Les corpus plurilingues, entre linguistique de corpus et linguistique de contact

Isabelle Léglise, Sophie Alby

Les corpus plurilingues, entre linguistique de corpus et linguistique de contact : réflexions et méthodes issues du projet CLAPOTY[*]

Isabelle Léglise[**] et Sophie Alby[***]

Le domaine de la linguistique de contact est en pleine expansion depuis une quinzaine d'années mais fragmenté en plusieurs traditions de recherches. Une première approche, diachronique, tend à se focaliser sur la description linguistique des conséquences du contact sur les langues (étude des *contact-induced language change*). Une seconde approche, synchronique, vise plus à décrire les effets du plurilinguisme et du sens socialement attribué par les locuteurs à l'alternance des langues (étude en particulier du *codeswitching*). Peu de travaux tentent de tenir compte des avancées de ces deux traditions de recherche. Nous voudrions ici montrer quelle méthodologie précise, dans l'analyse minutieuse des corpus, peut être mise en œuvre afin de prendre en compte à la fois les phénomènes synchroniques (de variation et mélanges de langues) et diachroniques (de changement) et quelles questions épistémologiques se posent dans le traitement des données. Pour ce faire, nous nous appuyons sur la méthodologie mise en place dans le projet CLAPOTY. Après avoir présenté le champ de la linguistique de contact et l'optique de la linguistique de corpus, nous présentons notre corpus plurilingue et discutons des choix effectués au regard des standards actuels des travaux en linguistique de corpus. Nous présentons ensuite les méthodes de repérage et d'analyse mises en place afin de pouvoir proposer des explications plurifactorielles au contact de langues.

1. LE DOMAINE DU CONTACT DE LANGUES

La prise en compte d'une pluralité de langues en présence sur le terrain est devenue un incontournable pour les linguistes travaillant à la description des

[*] Contacts de Langues : Analyses Plurifactorielles assistées par Ordinateur et conséquences TYpologiques (projet ANR-09-JCJC-0121-01). La tâche 1, présentée ici et dont nous sommes responsables, s'attache à la réalisation d'un corpus commun et à la création d'un modèle d'analyse de phénomènes de contact. Ont également participé à cette tâche : E. Adamou (CNRS, Lacito), C. Chamoreau (CNRS, SeDyL-CELIA), G. Ledegen (Rennes 2), B. Migge (UCDublin et SeDyL-CELIA), C. Saillard (Paris Diderot, LLF), D. Troiani (CNRS, SeDyL-CELIA), et P. Vaillant (Paris Nord, Lim&Bio).
[**] CNRS, SeDyL-CELIA. Courriel : leglise@vjf.cnrs.fr
[***] UAG, SeDyL. Courriel : Alby.sophie@gmail.com

langues en général ; les situations de multilinguisme sociétal sont la généralité et le monolinguisme individuel, un cas particulier (Wurm, 1996). Les situations de communication ordinaires prennent ainsi place, non pas dans des communautés linguistiques monolingues vues comme «homogènes», mais dans une «zone de contact» multilingue. Toutefois, les phénomènes de contact sont encore souvent traités à la marge, comme des épiphénomènes (Nicolaï, 2007 : 2). Issue de la linguistique historique moderne, la linguistique de contact (Goebl, Nelde, Starý & Wölck, 1996) met les phénomènes de contact au centre de ses préoccupations. Si toutes les langues sont mixtes, au sens faible du terme, (Thomason, 2003 : 21), beaucoup de travaux se sont penchés sur les langues mixtes, au sens fort, c'est-à-dire sur les langues qu'on ne peut génétiquement affilier à une seule langue. L'étude des créoles et pidgins a notamment démontré l'importance des facteurs socio-historiques dans le changement linguistique car les mécanismes et processus linguistiques intervenant lors de la genèse des créoles sont les mêmes que dans des situations «classiques» (Winford, 1997, Thomason, 1993). Winford (2003) insiste sur l'importance des facteurs sociaux dans la typologie des situations de contact qu'il propose. Toutefois, son travail s'intéresse essentiellement aux similarités et différences des résultats linguistiques du contact et peu aux éléments de contexte micro et macro-social – probablement parce qu'il étudie des situations anciennes pour lesquelles peu de données sociales sont disponibles (Léglise & Migge, 2005). Beaucoup de travaux se sont consacrés aux changements linguistiques induits par contact – *contact-induced language change* – en s'intéressant aux types de phénomènes susceptibles d'apparaître en fonction des caractéristiques typologiques des langues en présence (*cf.* Heine & Kuteva, 2005, Thomason, 2001b, Ross, 1999). En se focalisant sur des caractéristiques morphosyntaxiques ou typologiques, ces travaux ont toutefois laissé de côté les considérations sociales ou contextuelles du contact de langues. Dans une perspective fonctionnaliste par exemple Matras (2009) considère les répertoires plurilingues des locuteurs et les innovations individuelles comme agents du changement.

En synchronie, l'étude de l'alternance de langues et des parlers bilingues s'est développée dans une tradition autonome, se subdivisant elle-même en deux approches, l'une grammaticale, l'autre pragmatique. La première vise à déterminer la structure linguistique des productions bilingues (Poplack, 1980, Muysken, 1995, 2011, Myers-Scotton, 1993b, 2002, Backus, 2003). Différents modèles ont été proposés pour prédire la bonne formation des alternances et les contraintes linguistiques pesant sur elles (cf. par exemple celui de la langue matrice proposé par Myers-Scotton, 1993b). La seconde approche s'intéresse au rôle et aux significations sociales de l'alternance de langues (Auer 1995, 1999, Myers-Scotton, 1993a). Les travaux visent alors à déterminer la fonction communicative des alternances ainsi que leur fonction sociale, en tant que marque identitaire permettant de distinguer des groupes sociaux.

En France, la co-existence de deux expressions, l'étude des «contacts de langues» et celle des «langues en contact», dessine des lignes de partage tant disciplinaires que méthodologiques ou théoriques (Léglise, 2007a). L'étude des «contacts de langues» renvoie majoritairement, depuis une quinzaine d'années, à

des travaux dans une perspective de sociologie du langage et d'écologie des langues ayant mené à de très nombreuses publications (cf. entre autres Deprez, 1994, Juillard, 1995, Boyer, 1997, Canut & Caubet, 2002, Billiez, 2003). Ils s'intéressent à l'étude d'un ensemble de phénomènes : multilinguisme sociétal, diglossie, interactions plurilingues, ou abordent le contact à un niveau épistémologique, critiquant la notion saussurienne de langue et proposant de déplacer les frontières linguistiques à un autre niveau – celui des discours (Canut, 2001) ou celui des répertoires linguistiques (Nicolaï, 2005). Plus récemment, l'étude des «langues en contact» s'est développée parmi les linguistes descriptivistes avec un intérêt particulier pour les conséquences linguistiques des contacts dans une perspective structurale ou fonctionnaliste (cf. entre autres Kriegel, 2003, Chamoreau & Lastra, 2005, Chamoreau & Goury, 2012). Influencés par les travaux de Croft (2000), Field (2002), Heine & Kuteva (2005), Matras (2009), ou encore Ross (2007), ces auteurs se penchent sur des phénomènes de restructuration tels que l'emprunt, le calque, ou la grammaticalisation. Quelle que soit l'approche, on note un grand éclatement actuel des travaux, tant dans le domaine sociolinguistique où les études s'attachent à décrire des situations de contact variées, uniques et étrangères les unes aux autres, que dans le domaine descriptif, où les données sur lesquelles s'appuient les chercheurs sont spécifiques à la langue sur laquelle ils travaillent et ne sont que peu comparées entre elles.

On sait que le changement en situation de contact de langues a presque toujours des causes multiples (Thomason, 2001a). En diachronie, pour rendre compte d'évolutions phonétiques ou morphosyntaxiques, les travaux s'appuient sur des tendances internes aux langues et mentionnent souvent le contexte socio-historique qui permet de justifier un potentiel effet du contact de langues. Ces facteurs explicatifs sont toutefois encore rarement intégrés ensemble (Chamoreau & Léglise, 2012). En synchronie, les travaux descriptifs, en se concentrant sur des facteurs linguistiques ou typologiques laissent peu la place à des facteurs explicatifs liés au contexte des interactions et aux locuteurs qui y sont impliqués. Quant aux travaux sociolinguistiques, ils ne s'intéressent généralement pas à expliquer les conséquences linguistiques du contact de langues – mais se concentrent plutôt sur le contexte ou l'usage.

Le projet CLAPOTY (Léglise, 2009) part du constat de la fragmentation de ce champ de recherche en traditions n'ayant pas pour habitude de dialoguer. Ces deux traditions se rejoignent sur l'impact des facteurs sociaux sur le changement linguistique et sur les phénomènes de contact, ce qui constitue une avancée majeure dans la mesure où la linguistique historique s'est longtemps contentée d'étudier les motivations internes au changement (Thomason & Kaufman, 1988 : 1). Mais l'analyse précise des situations de contact, que ce soit au niveau macro-social ou au niveau micro-social, telle qu'appelée de ses vœux par Weinreich (1953) pour comprendre les phénomènes de contact, est loin d'être réalisée. Le projet CLAPOTY s'est donné pour objectif de prendre en compte des phénomènes synchroniques et diachroniques dans toute leur complexité en mettant en relation les facteurs sociaux habituellement pris en compte dans le cadre de l'anthropologie linguistique, de la pragmatique ou de la sociolinguistique, tout en

affinant les facteurs linguistiques traditionnellement pris en compte par la linguistique descriptive et typologique. L'ambition du projet est d'analyser, rendre compte et expliquer les phénomènes de contact en mettant en place une méthode d'analyse qui prenne en compte les connaissances issues de ces différents sous-domaines, et qui s'allie à des outils informatiques de recherche puissants élaborés spécifiquement pour ce programme. Par ce biais, les chercheurs impliqués espèrent pouvoir créer un cadre explicatif multi-niveaux et multi-factoriel des phénomènes de contact en se basant sur des langues typologiquement variées et des situations sociolinguistiques diverses.

Dans cet article, nous rendons compte de la démarche mise en œuvre et des méthodes et procédures mises en place pour analyser les conséquences linguistiques des contacts de langues au travers de cinq niveaux d'analyse : morphosyntaxique, interactionnel, sociolinguistique, pragmatique et typologique.

