú-ra-ni	
for	than=2	to.want-FT-HAB-INT	2IND	DEM-OBJ	to.do-CAUS-INF	

why do you want to do that?

[<para><ampe=ri ...>][13]

b) the linking of A and B is remarkable.

(9) ABC (corpus CLAPOTY_Lescure_Alby)

ot~~i~~	**réserve**	molo	**la**	**Basse Mana**
euh	reserve	DEM	DEF	Basse Mana

uh the reserve of the Low Mana

[<molo><la] basse mana>14

c) the presence of segment B inside segment A is remarkable.

(10) kosokoso (corpus CLAPOTY_Léglise)

efu yu wani sabi ala sani fa **la famille** da mi mu sabi fi yu seefi
if you want to know everything about the family, then I must also be able to know things about yours
<sabi ala sani fa[<la famille>]da mi...>15

d) the remarkable item is positioned inside segment A.

(11) je suis pas ton *blada* (corpus CLAPOTY_Léglise)

013.K: oh mais c'est le **ga** qui tire ça dans ma main **bay**
gars
oh but it's the guy who takes that from me, give it!
<tire ça dans ma main>
<take that in my hand>[16]

5.2-PRINT

Plurilingual corpora are also crammed with interesting phenomena at the interactional level, the level of the language choices made by the interlocutors. These choices and alternations can occur in individual utterances (within the same turn), in which case we treat them as PREMS, but they happen most often when taking turns –from one speaker to the next. We thus came to realise that here we have to apply the description "remarkable" to entire conversational sequences. To do this, we have opted for a sequentially based approach based on Auer (1995), in order to encode languages and interlocutors and highlight the interactional sequences in which changes in language occur. Specifically, the coding is done at the speech turn level. Each language is identified by a letter depending on the order in which it occurs in the corpus, and each speaker is identified by a number on the same principle, as can be seen in extract (12). Subsequently we focus on either the form of the entire interaction or that of the sequence of some exchanges. Thus, in the following example, the first language (French) is coded A and the first speaker (J.) is coded 1; the second language in order of appearance (Kali'na) is coded B and the second speaker in order of appearance (S.) is coded 2.[17]

(12) je lui parle en français? I speak to him/her in French? (corpus CLAPOTY_Alby)

23-S	une panier?	**A1**
	a basket?	
24-J	oui	**A2**
	yes	
25-S	(she sings) // j'ai fini / ça y est!	**A1**
	I've finished / it's done!	
26-J	**eneko te senepoya owa**	**B2**
	look, I show it to you	
27-S	**uwa**	**B1**
	no	
28-J	c'est bon?	**A2**
	is it okay?	

This exchange can be described as follows: A1 A2 A1 speaker 1 and speaker 2 both speak language or variety A down to line 26, where speaker 1 switches to language B and is followed by speaker 2 on the next line. Then, starting with line 28, the two speakers go back to using language A. Such an approach also enables us to deal with cases where several languages are used in the same utterance, as in example (13) where we see items from variety

B (French) inserted in a speech which seems to be organised with the morphosyntactic characteristics of variety A. The insert is marked by the use of brackets.

(13) informal conversation between ABC (corpus CLAPOTY_Lescure_Alby)

15-A	molo ot͟i͟ **nature garde**	**A[B]1**
	the gamekeeper	
16-A	ot͟i͟ **réserve** molo **la Basse Mana**	**A[B]1**
	uh the reserve of the Low Mana	
17-A	asito am͟i͟ man ne telapa moko kali'na / ot͟i͟ ͟inewala katako?	**A1**
	it's already a bit the Kali'na, uh how do you say that?	
18-A	moko kali'na ot͟i͟ **terrain de chass͟i͟l͟i͟** kanaiyan sipoli pamen.	**A[B]1**
	the hunting ground of the Kali'na like the white man says	

Thus A [B] 1 should be read as: "speaker 1 speaks in variety A (Kali'na, shown in normal type) while inserting elements from variety B (French, shown in boldface)." We take the same approach when three or more languages are present in the same corpus or in the same speech turn. Our corpora are particularly interesting because they mostly involve more than two speakers and more than two languages, a situation which has not been the subject of typological analysis in the literature thus far. In fact, the models proposed in the past are all based on two languages and two speakers: for example, the sequential analysis presented by Auer (1995) is based on the presence of two languages (A and B), and the typology of verbal interactions proposed by De Pietro (1988) is a bilingual, not a plurilingual, model.

The appeal of the mode of description we propose is that, like the PREMS, it requires a minimum of assumptions. It is only when we go back to review the whole of the corpora that we can draw on all the sequences identified in order to detect the structural organisation of the interactions in diverse plurilingual contexts which characterise our corpora. In the course of our analysis we can then try to ascertain how far the typologies proposed in the existing literature are validated or invalidated by our data. We are currently trying to solve one technical problem: how to annotate several lines of a corpus simultaneously. It is essential that the interactional characteristics be annotated in such a way that sequences with similar structures can be compared and easily connected to the metadata presented below. Achieving this would be a real breakthrough in the field, given that the models proposed thus far are based on "manual" methods or on content analysis (Alfonzetti, 1998, Alvarez-Caccamo, 1990).

5.3-PREDISC

There is an abundant literature in the field of language contact on the phenomena related to discourse markers, such as particles, connectives, fillers, and the like, which illustrate the points at which codeswitching occurs: they are often in one language while the rest of the utterance is in another (Matras, 1998). The following example illustrates this phenomenon with the use of "bon" and "quoi" in French in a bilingual Kali'na-French conversation.

(14a) informal conversation between ABC (corpus CLAPOTY_Lescure_Alby)

ami	**côté** molo **bon** palanaki͟l͟i ami-kon tameli tanepo man kali'na wa	
	kote	
from a certain viewpoint, the white man has shown his way of life to the Kali'na		

(14b) informal conversation between ABC (corpus CLAPOTY_Lescure_Alby)

wewe epeli ke soso molokon soso oti **frais quoi** soso oti		
the fruit of the trees with things like that, always fresh things, right! Always things.		

From this point of view, our corpora, many of which illustrate these phenomena, are not "remarkable" in the sense of being surprising – they illustrate expected phenomena and are thus remarkable especially because they are so typical. We have also developed a systematic annotation of all PREDISC so as to evaluate the claims of the existing literature in light of the diversity of our corpora and the information available for each one.

5.4-A method for the analysis of remarkable phenomena

A multilevel method of analysis has been proposed to understand the phenomena of each of our categories, PREMS, PREDISC, and PRINT. Linguists who identify remarkable phenomena are invited to consider a whole group of levels of analysis, beginning with an analysis of the possible role of each explanatory factor, before going on to show their interaction in the observed linguistic result, by systematically adopting different levels of analysis:

a) Analysis specific to language A: ask whether other examples of the same phenomenon are already documented in language A and propose an analysis: for example, a variation of this type was observed in situation X (for instance a geographic variation)

b) Analysis in connection with a group of languages (commonly but not necessarily from the same language family): ask whether other examples have previously been documented in languages close to language A: for example, the Romance languages generally exhibit simplification of the personal pronoun paradigms (many examples documented in such and such a variety)

c) Analysis in connection with contact (related to the linguistic or typological characteristics of the languages in contact): ask whether the PREMS might be related to a characteristic of language B: for example, the order of the constituents observed does not match that of language A in which the utterance is made but that of language B which is also present in the contact setting

d) Analysis in connection with each of the languages in different contact settings: ask whether other examples of the same type of effect are documented from language A or language B, in contact with different languages C and D: for example, if the contact situation being studied is that of French and Creole, does contact between French and African languages produce the same type of phenomena as those observed in the situation being studied, and does Creole in contact with a different language (Dutch) produce the same type of phenomena as those observed in the situation being studied?