2. LINGUISTIQUE DE CORPUS ET CORPUS PLURILINGUES

En linguistique de corpus, une longue tradition de travail s'est instituée ces quinze dernières années sur les corpus multilingues – c'est-à-dire des corpus comprenant des textes dans différentes langues, ces textes étant a priori chacun monolingue. Si possible, la linguistique de corpus sur corpus multilingues s'effectue sur des corpus de textes comparables (dans chaque langue le nombre et le genre ou type de textes sont comparables) (cf. notamment McEnery et al. 2000 ; Déjean, Gaussier et Sadat, 2002), parfois sur des corpus parallèles (c'est-à-dire, des textes et leurs traductions (Véronis, 2000 pour une présentation)) voire sur des corpus parallèles multilingues alignés (des corpus parallèles pour lesquels on a des relations d'équivalence de traduction entre des éléments qui composent les textes), cf. notamment (Véronis, 2002, Zweigenbaum et al., 2011).

Pour différencier de ces corpus multilingues, les corpus sur lesquels se penchent les linguistes de contact que nous sommes, nous utilisons ici le terme de «corpus plurilingues» c'est-à-dire de corpus comprenant plusieurs langues au sein de mêmes textes (interactions spontanées plurilingues illustrant des phénomènes de codeswitching ou de mélange entre plusieurs langues par exemple). Ces corpus plurilingues, à la différence des corpus multilingues précédemment cités, sont encore peu nombreux, peu disponibles à la communauté des linguistes, et peu «outillés» du point de vue des traitements informatisés disponibles. On peut citer la base ICOR de la plateforme CLAPI[1] qui comporte quelques données plurilingues, le projet LIPPS/LIDES[2] dont l'objectif était de développer des standards de transcription pour les langues mixtes et le codeswitching ou la base Bilingbank accesssible sous Talkbank[3].

Dans le domaine de la typologie linguistique et des études de la variation inter- ou translinguistique, les corpus parallèles ont progressivement fait leur apparition (Dahl, 2007, Stolz, 2007) à côté des travaux par comparaison de questionnaires

[1] *Cf.* http://clapi.univ-lyon2.fr et http://icar.univ-lyon2.fr/projets/corinte/.

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

[3] *Cf.* http://talkbank.org

(Matras et Sakel 2007). De larges corpus ont été constitués quelle que soit la méthode de recueil (traductions comparables et parallèles ou réponses à des questionnaires). Des bases de données ont également été développées mais, comme les corpus sur lesquels ces études se fondent, elles sont toutefois composées, pour la plupart, de données monolingues comparées[4]. Le phénomène de l'emprunt (*borrowing*) fait par exemple l'objet d'annotations particulières dans les corpus monolingues ; c'est le cas dans la base de données sur le romani[5] qui annote la «profondeur» de l'emprunt (Matras,White et Elšík, à paraître).

Les corpus plurilingues sont pourtant particulièrement intéressants car les problèmes de variation et de formes non-standard, souvent ignorés par les grands corpus, ou contrôlés par des paramètres généraux (comme les types de textes ou de discours recueillis), y sont centraux. Ils illustrent souvent non seulement ce qu'on considère généralement comme de la variation interne aux langues – par exemple des variations morphosyntaxiques ou lexicales (et que l'on peut parfois relier à des pratiques stylistiques ou dialectales) mais également des formes difficile à catégoriser. Les corpus présentant du codeswitching ou du code-mixing produits par des locuteurs plurilingues aux compétences variées (parfois en cours d'acquisition) posent en effet – comme nous le verrons plus loin – de redoutables problèmes non seulement d'identification des formes mais aussi de transcription et d'annotation. De même, la définition même des corpus plurilingues et le choix des situations de parole à documenter dépassent aussi ce qui est réalisé dans le domaine de la documentation des langues peu décrites (Migge & Léglise, 2013) et pour des grands corpus de référence.

Le projet CLAPOTY se situe dans une linguistique de corpus sensible aux corpus hétérogènes et a développé des outils pour travailler sur ces corpus. Toutes ces questions – de repérage, notation, annotation – ont fait l'objet de longues discussions parmi les membres du projet. Nous présentons quelques-unes des solutions retenues ci-dessous en explicitant les choix méthodologiques et épistémologiques qui se posaient à chaque fois que possible.

3. LE CORPUS CLAPOTY

3.1. La constitution d'un corpus commun : une nécessaire harmonisation

Pour remédier au manque de corpus plurilingues disponibles dans la littérature et pour avoir une base de travail commune à toute l'équipe, un corpus commun a été réalisé. Il est constitué de discours spontanés qui avaient été transcrits au départ selon des traditions et avec des objectifs assez différents comme le montrent les trois exemples ci-dessous :

(1) Clapoty_Léglise (nengee – **variété de français** non native ou langue seconde)
J — K̄en san i e suku e fuufeli a ini **maman chambre** anda ?
«Ken qu'est-ce que tu cherches ? tu es en train de déranger la chambre de maman»

[4] Par exemple Database Typological Research http://languagelink.let.uu.nl/tds/index.html
[5] Projet dirigé par Y. Matras (Romani Morpho-Syntactic Database Project, Université de Manchester, http://romani.humanities.manchester.ac.uk/rms/).

M — a na faansi i mu taki a djuka
«tu ne dois pas parler français mais ndyuka»
— Ken san i meki a sikoo tide ?
«Ken qu'est-ce que tu as fait à l'école aujourd'hui ?»

K — **ce que je faire à l'école ?** […] tide mi meki **bonhomme** a sikoo anga **plus**
«Ce que j'ai fait à l'école ? Aujourd'hui j'ai dessiné des bonhommes et aussi j'ai fait des additions»

M — pikin man i mu taki
«Petits hommes tu dois dire»

(2) Clapoty_Alby (français-**kali'na**)

1 E Aino' yemamɨ kapilɨ ɨwa man' yemamɨ molo man[6] ++ ayalanatoko loten[7]
«Attends ! Je dois faire mon travail. Ca c'est mon travail. Vous n'avez qu'à parler.»

2 D **Kosi'**
«Flute !»

3 Y Non' sérieux' **Oti poko awu wekatuya' +** c'est pas que **akinupewa to'**[8]
«Non, sérieux ! Pourquoi est-ce que je cours ? C'est pas que je sois paresseux !»

4 E [caf] man mei'
«Tu étais avec une fille ?»

(3) Clapoty_Chamoreau (purepecha-**espagnol**)[9]

inte	acha	**mas**	khéri-e-s-ti	**ke**	**de**	xo	anapu	yamintu
dem	homme	plus	grand-PRED-AOR-ASS3	que	de	ici	origine	tout

Cet homme est plus âgé que tous ceux d'ici. (Lit. Cet homme est plus âgé que de tous)

L'exemple 1 montre l'alternance entre un créole à base anglaise, le nengee, et des éléments en français. Ce qui intéressait son auteur à l'origine (Léglise, 2007b), c'était de pouvoir visualiser l'alternance entre ces différentes langues, d'où le choix de codes graphiques différents (gras pour marquer les éléments en français et times normal pour marquer les éléments en nengee). La notation des locuteurs était par ailleurs importante afin de repérer qui procède à quel type d'alternance et quelles sont les compétences linguistiques en compréhension et en production ainsi mises en œuvre par ces derniers. Dans l'extrait choisi, M s'exprime toujours en nengee, à la différence de K et de J qui intègrent des

[6] E. explique qu'il doit enregistrer leurs conversations.
[7] Il leur dit de parler, de dire n'importe quoi, de ne pas faire attention au magnétophone.
[8] Y. dit n'importe quoi, juste pour tester l'enregistrement, pour commencer à parler.
[9] Liste des abréviations utilisées : 2,3 (deuxième, troisième personne), ADP (adposition), ADV (adverbe), CAUS (causatif), DEM (démonstratif), DET (déterminant), FOC (focus), FT (formatif) GEN (génitif), GN (groupe nominal), HAB (habituel), IND (indépendant), INT (interrogatif), N (nom), OBJ (objet), PRT (particule), PRS (présent), PRTEN (particule énonciative).

éléments de français dans leurs énoncés en nengee. Mais, à la dernière ligne, M reformule «bonhomme» dans sa langue, montrant ainsi qu'elle comprend le français. Un autre élément intéressant concerne la forme des éléments français observés : à la première ligne, «maman chambre» suit l'ordre des constituants du nengee ; à la quatrième ligne, la forme verbale infinitive observée «ce que je faire» illustre la compétence partielle du jeune locuteur, en cours d'acquisition et d'apprentissage scolaire du français.

La transcription de l'exemple 2 visait au départ à fournir les indications nécessaires pour la réalisation d'une analyse interactionnelle portant sur les discours bilingues kali'na/français produits par des enfants et adolescents d'un village du nord-ouest de la Guyane (Alby, 2001). Les interlocuteurs à qui chaque locuteur s'adresse sont identifiés (E s'adresse à Y ligne 4, Y s'adresse à l'ensemble du groupe ligne 3), les prises de parole sont numérotées, de même que les montées intonatives, marquées par l'apostrophe (kosi' ligne 2), etc. Ces différents éléments permettent de mieux comprendre la fonction des alternances codiques, comme l'emphase (Y ligne 3) ; mais aussi de décrire les caractéristiques du mode mixte utilisé par les interlocuteurs lorsqu'ils sont au sein d'un groupe de pairs. Par exemple, ligne 3, on observe une absence de pause dans l'énonce "c'est pas que akinupe wa to" ce qui peut indiquer que le passage d'une langue à l'autre n'est pas à considérer comme un problème de compétence, mais comme le signe d'une variété bilingue correspondant à un groupe socialement identifié.

L'exemple 3 (Chamoreau 2012a) pour sa part – qui illustre également l'insertion d'une structure comparative espagnole «mas que… de» dans un énoncé en purépecha, une langue amérindienne du Mexique – avait été transcrit afin de donner à voir la composition morphosyntaxique de l'énoncé avec une ligne de glose interlinéaire obligatoire, mais sans mention des locuteurs ou des tours de parole. L'intérêt du chercheur était ici la description grammaticale des phénomènes sans prise en compte d'aspects interactionnels. C'est la comparaison de cet exemple 'décontextualisé' et d'autres exemples similaires (par exemple les énoncés présentés dans (Chamoreau, 1995) qui donne son sens au choix de transcription (morphème par morphème avec repérage des langues).

Comme on le voit, les éléments communs au départ parmi les différents linguistes participants au projet étaient minimes et se résumaient à 1) travailler sur des données de première main transcrites soit orthographiquement soit en API, 2) traduire ces données (avec au moins une traduction libre), 3) vouloir travailler sur l'hétérogénéité de ces corpus en traitant à la fois de faits de variation et de faits d'alternance des langues. Le projet CLAPOTY a donc nécessité de très longues discussions parmi ses membres sur l'harmonisation des transcriptions et des annotations[10].

[10] Nous entendons ici par «annotation» tout enrichissement, par les linguistes, de la transcription, en particulier les annotations qui précisent de quelle langue il s'agit, les annotations morphosyntaxiques, les parties du discours etc. Nous préciserons plus loin lorsqu'il s'agit de métadonnées de différents ordres (typologiques, sociolinguistiques etc.).

3.2. Un corpus plurilingue et hétérogène

Les enregistrements à verser dans notre corpus commun ont été choisis à partir des données des différents participants, en fonction de l'intérêt qu'ils représentaient du point de vue de leur hétérogénéité intrinsèque, mais aussi du point de vue de la diversité des langues en général (en terme de diversité géographique et typologique). Nous travaillons pour la plupart sur des langues (ou des variétés de langues) peu décrites[11] en contact parfois avec des langues pour lesquelles une grande tradition grammaticale existe. Le corpus Clapoty compte à ce jour 40 langues dont 25 pour lesquelles les membres sont «spécialistes» (les initiales des membres sont notées entre crochets).

Groupes de langues	Langues présentes dans les corpus
amérindiennes	kali'na [SA] ; nahuatl, purépecha (et différentes variétés géographiques) [CC]
créoles à base française	antillais (guadeloupéen, martiniquais) [PV] ; guyanais [IL, PV] ; réunionnais [GL], haïtien
créoles à base anglaise	nengee (ndjuka, pamaka, aluku) [BM] ; sranan tongo [BM]
créoles à base portugaise	casamançais [JFN]
romanes	français [IL, GL] (et différents variétés, stylistiques, et géographiques) ; espagnol (et différents variétés, notamment du Mexique) ; portugais (du Brésil)
germaniques	anglais (et différentes variétés, notamment de la Caraïbe) ; néerlandais
balkaniques	grec ; romani [EA] ; turc
est asiatiques	chinois (mandarin, minnan) [CS] ; langues aborigènes de Taiwan (amis, taroko) [CS] ; japonais
niger-congo (atlantique)	wolof

Tableau 1 : langues représentées dans le corpus Clapoty

A cette diversité typologique et géographique recherchée, s'ajoute une diversité souhaitée en termes de situations sociolinguistiques représentées dans les corpus ; ces derniers illustrent à la fois des contacts entre variétés dialectales de la même langue (par exemple différentes variétés de purepecha en contact, ou différentes variétés de nengee en contact), des contacts entre des variétés stylistiques d'une même langue (par exemple des éléments assimilables à une façon de «parler jeune» que nous identifions comme tels intégrés dans une façon «standard» ou «ordinaire» de parler telle langue), des contacts entre des langues vernaculaires, des contacts entre des langues vernaculaires et des langues véhiculaires, des contacts entre des langues dites «de grande diffusion» internationale, des contacts entre des langues considérées comme minoritaires ou majoritaires etc. (cf. plus loin la manière dont nous procédons pour prendre en compte ces différents paramètres).

L'hétérogénéité des corpus s'actualise enfin au niveau de la diversité des types d'interactions représentées. Notre corpus commun comptabilise actuellement 170

[11] Et qui nécessitent un travail de description minutieuse.

enregistrements transcrits, soit 170 transcriptions se présentant – pour reprendre la typologie proposée par Vion (1992) – sous la forme d'interactions à structure d'échange et d'interactions sans structure d'échange.

Ces interactions comportent toutes au moins deux (variétés de) langues, et comptent parfois jusqu'à une dizaine de langues en présence. La majorité des interactions à structure d'échange de notre corpus est par ailleurs composée d'au moins trois interlocuteurs et compte parfois une dizaine ou une trentaine d'interlocuteurs. Afin de donner à voir ici la diversité de situations de communication, nous pouvons répartir ces interactions en grands domaines (Fishman, 1972) et en types de textes : par exemple, les 13 interactions sans structure d'échange, peuvent être déclinées en quatre catégories : compte-rendu politique, monologue dans les médias, conte, récit. Nous pouvons par ailleurs sous-catégoriser les interactions à structure d'échange selon les relations que les interlocuteurs entretiennent et donc selon la symétrie ou l'asymétrie de leurs rôles respectifs : interactions symétriques et interactions asymétriques (Vion, 1992). On comptabilise ainsi 67 interactions symétriques et 77 interactions asymétriques. Les interactions symétriques correspondent à des interactions entre enfants dans des contextes familiaux ou scolaire, des interactions entre adolescents dans un cadre amical ou dans un cadre scolaire ; des interactions entre adultes dans un cadre amical, familial, dans les médias ou au travail.

Nous pouvons répartir les interactions asymétriques en quatre grandes catégories : interactions ayant eu lieu dans un cadre scolaire (qu'il s'agisse d'enregistrements réalisés dans des écoles, collèges ou lycées), interactions dans un cadre familial (repas de familles, discussions informelles entre plusieurs générations) ; situations d'entretiens ou d'interviews (en particulier avec un chercheur) ; interactions relevant du domaine professionnel ou des interactions de service (dans les domaine du commerce : sur le marché ou dans un magasin par exemple, de la santé avec un certain nombre d'enregistrements à l'hôpital entre soignants et patients, des services avec un certain nombre d'enregistrements entre agent et client). Le tableau suivant donne un aperçu de la diversité des interactions du corpus. Nous avons donc recherché une hétérogénéité maximale pour le corpus - hétérogénéité externe (liée aux types de textes et d'interactions du corpus) et interne (variation morphosyntaxique et présence d'éléments plurilingues).

Interactions à plusieurs participants	Total
en famille	12
à l'école	15
entre amis	24
dans les médias	15
en situation de travail	51
entretiens	27
Interactions sans structure d'échange	**Total**
discours politiques, récits, contes, …	13

Tableau 2 : types d'interactions du corpus

3.3. Annotation et encodage des corpus

Nos corpus sont balisés sous xml grâce à un éditeur, Jaxe[12], adapté par P. Vaillant au système d'annotation que nous avons élaboré. Nous renvoyons à Vaillant, Léglise & Alby (en préparation) pour de plus amples détails techniques sur le système d'annotation mis en place. A ce stade, il est important de noter que le schéma de document Corpus-Contact que nous avons créé s'inspire de normes de la TEI (Text Encoding Intiative)[13], adaptées pour nos besoins. Nous voudrions noter ici deux adaptations importantes.

La TEI découpe les textes en phrases. Dans notre cas, l'unité minimale dans laquelle nous souhaitons décomposer nos enregistrements est loin d'être la phrase, unité qui n'a pas de sens à l'oral, mais plutôt le tour de parole, suivant en cela la tradition initiée par Sacks, Schegloff & Jefferson (1974). Nous repérions déjà ces tours dans les exemples 1 et 2 présentés précédemment soit par l'initiale du locuteur, soit par le signe « – » voire par un numéro permettant ensuite de mieux citer le passage précis. Les prises de parole des locuteurs nous sont en effet apparues comme le découpage de nos transcriptions le moins discutable possible ; ce choix est conforme au cadre global des conventions adoptées par exemple par le groupe ICOR (2007) qui a développé des conventions de transcription en vue d'analyses interactionnelles plus spécifiques que nous ne suivons qu'en partie. Ce choix est également conforme aux choix réalisés dans d'autres projets de grande envergure sur l'oral, tels TalkBank et CHILDES sous CLAN (MacWhinney 2000, 2007).

La TEI propose de noter la langue de base de chaque phrase et, si un élément d'une autre langue intervient, elle le note entre chevrons, comme <élément étranger appartenant à la langue x>. Ce choix est également celui qui est réalisé dans le cadre du projet ANR CorpAfroas (Mettouchi & Chanard, 2010, cf. notamment Manfredi et al. (sous presse) qui s'intéressent à des phénomènes de codeswitching dans leur corpus de langues jusqu'alors peu décrites et peu documentées. Après avoir, dans un premier temps essayé d'associer aussi une langue[14] à chaque prise de parole, nous avons progressivement renoncé à l'attribution systématique d'une seule langue à chaque tour. Dans la plupart des cas, nous observons en effet plusieurs langues en présence dans le même tour, à l'intérieur de la même prise de parole par le même locuteur et nous avons décidé de noter ces tours comme «multilingues», et à l'intérieur de ces tours multilingues, nous identifions des «segments» associables à telle ou telle langue. Par exemple, la prise de parole suivante montre un début d'énoncé en français qui se poursuit en kali'na. Plutôt que de choisir – souvent arbitrairement – une langue matrice à l'énoncé, nous considérons que le tour est multilingue – ce que visuellement nous représentons par un surlignage jaune – et constitué de

[12] http://jaxe.sourceforge.net/fr/ dont les auteurs sont D. Guillaume, S. Ayadi, B. Tasche,O. Kykal, C. Dedieu, L. Guillon, B. Delacretaz, S. Kitschke

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

[14] Nous utilisons les codes ISO pour les langues. Cf. Vaillant, Léglise et Alby (en préparation) pour plus de détails.

plusieurs segments (ici deux langues différentes, repérées par les codes gras pour le français / normal pour le kali'na).

(4) match de foot (corpus Clapoty_Alby)

003-03	**ce n'est pas que**	akinupe	wa to↗
		paresseux	1.être PRTEN
	Je ne suis pourtant pas paresseux !		