e) Analysis in connection with contact but independent of the characteristics of the languages in contact: find out whether the existing literature reports identical phenomena in situations featuring other languages: for example, a grammaticalisation pattern that is normal across languages, or a gradual process already identified showing that the creation of articles follows a classic pattern, from the numeral to the indefinite, then the indefinite to the definite (Heine and Kuteva, 2003)

f) Sociolinguistic analysis: describe the communication setting in terms of the interlocutors (age, social or professional status, relationships), ask whether the utterance produced exemplifies a specific stylistic variety, and so on

g) Pragmatic analysis: if the phenomenon includes a change of language, ask what function can be attributed to the change, what is the topic of the exchange, what type of sequence (explanation, for example) is concerned, and so on.

The point of this approach, step by step and constraining, is to expand the explanations of phenomena beyond the usual confines of the transfer of structures from language B to language A (see Léglise 2013b as an example). By proposing a multilevel analysis, we anticipate that (most of the time) many of these levels are relevant to the linguistic outcomes observed. Following such an approach can make the different levels visible and identify new opportunities for explanation. The next step is to show that these levels interact and how they do so. At that point we will be able to offer multifactorial explanations for the phenomena observed.

6-Detailed and numerous metadata for multifactorial analysis

We have enriched each of the transcribed corpora with a large number of metadata relating to the contact settings, the languages, and the speakers. These draw on linguistic factors and social factors identified by contact linguistics (especially Thomason, 2001b, Winford, 2003), complemented by information from the fields of language typology, language acquisition, linguistic anthropology, and sociolinguistics. With an ambitious vision of the possibilities offered by the annotation of our data with this secondary information, we have sought to make these metadata as rich as possible, so that we can then examine each of these criteria as a potentially relevant factor in the production of the remarkable phenomena observed. To do this, the metadata have been organised in a database which is available for each text or corpus. Here we present the five main categories of metadata that we provide.

First, we have sought to classify each of our corpora according to three major typologies used in contact linguistics. Following the criteria given by each of the following authors, we have tried to fit our corpora into the typologies in question.

- The typology proposed by Winford (2003) for contact settings distinguishes between marginal contact settings (travel, exploration, conquests, media, foreign language learning, and so on), situations where speakers have grown up in the same community but there is contact between a dominant group and a minority group (immigration, invasion, military conquest, changes in national borders, intergroup contacts via trade, marriage, and so on), and situations where bilingualism is more “egalitarian.” Depending on the degree of contact, Winford seeks to assess the effects on languages, which can range from loanwords to large-scale structural borrowings with effects on the typology of the languages.

- The typology of verbal interactions proposed by De Pietro (1988) identifies different types of interaction situations involving bilingual or monolingual speakers, native or non-native speakers, those who share the same language and those who do not. He proposes a monolingual-bilingual axis and an endolingual-exolingual axis, so defining four scenarios. The linguistic phenomena observed in the interactions are thus explained according to the verbal communication situation as defined in the typology.

- The typology proposed by Auer (1999) for bilingual speech distinguishes between three situations: codeswitching, “in cases where the juxtaposition of the two codes is recognised and interpreted as locally significant by the participants”; language mixing, “where it is the juxtaposition of the two languages itself that is significant for the participants, not locally (contextually) but from the very fact that this type of speech is used”; and lastly, fused lects, in cases corresponding to stabilised mixed varieties “where speakers are no longer aware that their speech is mixed” and being mixed is intrinsic to the language thus created (Alby and Migge, 2007: 52).

The purpose of classifying our corpora following these three typologies is to verify whether the linguistic effects and phenomena expected in some of these situations do indeed match

those we see in our corpora; it is in some sense to "test" these typologies and verify whether they provide an explanation of the phenomena observed. Because of space and thematic orientation of this paper, we cannot discuss here the test of these typologies, only to say that but we may be in a position to supplement these typologies and test others.