De la même manière, des éléments peuvent appartenir à plusieurs langues possibles : dans des corpus d'alternance de langues enregistrés à la frontière entre la Guyane et le Surinam, l'adjectif [diʀɛkt] peut être considéré comme du français, de l'anglais, du néerlandais ou comme un emprunt à l'une de ces trois langues et, lors de nos transcriptions (ou de l'annotation de nos transcriptions), nous avons bien souvent expérimenté la difficulté qu'il y aurait à trancher. Ainsi, plutôt que de trancher, nous avons décidé d'étiqueter ces éléments comme eux-mêmes «multilingues» et d'identifier ensuite l'ensemble des possibilités associées. Visuellement, nous avons souhaité adopter un système de transcription qui montre les différentes possibilités. Pour des cas où les deux langues partagent un certain nombre de traits (en particulier lexicaux), comme une langue créole et sa langue lexificatrice, Ledegen (2012) a proposé d'utiliser une double transcription dite «flottante» afin de visualiser les deux interprétations possibles s'offrant au descripteur. Nous avons étendu cette notion à tous les cas où plusieurs transcriptions et plusieurs langues étaient possibles, même lorsque les langues ne sont a priori pas si «proches», par exemple, dans l'extrait suivant, un médecin tente de dire quelques mots dans la langue de sa patiente, il prononce ligne 11 «a go bon» (une forme que nous considérons comme non native) qui correspondrait à la forme standard «a e go bung» en nengee (la prononciation attendue de la finale nasale 'ung' étant un peu moins ouverte que le 'on' français). Il nous semble alors important de noter – lors de la transcription alternative – la proximité de ce qui a été prononcé avec d'une part l'adjectif «bon» qui semble sélectionné et d'autre part avec l'adjectif «bung» qui est peut-être la forme standard visée par le locuteur.

(5) *kosokoso* (corpus Clapoty_Léglise)

010 Inf1:	tu parles qu'elle va mieux qu'hier !	
011 Méd:	a go	*bon* ?
		bung
	aller bien ?	
012 F1:	**a e mama / mama fu mi e go ?**	
	ma mère elle va ?	
013 Méd:	mama ça *go bon* ?	
		go bung ?
	la mère ça aller bien ?	

Dans cet exemple, les transcriptions alternatives font parfois intervenir plus que deux langues, comme c'est le cas de la ligne 013 qui comprend des éléments attribuables au français (*ça* et *bon*), des éléments de nengee non natif (*mama* et *bung*), et un élément, *go*, qui peut être catégorisé comme du nengee ou comme de l'anglais. L'utilisation de cette méthode a ainsi permis de mettre en évidence que,

dans certains corpus, la quasi-totalité des prises de parole pouvait être attribuées à l'une ou l'autre langue comme dans l'exemple 6 :

(6) discussion entre hommes dans un bar à Saint-Laurent (corpus Clapoty_Migge)

002.B:	*a*	*fu den man*	**dati**	*ya*
	a	**fu den man**		
	FOC	pour l'homme	DEM	oui
	C'est à cause de ces hommes, oui			
003.C:	*i wani go*	na *dape*	*a didon*	
	i wani go		**a didon**	
	2 vouloir aller	là	3 allonger	
	Tu veux aller là où il est allongé ?			

Notre choix assumé de faire figurer l'ensemble des possibles sur les transcriptions transforme ainsi le regard que nous portions sur les corpus. Plutôt que de considérer l'extrait 6 comme du nengee dans lequel quelques éléments de sranan tongo s'insèrent, on peut dès lors considérer que les locuteurs utilisent préférentiellement des éléments communs aux deux langues lorsqu'ils s'expriment et que par moment, ils sélectionnent telle ou telle marque dans l'autre des langues (Migge & Léglise, 2011) appartenant à leur répertoire linguistique.

4. MÉTHODES POUR UNE DESCRIPTION ET UNE ANALYSE MULTINIVEAUX DES «PHÉNOMÈNES REMARQUABLES»

L'un des défis majeurs posés par ce type de données est de se doter d'outils pour décrire (puis analyser et expliquer) les phénomènes linguistiques observés dans des corpus plurilingues, qu'il s'agisse de phénomènes attribuables à de la variation (classiquement considérée comme interne ou due au contact de langues (Léglise et Chamoreau, 2013) ou de phénomènes de contacts plus évidents tels que le codeswitching ou le code-mixing. Le choix méthodologique que nous avons effectué est de qualifier (et d'annoter) tous les phénomènes sur lesquels nous souhaitions travailler comme «phénomènes remarquables». Nous utilisons «remarquable» dans les deux sens de l'adjectif : soit les phénomènes observés sortent de l'ordinaire (de la langue ordinaire) – et nous partons d'un sentiment d'écart par rapport à la forme attendue ou de référence (Léglise, 2012) pour qualifier la forme observée de «remarquable», c'est-à-dire digne d'un intérêt particulier, soit les phénomènes observés nous paraissent exemplaires de phénomènes connus et bien décrits dans la littérature sur le contact de langues – et nous partons d'un sentiment de fréquence ou d'exemplarité. Ce choix est minimaliste du point de vue de la terminologie choisie, il évite ainsi tous les termes particulièrement foisonnants dans le domaine du contact de langues et très souvent contradictoires (d'un auteur à l'autre ou d'un cadre à l'autre).

Il permet de ne pas entrer dans des débats terminologiques sans fin comme ceux qui existent sur la distinction entre emprunt et codeswitching[15] par exemple

[15] Plusieurs années de débats internes et infructueux entre emprunts et codeswitching nous ont finalement convaincus de la nécessité de ne pas nommer les phénomènes mais plutôt d'observer leur fonctionnement.

ou encore calque, interférence ou transfert, etc. (Mackey, 1976 ; Zentella, 1997). La position retenue est d'employer des termes les plus «neutres» possibles en regroupant les phénomènes intéressants à observer en méta-catégories regroupées selon leur comportement ou leurs caractéristiques : phénomènes remarquables morphosyntaxiques (PREMS), phénomènes remarquables interactionnels (PRINT), phénomènes remarquables discursifs (PREDISC). Une fois les phénomènes remarquables annotés, ils apparaissent alors surlignés en gris comme en (7), il s'agit de les décrire, par une approche multi-niveaux, puis de les expliquer (cf. 4.4 ci-dessous).

(7) conseil municipal (corpus Lescure /Alby)[16]

001	ot~~i~~ nok~~i~~ molo	**CCOG compt~~i~~ l~~i~~**	**rappeler** poko	s-~~i~~-yan	o'wainen
	euh euh DEM CCOG compte-GEN		rappeler occupé.à	1-mettre-PRS	2.vous
	je vous rappelle le compte de la CCOG				

4.1. Les PREMS

Le premier niveau d'appréhension des phénomènes remarquables concerne le matériau linguistique produit, au niveau de la chaine parlée, et la suite de ses caractéristiques morphosyntaxiques habituelles ou inhabituelles[17]. Pour décrire ce qui est remarquable, nous proposons d'utiliser une notation fondée sur nos catégories d'annotation des langues. Ce premier descripteur concerne l'endroit où se situe le phénomène remarquable dans la chaine parlée : ci-dessous le phénomène remarquable est noté entre crochet [ ], il peut concerner la présence consécutive d'un élément d'une langue A et d'un élément d'une langue B, ou la forme particulière d'un élément d'une langue A, ou encore l'insertion d'un élément d'une langue A dans une langue B etc. Nous présentons un exemple de chaque type ci-dessous :

a) [<segment de langue A><segment de langue B>] : la suite A + B est remarquable.

(8) le président Cardenas à Tanaco (corpus Clapoty Chamoreau)

006.M	para ampe=i	wé-ka-sïn-i	t'u	ima-ni	ú-ra-ni
	pour que=2	vouloir-FT-HAB-INT	2IND	DEM-OBJ	faire-CAUS-INF
	pourquoi veux tu faire cela ?				

[<para><ampe=ri ...>][18]

[16] Cet extrait illustre des phénomènes de contact entre le kali'na (qui apparaît en times normal) et le français (noté en gras). L'abréviation CCOG est couramment utilisée pour : Communauté des Communes de l'Ouest Guyanais.

[17] Les caractéristiques phoniques et prosodiques bien qu'intéressantes ne sont pas traitées actuellement et feront partie de développements futurs.

[18] Selon C. Chamoreau, ce qui est remarquable c'est la suite *para* + *ampé* «que, quoi» pour former un interrogatif *para ampé* «pourquoi» alors qu'il existe en purepecha est un interrogatif *anti* «pourquoi» dont l'utilisation est fréquente.

b) <segment de langue A [><] segment de langue B> : la jointure entre A et B est remarquable.

(9) ABC (corpus Clapoty Lescure_Alby)
001-16 ot~~i~~ réserve molo la Basse Mana
euh réserve DEM la Basse Mana
euh la réserve de la Basse Mana
[<molo><la] basse mana>[19]

c) <SgA [<SgB>]> : la présence du segment B à l'intérieur du segment A est remarquable.

(10) *kosokoso* (corpus Clapoty Léglise)
132.F1: **efu yu wani sabi ala sani fa** la famille **da mi mu sabi fi yu seefi**
si tu veux tout savoir de la famille, alors je dois aussi pouvoir savoir des choses sur la tienne
<sabi ala sani fa[<la famille>]da mi …>[20]

d) <SgA [ ]> : ce qui est remarquable se situe à l'intérieur du segment A.

(11) je suis pas ton *blada* (corpus Clapoty Léglise)
013.K: oh mais c'est le **ga** gars qui tire ça dans ma main **bay**
oh mais c'est le type qui me prend ça, donne !.
<tire ça dans ma main>[21]

C'est seulement à l'issue de cette première étape descriptive que nous expliquons en quoi le phénomène nous paraît intéressant et que nous renvoyons à la littérature sur la question si le phénomène est déjà traité. Nous identifions comme PREMS de nombreux phénomènes, qu'il est possible ensuite de catégoriser en fonction des éléments morphosyntaxiques concernés. Nous avons ainsi adopté une typologie des PREMS afin de définir des espaces de comparaison parmi tous nos corpus : PREMS concernant le groupe verbal, PREMS concernant le groupe nominal etc. Par exemple, nous sommes en train de travailler collectivement sur les PREMS touchant les groupes nominaux (détermination, possession etc.). Le groupe nominal constitue un des domaines majeurs des structures grammaticales (Heine et Kuteva, 2008 ; MacSwan, 1997). Du fait de la définition large et a minima des «phénomènes remarquables» présentée plus haut, nous nous intéressons tant à des GN où deux langues sont

[19] Selon S. Alby, ce qui est remarquable c'est la jonction entre un démonstratif médial inanimé kali'na (démonstratifs qui par ailleurs pourraient être en phase de changement linguistique de type démonstratif vs. article défini) et un article défini français (dont la fonction grammaticale pose par ailleurs question).
[20] Selon I. Léglise, l'insertion d'un élément français composé d'un nom et de son déterminant dans un énoncé en nengee est remarquable, la plupart des autres cas de N français dans des environnements de nengee apparaissent en effet sans déterminant.
[21] D'après I. Léglise, la forme attendue serait «prendre ça dans ma main» plutôt que «tirer ça dans ma main» observé, et donc remarquable.

exprimées comme dans : [<**owi>** <maison>] (une maison) ou [<titre de propriété> <**a>**] (le titre de propriété), qu'à des GN où plusieurs langues possibles ont été identifiées (comme dans [**tout** papier/**papié a**] analysable comme <**tout**><papier> <a> ou <**tout papié a**> (tous les papiers)) qu'à des variations au sein d'une même langue (comme dans l'exemple <[le] maison>]) ou encore à des phénomènes de changement (comme dans l'exemple [<ma achati[22]>] (un homme)).

Les exemples ici présentés montrent des phénomènes de contacts sur lesquels une longue littérature existe en linguistique de contact, par exemple en ce qui concerne l'ordre des mots (qui suit ou non celui de la langue où le nom est exprimé) dans les recherches sur le codeswitching (entre autres Belazi, Rubin et Toribio, 1994 ; Nishimura, 1985 ; Bentahila et Davies, 1983 ; Mahootian, 1993 ; Gumperz, 1976 ; Bokamba, 1989 ; Poplack, 1981, MacSwan, 1997); mais aussi en ce qui concerne des phénomènes de restructuration comme la grammaticalisation du numéral en article indéfini ou celle du démonstratif en article défini (cf. entre autres Givon, 1981; Heine, 1997; Dryer, 2005a, 2005b).

4.2. Les PRINT

Les corpus plurilingues regorgent également de phénomènes intéressants au niveau interactionnel, c'est-à-dire au niveau des choix de langues effectués par les interlocuteurs. Ces choix et alternances peuvent se situer au sein même des tours ou prises de parole, auquel cas nous traitons le phénomène comme PREMS, mais ils se situent surtout, et le plus souvent, entre différentes prises de parole. Il s'est donc avéré nécessaire de pouvoir qualifier certaines séquences comme «remarquables». Pour ce faire, nous avons opté pour une approche structurelle des interactions basée sur le principe de leur séquentialité.

Les alternances codiques produisent des effets variés dans l'interaction (Auer, 1995 ; Alby et Migge, 2007) qui ont fait l'objet de nombreuses descriptions dans le but d'identifier les fonctions communicatives qui leur sont sous-jacentes[23]. Ces études ont conduit à la création de listes de fonctions attribuables aux alternances conversationnelles[24] dépendantes soit des langues soit des situations observées,

---

[22] Grammaticalisation en purepecha du numéral "un" en un article indéfini (Chamoreau, 2012b).

[23] Notons en outre que des auteurs comme Poplack et Sankoff (1988) insistent sur la nécessité de ne pas différencier sur ce point les stratégies conversationnelles des monolingues (changement de registre) de celles des bilingues, la seule différence se situant dans l'utilisation de deux variétés de langues différentes chez les bilingues.

[24] Zentella (1997) en dénombre 22 regroupées dans trois catégories (alternances liées à un changement de rôle des interlocuteurs ou au contrôle du comportement de l'interlocuteur, alternances visant à une clarification ou à une emphase, alternances «béquilles» qui servent à combler une lacune lexicale ou autre). Pour Alvarez-Caccamo (1990) c'est le style conversationnel qui permet d'expliquer les passages d'une langue à l'autre (mode humoristique, de la dispute, du discours rapporté). Tandis que Deprez (1991) considère que le «code-switching conversationnel» permet de produire des effets de sens variés dirigés vers l'interlocuteur ou vers le propos.

or ces listes sont infinies du fait même de la créativité inhérente à l'alternance. Auer (1995, 1999) propose comme alternative une typologie basée sur la séquentialité qui s'appuie sur le modèle de l'analyse conversationnelle. Nous l'avons adoptée pour coder, dans nos corpus plurilingues, les langues et les interlocuteurs par des chiffres et des lettres afin de pouvoir mettre en évidence les séquences interactionnelles faisant apparaître des changements de langues. Concrètement, la codification se situe au niveau de la prise de parole. Chaque langue est identifiée par une lettre dans l'ordre d'apparition du corpus (il n'y a pas de hiérarchisation du type la langue «A» est la langue «la plus utilisée») et chaque locuteur est identifié par un numéro selon le même principe comme nous le voyons dans l'extrait (12). Par la suite nous portons notre attention soit sur la forme de l'ensemble de l'interaction, soit sur des séries d'échanges. Ainsi, dans l'exemple suivant, la première langue (français) est codée A et le premier locuteur (J.) est codé 1, la deuxième langue dans l'ordre d'apparition (kali'na) est codée B et le deuxième locuteur dans l'ordre d'apparition (S.) est codé 2[25].

(12) je lui parle en français ? (corpus Clapoty Alby)

023.S:	une panier?	**A1**
024.J:	oui	**A2**
025.S:	(elle chante) // j'ai fini / ça y est!	**A1**
026.J:	**oeneko te senepoya owa** *regarde, je te le montre*	**B2**
027.S:	**uwa** *non*	**B1**
028.J:	c'est bon?	**A2**

L'échange peut se décrire de la façon suivante : «A1 A2 A1 A2 : le locuteur 1 et le locuteur 2 parlent tous les deux dans la variété A» jusqu'à la ligne 026 où le locuteur 1 passe à la variété B et est suivi par le locuteur 2 à la ligne suivante. Puis à partir de la ligne 028, les deux locuteurs recommencent à utiliser la variété A. Une telle approche permet de traiter aussi des cas où plusieurs langues sont employées dans la même prise de parole comme dans l'exemple (13) où l'on observe l'insertion d'éléments de la variété B (français) dans un discours qui semble par ailleurs être organisé selon les caractéristiques morphosyntaxiques de la variété A. L'insertion est marquée par l'utilisation des crochets.

(13) conversation informelle entre ABC (corpus Clapoty Lescure_Alby)

001-15	molo ot~~i~~ **nature garde** *le garde chasse*	**A[B]1**
001-16	ot~~i~~ **réserve** molo **la Basse Mana** *euh la réserve de la Basse Mana*	**A[B]1**
001-17	asito am~~i~~ man ne telapa moko kali'na / ot~~i~~ ~~i~~newala katako? *c'est un peu déjà le Kali'na, euh comment dit-on ?*	**A1**
001-18	moko kali'na ot~~i~~ **terrain de chass~~i~~l~~i~~** kanaiyan sipol~~i~~ pamen *le terrain de chasse du Kali'na comme dit le Blanc*	**A[B]1**

[25] Lorsque deux variétés d'une même langue sont présentes nous proposons d'employer les formes A', A'', etc.

Ainsi A[B]1 se lit «le locuteur 1 parle en variété A (kali'na, apparaissant en times normal) en insérant des éléments de la variété B (français, noté en gras)». Nous traitons de la même manière les cas où trois, quatre, cinq langues sont en présence dans un même corpus ou au sein du même tour. Nos corpus sont particulièrement intéressants car ils font la plupart du temps intervenir plus de deux locuteurs et plus de deux langues ce qui n'a jusqu'à présent pas fait l'objet de typologies dans la littérature. En effet, les modèles proposés se basent sur deux langues et deux locuteurs, par exemple l'analyse séquentielle proposée par Auer (1995, 1999) se base sur la présence de deux langues (A & B), de même que la typologie des interactions verbales proposée par de Pietro (1988) est un modèle bilingue et non plurilingue.

L'intérêt d'un tel mode de description est que, comme pour les PREMS, elle est réalisée a minima. Ce n'est que dans un retour réflexif sur la totalité des corpus, que nous pourrons, sur la base de toutes les séquences identifiées, revenir sur l'organisation structurelle des interactions dans les contextes plurilingues variés qui caractérisent nos corpus. Au cours de l'analyse, nous chercherons ensuite à vérifier dans quelle mesure les typologies proposées dans la littérature sont validées ou invalidées par nos données. Nous sommes actuellement en train d'essayer de résoudre le problème technique qui consiste à pouvoir annoter plusieurs lignes de corpus simultanément. Il est en effet essentiel que les caractéristiques interactionnelles puissent être annotées de manière à pouvoir ensuite comparer des séquences ayant des structures similaires et à pouvoir les relier aisément aux métadonnées présentées en 4.5. Obtenir ce résultat constituerait une véritable avancée dans le domaine dans la mesure où jusqu'à présent les modèles proposés se fondent sur des outils «manuels» ou sur des analyses de contenu (Matthey et de Pietro, 1997 ; Alfonzetti, 1998 ; Alvarez-Caccamo, 1990).

4.3. Les PREDISC

La littérature sur les contacts de langues abonde sur les phénomènes touchant les «petits mots» du discours (Vincent, 1993; Traverso, 1999) tels que les ponctuants, les ligateurs, les marqueurs de structuration discursive, etc. qui illustrent des points d'alternance : ils sont très souvent dans une langue alors que le reste de l'énoncé est dans une autre (Matras, 1998). L'exemple suivant illustre ce phénomène avec l'emploi de «bon» et de «quoi» en français dans un discours bilingue kali'na-français.

(14a) conversation informelle entre ABC (corpus Clapoty Lescure_Alby
057 am~~i~~ **côté** molo **bon** palanak~~i~~l~~i~~ am~~i~~-kon tamel~~i~~ tanepo man kali'na wa kote
d'un certain côté, le Blanc a montré son mode de vie au Kali'na

(14b) conversation informelle entre ABC (corpus Clapoty Lescure_Alby)
E: wewe epel~~i~~ ke **soso** molokon **soso** ot~~i~~ **frais** quoi **soso** ot~~i~~
le fruit des arbres avec des choses comme ça, toujours des choses fraîches quoi ! Toujours des choses.