Second, when considering the different languages found in our recordings, it has seemed important to note the genetic or typological relationships these languages exhibit: are they related (in the same language family, mutually comprehensible, stylistic or dialectal varieties of the same language, and so on)? Can they be considered typologically "close" or remote, and by what criteria? The question of typological distance between languages is not a trivial one, and can be addressed in different ways, notably in terms either of objective distance – which some have attempted to measure but we note only locally in the context of linguistic subsystems or specific areas – or in terms of subjective or perceived distance (Kellerman and Sharwood Smith 1986, Giacalone Ramat 1994), this being particularly important in language acquisition settings, and which we also take into account in our analyses (see the linguistic ideology referred to below). Although this is a complex issue, it is an essential one in the field of contact linguistics. Thomason (2010: 40) stresses the importance of knowing how much typological distance there is between the particular subsystems (or domains) of the languages in contact, since this helps to predict the type of interference (what we would call remarkable phenomenon) that may occur; this is also a function of the intensity of the contact. When the typological distance is small, subsystems – in which one rarely observes contact-induced change – may be affected by contact. Thomason cites the case of inflectional morphology, which is usually little affected by contact. A "minimum" typological distance is responsible for the frequency of interdialectal interference, which involves inflectional features rarely transferred in the case of more distant languages. This, according to Thomason, is not a mere trigger effect but an important explanatory factor.

Here are some metadata that we propose specifically for genetic or typological relationships. For a corpus illustrating contacts between Kali'na (a Carib language) and French (a Romance language), we note that these languages are typologically distant when we look at the order of the constituents in a sentence, especially the Verb Phrase. For a corpus illustrating contacts between Pamaka (an English-based Creole) and Aluku (an English-based Creole), we consider that they are dialectal varieties of the same language (Nengee), while for contacts between Pamaka (an English-based Creole) and English (a Germanic language), we consider that at the genetic level (using "genetic" here in the broad sense), the contacts are between a creole language and its lexifier, but that at the typological level these languages' features are relatively distant, for example with respect to the expression of TMA markers. This kind of information seems important for verifying whether the observed consequences of language contact are related to genetic or typological relationships and similarities.

Third, the existing literature on contact stresses that the duration and stability of language contact is an important criterion affecting the outcome of contact (Thomason, 2001a, Winford, 2003). It is usually the social and sociolinguistic data included in analyses of contact, the relevant "social factors" or "contact scenarios," which have explanatory power. We have decided to define these elements for each pair or trio of languages in our corpus. We apply this annotation on two levels: at the level generally included in the literature, namely the "linguistic community," and at the level which seems equally relevant to explain the phenomena in question, namely the speakers and their families.

Fourth, our close relationship to fieldwork and knowledge of work in language acquisition and linguistic anthropology have led us to specify a number of secondary data which seem to play an important explanatory role in the results of contact, a role we would in any case like to test. Where and how languages have been acquired seem to us to be important data and we have sought to record them systematically: for instance, was this language transmitted at home when the speaker was a child, was it the primary language (or one of the languages) of socialisation, did the speaker learn the language at school or in a similar formal context, or was it acquired in informal contexts (among peers) or in the public sphere, or was this a case of language shift or disruption of intergenerational transmission?

Fifth, the status of different languages in the communication situation of the corpus is also a factor to be taken into account, and we also seek to test the role that the following factors can play. What are the functions performed by the different languages in the geographical region in question? What is their respective status (in principle and in practice)? How are they balanced quantitatively (as numerically majority or minority languages in the microsituation concerned, in the town where the recording is made, or in the region as a whole)? What are the ideological relationships between the languages? In the region as a whole, is language A ideologically a minority or devalued language? In the town or district, are language A and language B equally valued? In the microsituation, is language B viewed as situationally more appropriate or valued than language A, for example?

All these questions seem to us to be relevant, and we view them as secondary data worth noting and then examining in order to confirm or disconfirm their role in the observed language results, and, if they turn out to have a role, to explain these results more effectively.

7-CONCLUSION

On one hand, with a specific attention to heterogeneous language practices, we hope to offer a wider but more precise perspective on language contact and multilingualism. The excerpts presented here are not sporadic instances of codeswitching or mere variations maybe leading to change but rather part of a broader depiction of the use of various linguistic resources by plurilinguals. It also precisely shows how poly- or translanguaging occurs in everyday life among individuals and may sometimes lead to style creation or language change. For this, we not only look at the level of utterances through insertions and alternations but we propose a turn-taking approach to examine polylanguaging in interaction.

On the other hand, our corpora and methods enable us to work on heterogeneous data from everyday communication, whether these use resources which may be attributed to various languages, dialects or styles, or exhibit variations within what are usually viewed as monolingual productions. The method of corpus annotation that we have established reveals the presence of heterogeneity, because the step-by-step approach requires the researcher to raise questions that did not necessarily arise during the transcription; it also requires the transcriber to constantly widen the universe of possibilities by asking whether an alternative transcription is possible and if the item being annotated might belong to a language other than the one that first comes to mind.

Similarly, the method of analysis of remarkable phenomena and the inclusion of social data also require to examine the data with an open-ended set of possibilities in mind. Only this spirit of openness will be sure to produce multifactorial explanations and multilevel analyses. We are firmly committed to analysis – of both linguistic data and social data – at a “micro-micro” level, which requires painstaking recording and analysis. We believe that only thus can

we find regularities (especially statistical ones) and explanations for the phenomena we observe.

By combining methods and perspectives drawn from different traditions in linguistics, and from different levels (morphosyntactic, interactional, pragmatic, sociolinguistic, typological) we make use of two approaches, the inductive and the deductive. This methodology seeks to open up the range of possibilities for explanation by means of complex manual analysis of the phenomena and at the same time to test hypotheses via computerised verification using databases created from hundreds of manually entered annotations.

We believe corpora and data exchange are a good way to encourage a discussion between and cross-fertilise various research traditions such as linguistic typology, contact linguistics, pragmatics and sociolinguistics although they are all interested in multilingualism and the consequences of language contact. It is our hope our devised methodology and database will serve them all for testing hypotheses and sharing future corpora and experiences.

About the authors

Isabelle Léglise is Senior Researcher at the French National Centre for Scientific Research (CNRS) where she runs programmes on Multilingualism, Language Variation and Contact at the SeDyL (Structure et Dynamique des Langues) Research Lab she is heading. Her research projects in French Guiana, Suriname and Brazil address multilingualism and language policy related to migration, education and health. She published widely on multilingualism, language variation, discourse analysis and language policy in postcolonial settings. Her last publications include *Exploring Language in a Multilingual Context: Variation, Interaction and Ideology in language documentation* (2013, Cambridge University Press) with B. Migge, and the co-edition of books such as *The Interplay of Variation and Change in Contact Settings* (2013, John Benjamins) or *In and out of Suriname: language, mobility and identity* (2015, Brill).

Sophie Alby is Associate Professor in the University of French Guiana. She is member of the SeDyL (CNRS-IRD-INALCO). Her research interests are in synchronic language contact focusing on multilingual contexts, and language and education with respect to lesser-used languages in French Guiana. Her last publications include 'Kali'na NP's in contact. Variation or linguistic change?' To be published in *STUF,* 'Multilingualism as a resource for teaching and learning in French Guiana' (2016, with I. Léglise) In *The multilingual edge of education*. London: Palgrave Macmillan, 'Code-switching. Alternances et mélanges codiques.' (2013). *Sociolinguistique du contact. Dictionnaire des termes et concepts*. Lyon : ENS.

References

Alby, S. (2001) Contacts de langues en Guyane française : une description du parler bilingue kali'na-français, Thèse de doctorat Université Lumière Lyon II.

Alby, S. and Migge, B. (2007) Alternances codiques en Guyane française. Les cas du kali'na et du nenge, in I. Léglise, B. Migge (ed), *Pratiques et représentations linguistiques en Guyane: regards croisés*, 49--72. Paris: IRD Editions.