De ce point de vue, nos corpus, dont beaucoup illustrent ces phénomènes, ne sont pas «remarquables» au sens où ils seraient étonnants – ils illustrent des phénomènes attendus et sont donc particulièrement remarquables en terme d'exemplarité. Nous avons développé également une annotation systématique de tous les PREDISC de manière à pouvoir discuter des propositions de la littérature sur la base de la diversité de nos corpus et des informations disponibles pour chacun d'eux.

4.4. Méthode d'analyse des phénomènes remarquables

Une fois les différents phénomènes remarquables identifiés, il s'agit de les décrire, de les analyser puis de tenter de les expliquer. Une méthode d'analyse multi-niveaux a été proposée pour essayer de rendre compte des phénomènes de chacune de ces catégories, PREMS, PREDISC et PRINT. A titre d'exemple, nous citons la démarche élaborée pour les PREMS ; elle s'appuie sur la démarche proposée par Léglise (2012, 2013), qui favorise une analyse séparée des facteurs d'explication (inter et intrasystémiques) avant de montrer leur interaction dans le résultat linguistique observé, en systématisant l'entrée par différents niveaux d'analyse. Les linguistes identifiant des phénomènes remarquables sont invités à se pencher sur un ensemble de niveaux d'analyse :

a) Analyse propre à la langue A : se demander s'il existe d'autres exemples déjà documentés du même phénomène dans la langue A et proposer une analyse – exemple : une variation de ce type a été observée dans la situation X (variation géographique par exemple)

b) Analyse liée à un groupe de langues (au sens de familles linguistiques mais pas exclusivement) : se demander s'il existe d'autres exemples déjà documentés dans des langues proches de la langue A) – exemple : les langues romanes connaissent généralement des réductions de paradigme au niveau des pronoms personnels (nombreux exemples attestés dans telle et telle variété)

c) Analyse liée au contact (en fonction des caractéristiques linguistiques ou typologiques des langues en contact) : se demander si le PREMS peut être lié à une caractéristique de la langue B – par exemple l'ordre des constituants observé ne correspond pas à celui de la langue A dans laquelle l'énoncé est produit mais à celui de la langue B également présente dans la situation de contact

d) Analyse liée à chacune des langues dans d'autres situations de contact : se demander si d'autres exemples sont documentés – et de la langue A – et de la langue B, en contact avec d'autres langues C, D, et produisant le même type d'effets – exemple : la situation de contact actuelle comprend du français et du créole ; le français, en contact avec des langues africaines, produit le même type de phénomènes que ceux observés dans la situation actuelle ; par ailleurs, le créole, en contact avec une autre langue (le néerlandais), produit / ou ne produit pas le même type de phénomènes que ceux observés dans la situation actuelle.

e) Analyse liée au contact indépendante des caractéristiques des langues en contact : vérifier si la littérature fait état de phénomènes identiques dans des situations mettant en scène d'autres langues – par exemple, un schéma de grammaticalisation habituel dans les langues, ou un processus graduel déjà

identifié montrant que la création des articles suit un schéma classique, du numéral vers l'indéfini, puis de l'indéfini vers le défini (Heine et Kuteva 2003, Dryer 2005a, 2005b)

f) Analyse sociolinguistique : décrire la situation de communication en terme d'interlocuteurs (leur âge, leur situation sociale ou professionnelle, leurs rapports), se demander si l'énoncé produit renvoie à une variété stylistique particulière, etc.

g) Analyse pragmatique : si le phénomène inclut un changement de langue, se demander quelle fonction on peut attribuer à ce changement, quel est le thème de l'échange, quel type de séquence (explicitation par exemple) est concerné etc.

L'idée de cette démarche, pas à pas, et contraignante pour le descripteur, est d'étendre la possibilité d'explication des phénomènes qui reste trop souvent confinée au transfert de structures de la langue B vers la langue A (Léglise, 2013). En proposant une analyse multi-niveaux, on fait le pari que plusieurs de ces niveaux sont (la plupart du temps) concernés dans les résultats linguistiques observés. Suivre une telle démarche permet de rendre visibles ces différents niveaux et d'identifier des possibilités d'explication. L'étape suivante consiste à montrer que ces niveaux interagissent et à démonter comment. On sera alors en mesure de proposer des explications plurifactorielles aux phénomènes observés.

4.5. Des métadonnées fines et nombreuses pour une analyse plurifactorielle

Nous avons enrichi chaque corpus transcrit d'un grand nombre de métadonnées qui concernent la situation de contact, les langues et les locuteurs concernés. Elles s'inspirent des facteurs linguistiques et des facteurs sociaux identifiés par la linguistique de contact (en particulier Thomason 2001b, Winford 2003) augmentés de connaissances issues des domaines de la typologie des langues, de l'acquisition/apprentissage des langues, de l'anthropologie linguistique et de la sociolinguistique. Avec une vision maximaliste des possibilités offertes par l'annotation d'informations secondaires sur nos données, nous avons souhaité que ces métadonnées soient les plus riches possibles afin que nous puissions ensuite interroger chacun de ces critères comme un facteur potentiellement pertinent dans la réalisation des phénomènes remarquables observés. Pour ce faire, elles ont été structurées par P. Vaillant dans une base de données qui est renseignée pour chaque texte ou corpus. Nous renvoyons à Vaillant, Léglise & Alby (en préparation) pour des détails sur la conception et l'architecture de la base de données. Nous présentons ci-dessous cinq grandes catégories de métadonnées que nous renseignons.

Premièrement, nous avons voulu catégoriser chacun de nos corpus selon trois typologies majeures de la linguistique de contact. En fonction des critères donnés par chacun des auteurs suivants, nous avons tenté d'inscrire nos corpus dans les typologies concernées :

- La typologie de Winford (2003) sur les situations de contacts distingue entre des situations de contacts marginales (voyages, explorations, conquêtes, médias, apprentissage de langues étrangères, etc.), des situations où les locuteurs évoluent

dans la même communauté mais avec un contact entre un groupe dominant et un groupe minoritaire (immigration, invasion, conquête militaire, modifications des frontières des états, contacts intergroupaux liés à du commerce, des mariages, etc.), et des situations de bilinguisme plus «égalitaires». Selon le degré du contact, Winford cherche à évaluer les effets sur les langues pouvant aller d'emprunts lexicaux uniquement jusqu'à des emprunts structuraux massifs ayant des effets sur la typologie des langues.

- La typologie des interactions verbales de de Pietro (1988) propose d'identifier différents types de situations d'interactions entre des bilingues ou des monolingues, entre des locuteurs natifs ou non, partageant ou non les mêmes langues. Il propose un axe unilingue-bilingue et un axe endolingue-exolingue définissant ainsi quatre cas de figures. Les observables linguistiques identifiés dans les interactions sont ainsi explicités en fonction de la situation de communication verbale telle qu'elle a été définie dans la typologie.

- La typologie de Auer (1999) sur les discours bilingues distingue entre l'alternance conversationnelle (ou *codeswitching*) «pour les cas où la juxtaposition des deux codes est perçue et interprétée comme localement significative par les participants», le mélange de langues (ou *language mixing*) «où c'est la juxtaposition des deux langues en elles-mêmes qui est significative pour les participants, non pas localement (contextuellement) mais dans le fait même d'employer ce type de discours», et enfin la fusion de lectes (ou *fused lects*) pour les cas correspondant à des variétés mixtes stabilisées «où les locuteurs n'ont plus conscience de la mixité de leur discours», où la mixité est constitutive de la langue ainsi créée (Alby et Migge, 2007 : 52).

L'objectif de la catégorisation de nos corpus selon ces trois typologies est de vérifier si les effets et phénomènes linguistiques attendus dans certaines de ces situations correspondent bien à ceux que nous observons dans nos corpus, il s'agit donc en quelque sorte de «tester» ces typologies et de vérifier si elles permettent d'expliciter les phénomènes observés. Le cas échéant nous serons peut-être à même de compléter ces typologies.

Deuxièmement, en ce qui concerne les différentes langues présentes dans nos enregistrements, il nous a paru important de noter quelles relations ces différentes langues entretiennent d'un point de vue génétique ou typologique : sont-elles apparentées ? (même famille linguistique ?, intercompréhension relative ?, variétés stylistiques ou dialectales de la même langue ? etc.), peut-on considérer – et selon quel critère – qu'elles sont typologiquement «proches» ou éloignées ? La question de la distance typologique entre les langues n'est pas triviale et peut-être abordée de différentes manières, notamment en fonction d'une distance objective – que certains tentent de mesurer et que nous nous contentons de noter localement dans le cadre de sous-systèmes linguistiques ou domaines particuliers – ou d'une distance subjective ou perçue (Kellerman & Sharwood Smith, 1986, Giacalone Ramat, 1994) particulièrement importante en situation d'acquisition des langues et que nous prenons également en compte dans nos analyses (cf. idéologie linguistique mentionnées plus bas). Bien que complexe, c'est une question essentielle dans le domaine du contact. Thomason (2010 : 40) insiste sur l'importance de connaître le degré de distance typologique entre des sous-

systèmes (ou domaines) particuliers des langues en contact car cela aide à prédire le type d'interférence (nous dirions ici de phénomène remarquable) qui peut se produire –en fonction également de l'intensité du contact. Lorsque la distance typologique est étroite, des sous-systèmes - pour lesquels on observe rarement de changement induit par contact – peuvent être affectés par le contact. Thomason donne le cas de la morphologie inflexionnelle qui est habituellement peu touchée par le contact. Une distance typologique «minimale» est responsable de la fréquence d'interférences interdialectales impliquant des traits inflexionnels rarement transférés dans le cas de langues plus distantes. Ce n'est, selon l'auteur, pas l'effet déclencheur mais un facteur explicatif important.

Voici quelles métadonnées nous renseignons concrètement pour les relations génétiques ou typologiques : pour un corpus illustrant des contacts entre le kali'na (langue amérindienne de la famille caribe) et le français (langue romane), nous notons qu'il s'agit de langues typologiquement éloignées par exemple du point de vue de l'ordre des constituants dans la phrase – et plus particulièrement de l'ordre dans le groupe nominal. Pour un corpus illustrant des contacts entre le pamaka (créole à base anglaise) et l'aluku (créole à base anglaise), nous considérons qu'il s'agit de variétés dialectales de la même langue (le nengee) et pour le contact entre le pamaka (créole à base anglaise) et l'anglais (langue germanique), on considère qu'au niveau génétique (cet adjectif étant pris ici au sens large), il s'agit du contact entre une langue créole et sa langue lexificatrice mais qu'au niveau typologique, ces langues ont des caractéristiques relativement éloignées du point de vue de l'expression des marques de TAM par exemple. Ces informations nous semblent importantes pour vérifier si les effets observés du contact peuvent être liés à des apparentements et ressemblances génétiques ou typologiques.