Alfonzetti, G. (1998) The conversational dimension in codeswitching between Italian and dialect in Sicily. In P. Auer (ed) *Codeswitching in conversation* 180--214. London: Routledge.

Alvarez Caccamo, C. (1990) Rethinking conversational code-switching: codes, speech varieties and contextualisation. *Proceedings of the sixteenth annual meeting of the Berkeley Linguistics Society*. Berkeley.
Auer, P. (1995) The pragmatics of code-switching: a sequential approach. In L. Milroy and P. Muysken (ed) *One speaker, two languages: cross disciplinary perspectives on code-switching* 115--135. Cambridge: Cambrige University Press.
Auer, P. (1999) From codeswitching via language mixing to fused lects: toward a dynamic typology of bilingual speech. *The International Journal of Bilingualism* 3-4: 309--332.
Backus, A. (2003) Units in codeswitching: evidence for multimorphemic elements in the lexicon. *Linguistics* 41(1):83--132.
Chamoreau, C. (1995) La comparaison en purepecha. Un exemple d'évolution syntaxique. *Faits de Langues* 5: 140--143.
Chamoreau, C. (2012) Constructions périphrastiques du passif en purepecha. Une explication multifactorielle du changement linguistique. In C. Chamoreau and L. Goury (ed) *Changement linguistique et langues en contact. Approches plurielles du domaine prédicatif* 251--270. Paris: CNRS Editions.
Déjean, H., Gaussier, E. and Sadat F. (2002) Bilingual terminology extraction: an approach based on a multilingual thesaurus applicable to comparable corpora. *Proceedings of 19th COLING* (International Conference on Computational Linguistics), Howard International House and Academia Sinica, Taipei, Taiwan (DOI 10.3115/1072228.1072394).
De Pietro, J-F. (1988) Vers une typologie des situations de contacts linguistiques. *Langage et Société* 43: 65--89.
García, O. (2009) Education, Multilingualism and Translanguaging. In Ajit Mohanty, Minati Panda, Robert Phillipson, and Tove Skutnabb-Kangas, *Multilingual Education For Social Justice: Globalising the Local*, 140--58. New Delhi: Orient Blackwan
Giacalone Ramat, A. (1994) Il ruolo della tipologia linguistica nell'acquisizione di lingue seconde. In A. Giacalone Ramat and M. Vedovelli (eds) *Italiano lingua seconda/lingua straniera* 27--43. Roma: Bulzoni.
Goebl, H., Nelde, P.H., Stary, Z. and Wölck W. (ed). (1996) *Contact linguistics. An international handbook of contemporary research*. Berlin: De Gruyter.
Heine, B. and Kuteva, T. (2003) On contact-induced grammaticalization. *Studies in Language* 27(3): 529--572.
Heine, B. and Kuteva, T. (2005) *Language contact and grammatical change*. Cambridge: Cambridge University Press.
Jacquemet, M. (2005) Transidiomatic Practices: Language and Power in the Age of Globalization. *Language and Communication*, 25 (3): 257--77.
Jørgensen, J. N., Karrebæk, M. S., Madsen, L. M., and Møller, J. S. (2011) Polylanguaging in Superdiversity. *Diversities* 13(2). www.unesco.org/shs/diversities/vol13/issue2/art2.
Kellerman, E. and Sharwood Smith, M. (eds). (1986) *Crosslinguistic influence in Second Language Acquisition.* New York: Pergamon Press.
Ledegen, G. (2012) Prédicats 'flottants' entre le créole acrolectal et le français à la Réunion : exploration d'une zone ambiguë. In C. Chamoreau and L. Goury (ed) *Changement linguistique et langues en contact. Approches plurielles du domaine prédicatif,* 251--270. Paris: CNRS Editions.
Léglise, I. (2007) Des langues des domaines, des régions. Pratiques, variations, attitudes linguistiques en Guyane, in I. Léglise, B. Migge (ed), *Pratiques et représentations linguistiques en Guyane : regards croisés*, 29--47. Paris: IRD Editions.

Léglise, I. (2013a) Multilinguisme, variation, contact. Des pratiques langagières sur le terrain à l'analyse de corpus hétérogènes. Paris: INALCO. https://tel.archives-ouvertes.fr/tel-00880500
Léglise, I. (2013b) The interplay of inherent tendencies and language contact on French object clitics: an example of variation in a French Guianese contact setting, in I. Léglise I., C. Chamoreau, (ed), *The interplay of variation and change in contact settings*, 137--163. Amsterdam: John Benjamins.
Léglise, I. (to appear) Pratiques langagières plurilingues et frontières de langues, in M. Auzanneau, L. Greco (ed), *Dessiner les frontières*, Lyon: ENS Editions.
Mackey, W.F. (1976) *Bilinguisme et contact des langues*. Paris: Klincksieck.
MacWhinney, B. (2000) *The Childes Project: tools for analyzing talk*. Mahwah, NJ: Lawrence Erlbaum Associates.
MacWhinney, B. (2007) The TalkBank Project. *Departement of Psychology*, Paper 174, http://repository.cmu.edu/psychology/174.
Matras, Y. (1998) Utterance modifiers and universals of grammatical borrowing. *Linguistics* 36(2): 281--331.
McEnery, A.M. and Wilson A. (ed). (2000) *Multilingual corpora in teaching and research*. Amsterdam: Rodopi.
Migge, B. and Léglise, I. (2011) On the emergence of new language varieties: The case of the Eastern Maroon Creole in French Guiana, in L. Hinrichs, J. Farquharson (ed), *Variation in the Caribbean*, 181--199. Amsterdam: John Benjamins.
Migge, B. and Léglise, I. (2013) *Exploring Language in a Multilingual Context: Variation, Interaction and Ideology in language documentation*. Cambridge: Cambridge University Press.
Myers-Scotton, C. (1993a) *Social motivations for code-switching: evidence from Africa*. Oxford: Clarendon Press.
Myers-Scotton, C. (1993b) *Duelling languages: grammatical structure in codeswitching*. Oxford: Clarendon Press.
Muysken, P. (1995) Code-switching and grammatical theory. In L. Milroy and P. Muysken (ed) *One speaker, two languages: cross-disciplinary perspectives on code-switching* 177--1998. Cambridge: Cambridge University Press.
Poplack, S. (1980) Sometimes I'll start a sentence in Spanish y termino en Espanol. *Linguistics* 18: 581--618.
Rampton, B. (2005) Crossing: Language and Ethnicity among Adolescents. 2nd ed. Manchester, UK and Northampton MA: St. Jerome Publishing
Ross, M. (1999) Exploring metatypy: how does contact-induced typological change come about? *Australian Linguistic Society's annual meeting*, Perth (http://rspas.anu.edu.au/linguistics/mdr/Metatypy.pdf).
Sacks, H., Schegloff, E. and Jefferson G. (1974) A simplest systematics for the organisation of turn-taking for conversation. *Language* 50(4): 696--735.
Thomason, S. (2001a) *Language contact: an introduction*. Edinbourg: Edinburg University Press.
Thomason, S. (2001b) Contact-induced typological change. In M. Haspelmath, E. Koenig, W. Oesterreicher and W. Raible (ed) *Language typology and language universals, Sprachtypologie und sprachliche universalien, vol.*2 1640--1648. Berlin/New York: Walter de Gruyter.