Troisièmement, la littérature sur le contact insiste sur la durée et la stabilité du contact entre les langues comme un critère important intervenant dans les résultats de ce contact (Thomason 2001a, Winford 2003). Ce sont généralement les données sociales ou sociolinguistiques intégrées dans les études sur le contact, les «social factors» du «scenario de contact» pertinents, qui ont un pouvoir explicatif. Nous avons décidé de préciser ces éléments pour chacune des paires ou trio de langues présents dans nos corpus – nous réalisons cette annotation à deux niveaux : au niveau généralement considéré dans la littérature, qui est celui de la «communauté linguistique», et au niveau qui nous paraît également pertinent pour expliquer les phénomènes, celui du locuteur et de sa famille.

Quatrièmement, notre connaissance des terrains et des travaux en acquisition et anthropologie linguistique nous ont fait préciser un certain nombre de données secondaires qui nous semblent pouvoir jouer un rôle explicatif important dans les résultats du contact – rôle que nous souhaitons en tout cas tester. Le lieu et le moyen d'acquisition des langues par les locuteurs nous semblent des données importantes et nous avons souhaité les noter systématiquement : telle langue a-t-elle été transmise en famille lorsque le locuteur était enfant, est-ce la (ou l'une des) langue(s) de socialisation majoritaire pour lui, a-t-il appris cette langue à l'école ou dans un contexte formel comparable, a-t-il acquis cette langue dans des

contextes informels (avec des pairs) ou dans l'espace public, ou encore est-on dans un cas de rupture de transmission intergénérationnelle ?

Cinquièmement, le statut des différentes langues dans la situation de communication correspondant au corpus est également un élément à prendre en compte – et nous souhaitons également tester le rôle que ces éléments peuvent jouer : quelles sont les fonctions jouées par les différentes langues sur le territoire concerné ? Quels sont leurs statuts (de jure et de dicto) respectifs ? Quels sont les équilibres numériques en présence (langue majoritaire ou minoritaire numériquement parlant dans la micro-situation concernée, dans la ville où l'enregistrement est réalisé, sur le territoire global) ? Quels sont les rapports idéologiquement parlant entre les langues : au niveau du territoire, la langue A est-elle idéologiquement minoritaire ou dévalorisée ?, au niveau de la région ou de la ville concernée, la langue A et la langue B sont-elles également valorisées ? au niveau de la micro-situation considérée, la langue B est-elle considérée comme appropriée à la situation / valorisée, à la différence de la langue A par exemple ?

Toutes ces questions nous semblent pertinentes, nous les considérons comme autant de données secondaires intéressantes à noter – et à interroger ensuite, pour valider ou invalider leur rôle dans les résultats linguistiques observés, voire, si leur rôle s'avère montré, permettre de mieux expliquer ces résultats.

CONCLUSION

Nos corpus et méthodes permettent de travailler sur des données hétérogènes, qu'elles soient plurilingues, pluri-dialectales, ou pluri-stylistiques ou qu'il s'agisse de variations observées dans ce que l'on considère habituellement comme des productions monolingues. La méthode d'annotation du corpus que nous avons mise en place est un révélateur d'hétérogénéité car la démarche pas à pas oblige le linguiste à se poser des questions qu'il ne se posait pas forcément lors de la transcription, elle oblige également à ouvrir l'univers des possibles, à chaque instant en se demandant si une transcription alternative est possible et si l'élément ainsi noté pourrait appartenir à d'autres langues que celle qui vient spontanément à l'esprit du transcripteur.

De la même manière, la méthode d'analyse des phénomènes remarquables et le renseignement des données sociales obligent également le linguiste, par une démarche pas à pas, à s'intéresser à ses données en ayant en tête un ensemble ouvert de possibilités à aller chercher et renseigner. Seul cet esprit d'ouverture est garant de possibles analyses multiniveaux et explications plurifactorielles. Le parti pris résolument choisi est celui d'analyses à un niveau «micro du micro» - tant au niveau des données linguistiques que des données sociales - qui nécessitent un travail de fourmi sur les enregistrements et dans les analyses, nous croyons, à la suite de Léglise (1999) que c'est à ce prix que l'on peut trouver des régularités (en particulier statistiques) et des explications aux phénomènes observés.

Le projet Clapoty constitue une fabuleuse aventure humaine. En croisant des méthodes et des points de vue, issus de plusieurs traditions en sciences du

langage, il adopte de fait deux approches, l'une inductive, l'autre déductive. Il cherche à la fois à ouvrir l'éventail des possibilités d'explication par une analyse manuelle et complexe des phénomènes repérés et d'autre part à tester des hypothèses par des vérifications informatiques à partir des bases de données créées à partir des centaines d'annotations manuelles réalisées.

BIBLIOGRAPHIE

Alby S., 2001, *Contacts de langues en Guyane française : une description du parler bilingue kali'na-français*, Thèse de doctorat sous la direction de J-C. Pochard, Université Lumière Lyon II, Lyon.

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

Alfonzetti G., 1998, The conversational dimension in codeswitching between italian and dialect in Sicily, in P. Auer (ed), *Codeswitching in conversation*, Londres, Routledge, p. 180-214.

Alvarez Caccamo C., 1990, Rethinking conversational code-switching: codes, speech varieties and contextualisation, Communication au colloque *Proceedings of the sixteenth annual meeting of the Berkeley Linguistics Society*, 16-19 février, Berkeley.

Auer P., 1995, The pragmatics of code-switching: a sequential approach, in L. Milroy & P. Muysken (eds), *One speaker, two languages: cross disciplinary perspectives on code-switching*, Cambridge, Cambrige University Press, p. 115-135.

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, p. 309-332.

Backus A. 2003. Units in codeswitching: evidence for multimorphemic elements in the lexicon. *Linguistics*, *41*(1), p. 83-132

Belazi H. M., Rubin E. & Toribio A. J., 1994, Code Switching and X-Bar theory: the functional head constraint, *Linguistic Inquiry* 25-2, p.221-237.

Bentahila A. & Davies E. E., 1983, The syntax of Arabic-French code-switching, *Lingua*, 59, p. 301-330.

Billiez J. (éd), 2003, *Contacts de langues. Modèles, typologies, interventions*, Paris, L'Harmattan.

Bokamba E. G., 1989, Are there syntactic constraints on Code-Mixing ?, *World Englishes* 8, p. 277-293.

Boyer H., 1997, Conflits d'usages, conflits d'images, in H. Boyer (éd), *Plurilinguisme : «contact» ou «conflit» de langues*, Paris, L'Harmattan, p. 9-35.

Canut C., 2001, A la frontière des langues. Figures de la démarcation, *Cahiers d'Etudes Africaines* 163-164, p. 443-463.

Canut C. & Caubet D. (éds), 2002, *Comment les langues se mélangent. Codeswitching en francophonie*, Paris, L'Harmattan.

Chamoreau C., 1995, La comparaison en purepecha. Un exemple d'évolution syntaxique, *Faits de Langues* 5, p. 140-143.

Chamoreau C., 2012a, Constructions périphrastiques du passif en purepecha. Une explication multifactorielle du changement linguistique, in C. Chamoreau & L. Goury (éds), *Changement linguistique et langues en contact. Approches plurielles du domaine prédicatif*, Paris, CNRS Editions, p. 251-270.
Chamoreau C., 2012b, Développement de l'article indéfini *ma* en purepecha, Communication au séminaire de la Fédération "Typologie et Universaux Linguistiques", *Evolution des structures morphosyntaxiques. Vers une typologie intégrative*, 10 mai 2012.
Chamoreau C. & Lastra Y. (ed), 2005, *Dynamica linguistica de las lenguas en contacto*, Sonora, Universidad de Hermosillo.
Chamoreau C. & Goury, L. (éds), 2012, *Changement linguistique et langues en contact. Approches plurielles du domaine prédicatif*, Paris, CNRS Editions.
Chamoreau C. & Léglise, I., 2012, A multi-model approach to contact-induced language change, in C. Chamoreau & I. Léglise (éds) *Dynamics of contact-induced language change*, Mouton de Gruyter, p. 1-15.
Croft W., 2000, *Explaining language change*, Harlow, Pearson Education Limited.
Dahl Ö., 2007. From questionnaires to parallel corpora in typology, *Sprachtypologie und Universalienforschung* 60 (2), p. 172-81.
Déjean H., Gaussier E. & Sadat F., 2002, Bilingual terminology extraction: an approach based on a multilingual thesaurus applicable to comparable corpora, *Proceedings of COLING' 2002*, Taipei, Japon.
De Pietro J.-F., 1988, Vers une typologie des situations de contacts linguistiques, *Langage et Sociétés* 43, p. 65-89.
Deprez C., 1994, *Les enfants bilingues : langues et familles*, Paris, Didier.
Dryer M. S., 2005a, Definite articles, in M. S. Dryer & M. Haspelmath (eds), *The world atlas of language structures*, **LIEU, EDITEUR?,** p. 154-155.
Dryer M.S., 2005b, Indefinite articles, in M. S. Dryer & M. Haspelmath (eds), *The world atlas of language structures*, **LIEU, EDITEUR?,** p. 158-159.
Fishman, J. 1972, Domains and the relationship between micro and macrosociolinguistics, in J. J. Gumperz & D. Hymes, Directions in sociolinguistics. The ethnography of communication, New York, Holt, Rinehart & Winston, p. 435-453.
Field F. W., 2002, *Linguistic borrowing in bilingual contexts*, Amsterdam/Philadelphia, John Benjamins Publishing Company.
Giacalone Ramat A., 1994, Il ruolo della tipologia linguistica nell'acquisizione di lingue seconde, in A. Giacalone Ramat & M. Vedovelli (eds), *Italiano lingua seconda/lingua straniera.* Atti del XXVI Congresso della Società di Linguistica Italiana, Roma, Bulzoni, p. 27-43.
Givon T., 1981, On the development of the numeral 'one' as an indefinite marker, *Folia Linguistica Historica* 2-1, p. 35-53.
Goebl H., Nelde P. H., Stary Z. & Wölck W. (eds), 1996, *Contact linguistics. An international handbook of contemporary research, vol.1*, Berlin/New York, De Gruyter.
Gumperz J. J., 1976, The sociolinguistic significance of conversational Code-Switching, *Papers on Language and Context: Working Papers* 46, p. 1-46.
Groupe ICOR, 2007, Variations interactionnelles et changement catégoriel : l'exemple de '*attends*', in **AUTEUR(S)?,** *La mise en œuvre des langues dans l'interaction*, Paris, L'Harmattan, p. 299-320.