Thomason, S. (2003) Social factors and linguistic processes in the emergence of stable mixed languages, in *The Mixed language debate*, Yaron Matras & Peter Bakker (eds), Berlin, New York: Walter de Gruyter, 21-39.
Thomason, S. (2010) Contact Explanations in Linguistics. In R. Hickey (ed) *The Handbook of Language Contact* 31--47. Wiley-Blackwell.
Thomason, S. and Kaufman T. (1988) *Language contact, creolization, and genetic linguistics*. Oxford/Berkeley: University of California Press.
Vaillant, P. and Léglise, I. (2014) A la croisée des langues : Annotation et fouille de corpus plurilingues, *RNTI Revue des Nouvelles Technologies de l'Information*, Editions Hermann Paris, 81—100.
Weinreich, U. (1953) *Languages in contact: findings and problems*. New York: The Linguistic Circle of New York.
Winford, D. (2003) *An introduction to Contact Linguistics*. Oxford: Blackwell.
Wurm, S.A. (1996) *Atlas des langues en péril dans le monde*. Paris/Camberra: Editions UNESCO/Pacific Linguistics.
Zentella, A-C. (1997) *Growing up bilingual: Puerto Rican children in New York*. Oxford: Blackwell Publishers.

Appendix: Abbreviation used for the interlinear gloss:

1	first person
2	second person
3	third person
AOR	aorist
ASS	associative
CAUS	causative
DEF	definite
DEM	demonstrative
FOC	focus
FT	formative
GEN	genitive
HAB	habitual
IND	indicative
INF	infinitive
INT	interrogative
OBJ	object
PRED	predicative
PRS	present
PRTEN	enunciative particle

[1] This paper is a new version, translated, condensed and revised, from Léglise and Alby (2013), Les corpus plurilingues, entre linguistique de corpus et linguistique de contact : réflexions et méthodes issues du projet CLAPOTY, Faits de Langues n°41, 95-122. We would like to thank two anonymous reviewers for their valuable comments and suggestions.

[2] "Towards a multi-level, typological and computer-assisted analysis of contact-induced language change" is a project funded from 2009 to 2014 by the ANR under the number 09-JCJC-0121-01. Task 1, which is presented here and which we directed, involved the building of a common corpus and the creation of a model for analysing contact phenomena. The other

researchers actively contributing to Task 1 were E. Adamou (CNRS, Lacito), C. Chamoreau (CNRS, SeDyL), G. Ledegen (Rennes 2), B. Migge (UCDublin), C. Saillard (Paris Diderot, LLF), D. Troiani (CNRS, SeDyL), and P. Vaillant (Paris Nord, Lim&Bio).

[3] Cf. http://www.ling.lancs.ac.uk/staff/ruthanna/lipps/lipps.htm.

[4] Cf. http://talkbank.org

[5] E. explains that he has to record their conversations.

[6] He tells them to talk, to say anything, not to pay attention to the tape-recorder.

[7] Y. talks at random, just to test the recording before beginning the conversation.

[8] By "annotation" here we mean any sort of supplementary material added to the transcription by the linguists, especially the annotations that indicate the language spoken, morphosyntactic gloss, identification of parts of speech, etc. We will point out later where metadata of various sorts (typological, sociolinguistic, etc.) are concerned.

[9] http://jaxe.sourceforge.net/fr/ created by D. Guillaume, S. Ayadi, B. Tasche, O. Kykal, C. Dedieu, L. Guillon, B. Delacretaz, and S. Kitschke.

[10] http://www.tei-c.org .

[11] We use the ISO codes for the languages. Cf. Vaillant and Léglise (2014) for more details.

[12] This extract illustrates phenomena of contact between Kali'na (shown in normal type) and French (in bold).

[13] The sequence *para* + *ampé* used to form the interrogative *para ampé* (what for) is remarkable, because Purepecha has its own frequently used interrogative *anti* "why".

[14] The remarkable item is the linking of a medial inanimate demonstrative in Kali'na (demonstratives that in any case may be undergoing a linguistic change from demonstrative to definite article) with a French definite article (whose grammatical function is open to question).

[15] The insertion of a French item composed of a noun and its determinant into an utterance in Nengee is remarkable, since the majority of occurrences of French nouns in Nengee environments are without determinants.

[16] The expected form would be 'prendre ça dans ma main' so that the observed 'tirer ça dans ma main' is remarkable.

[17] When two varieties of the same language are presented, we use the forms A', A'', etc.