Heiden S., 2006, Un modèle de données pour la textométrie : contribution à une interopérabilité entre outils in J-M. Viprey et al. *Archives, Bases, Corpus*, vol 1, Presses Universitaires de Franche-Comté, Besançon, p. 747-487.
Heine B., 1997, *Cognitive foundations of grammar*, Oxford, Oxford University Press.
Heine B. & Kuteva T., 2003, On contact-induced grammaticalization, *Studies in Language* 27-3, p. 529-572.
Heine B. & Kuteva T., 2005, *Language contact and grammatical change*, Cambridge, Cambridge University Press.
Heine B. & Kuteva T., 2008, Constraints on contact-induced linguistic change, *Journal of Language Contact,* Thema 2, p. 57-90.
ICOR, 2007, *Conventions de transcription. Accessible en ligne*, http://icar.univ-lyon2.fr/projets/corinte/bandeau_droit/convention_icor.htm (consulté le 06/12/2012).
Juillard C., 1995, *Sociolinguistique urbaine : la vie des langues à Zinguichor*, Paris, CNRS Editions.
Kellerman E. & Sharwood Smith M. (eds), 1986, *Crosslinguistic influence in Second Language Acquisition.* New York, Pergamon Press.
Kriegel S. (ed), 2003, *Grammaticalisation et réanalyse. Approche de la variation créole et française*, Paris, CNRS Editions.
Ledegen G., 2012, Prédicats «flottants» entre le créole acrolectal et le français à la Réunion : exploration d'une zone ambiguë, in C. Chamoreau, L. Goury (ed), *Changement linguistique et langues en contact. Approches plurielles du domaine prédicatif*, Paris, CNRS Editions, p. 251-270.
Léglise I., 1999, *Contraintes de l'activité de travail et contraintes sémantiques sur l'apparition des unités et l'interprétation des situations*, Thèse de doctorat, Université Paris 7-Denis Diderot.
Léglise I., 2007a., Explaining language contact phenomena in a prospective diachronic perspective: discussion of a methodological frame, *Language Contact Symposium*, Max Planck Institut, Leipzig, May 10-13.
Léglise I., 2007b, Des langues des domaines, des régions. Pratiques, variations, attitudes linguistiques en Guyane, in I. Léglise & B. Migge (éds), *Pratiques et représentations linguistiques en Guyane : regards croisés*, Paris, IRD Editions, p. 29-47.
Léglise I., 2009, *Contacts de langues : analyses plurifactorielles assistées par ordinateurs et conséquences typologiques*, Projet de recherche soumis à l'ANR.
Léglise I., 2012, Variations autour du verbe et de ses pronoms objets en français parlé en Guyane : rôle du contact de langues et de la variation intrasystémique, in C. Chamoreau et L. Goury (éds) *Changement linguistique et langues en contact*, CNRS editions, p. 203-230.
Léglise I., 2013, 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, (eds), *The interplay of variation and change in contact settings – Morphosyntactic studies,* John Benjamins, p. 137-163.
Léglise I. & Chamoreau C., 2013, The interplay of variation and change in contact settings, in I. Léglise & C. Chamoreau, (eds), *The interplay of variation and change in contact settings* John Benjamins, p. 1-20.
Léglise I. & Migge B., 2005, Pour une étude des contacts de langues en synchronie : quelques exemples tirés du cas guyanais, *TRACE* 47, p. 113-131.

Mackey W. F., 1976, *Bilinguisme et contact des langues*, Paris, Klincksieck.
MacSwan J., 1997, *A minimalist approach to intrasentential code switching: Spanish-Nahuatl bilingualism in Central Mexico*, University of California, Los Angeles.
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.
Mahootian S., 1993, *A null theory of codeswitching*, thèse de doctorat, Northwestern University.
Matras Y., 1998, Utterance modifiers and universals of grammatical borrowing, *Linguistics*, 36-2, p. 281-331.
Matras, Y., 2009, *Language contact*, Cambridge, Cambridge University Press.
Matras Y. & Sakel J. (eds), 2007, *Grammatical Borrowing in Cross-Linguistic Perspective*, Berlin, Walter de Gruyter.
Matras Y., White C. & Elšík V. à paraître, The Romani Morpho-Syntax (RMS) Database, in M. Everaert & S. Musgrave (eds). *Linguistic databases*. Berlin, Mouton de Gruyter.
Matthey M. & De Pietro J.-F., 1997, Utopie souhaitable ou domination acceptée ?, in H. Boyer (éd), *Plurilinguisme : 'contact' ou 'conflit' de langues*, Paris, L'Harmattan, p. 133-190.
Botley S. P., McEnery A. M. & Wilson A. (eds), 2000, *Multilingual corpora in teaching and research*, Amsterdam, Rodopi.
Manfredi S., Simeone-Senelle M. C. & Tosco M., sous presse, Codeswitching and borrowing in CorpAfroAs, in Mettouchi A. et al. *CorpAfroAs: A Corpus for Afro-Asiatic Languages*, Amsterdam, Benjamins.
Mettouchi A. & Chanard C., 2010, From fieldwork to annotated corpora: the CorpAfroAs Project, *Faits de Langue – Les Cahiers* 2, p. 255-265.
Migge B. & 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 (eds), *Variation in the Caribbean*, Amsterdam, John Benjamins, p. 181-199.
Migge B. & Léglise I., 2013, *Exploring Language in a Multilingual Context: Variation, Interaction and Ideology in language documentation*, 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.
Myers-Scotton C., 2002, Contact Linguistics: Bilingual Encounters and Grammatical Outcomes. Oxford: Oxford University Press.
Muysken P., 1995, Code-switching and grammatical theory, in L. Milroy & P. Muysken (eds), *One speaker, two languages: cross-disciplinary perspectives on code-switching*, Cambridge, Cambridge University Press, p. 177-198.
Muysken P., 2011, Codeswitching, in R. Mesthrie (ed.), The Cambridge Handbook of sociolinguistics, Cambridge University Press, 301-314.
Nicolaï R., 2005, Language processes, theory and description of language change, and building on the past: lessons from Songhay, in Z. Frajzyngier, A. Hodges & D. S. Rood (eds), *Linguistic diversity and language theories*, Amsterdam/Philadelphia, John Benjamins, p. 81-104.

Nicolaï R., 2007, Le contact des langues, point aveugle du 'linguistique', *Journal of Language and Contact* 1, p. 1-21.
Nishimura M., 1985, *Intrasentential Code-Switching in Japanese and English*, Thèse de doctorat, Université de Pennsylvania.
Poplack S., 1980, Sometimes I'll start a sentence in Spanish y termino en Espanol, *Linguistics*, 18, p.581-618.
Poplack S., 1981, The syntactic structure and social function of code-switching, in R. P. Duran (ed), *Latino language and communicative behavior*, Norwood, New Jersey, Ablex, p. 169-184.
Ross M., 1999, Exploring metatypy: how does contact-induced typological change come about?, Communication à *Australian Linguistic Society's annual meeting*, Perth (http://rspas.anu.edu.au/linguistics/mdr/Metatypy.pdf).
Ross, M., 2001, Contact-induced change in Oceanic languages in North-West Melanesia, in A. Aikhenvald & R. Dixon (eds), *Areal diffusion and genetic inheritance*, Oxford, Oxford University Press, p. 134-166.
Ross, M., 2007, Calquing and metatypy, *Journal of Language Contact* (http://cgi.server.unifrankfurt.de/fb09/ifas/JLCCMS/issues/THEMA_1/JLC_THEMA_1_2007_06R_oss.pdf)
Sacks H., Schegloff E. & Jefferson G., 1974, A simplest systematics for the organisation of turn-taking for conversation, *Language*, 50-4, p. 696-735.
Stolz T., 2007, Harry Potter meets *Le Petit Prince* – On the usefulness of parallel corpora in crosslinguistic investigations, *Sprachtypologie und Universalienforschung* 60 (2), p. 100-117.
Thomason S., 1993, On identifying the sources of creole structures, in S. Mufwene (ed), *Africanisms in Afro-American language varieties*, Athens, GA, University of Georgia Press, p. 280-295.
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 & W. Raible (eds), *Language typology and language universals, Sprachtypologie und sprachliche universalien, vol.*2, Berlin/New York, Walter de Gruyter, p. 1640-1648.
Thomason S., 2010, Contact Explanations in Linguistics, in R. Hickey **(ED?)**, *The Handbook of Language Contact,* Wiley-Blackwell, p.31-47
Thomason S. & Kaufman T., 1988, *Language contact, creolization, and genetic linguistics*, Oxford/Berkeley, University of California Press.
Traverso V., 1999, *L'Analyse des conversations*, Paris, Nathan.
Vaillant, P., Léglise I. & Alby S., en préparation, *Le schéma de document Corpus Contact*.
Véronis J. (ed) 2000, *Parallel Text Processing*. Alignment and Use of Translation Corpora, Kluwer Academic Publishers.
Véronis J. (ed), 2002, Alignement lexical dans les corpus multilingues, *Lexicometrica* (http://lexicometrica.univ-paris3.fr/thema/thema6.htm).
Vion R., 1992, *La communication verbale. Analyse des interactions*, Paris, Hachette.
Vincent D., 1993, *Les ponctuants de la langue et autres mots du discours*, Québec, Nuits Blanches.
Weinreich U., 1953, *Languages in contact: findings and problems*, New York, The Linguistic Circle of New York.
Winford D., 1997, Creoles in the context of contact linguistics, *Journal of Pidgin and Creole Languages* 12, p. 131-151.

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.
Zweingebaum P., Rapp R., Sharoff S. (ed), 2011, *Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web*, Association for Computational Linguistics, Portland.