À la croisée des langues. Annotation et fouille de corpus plurilingues

Pascal Vaillant, Isabelle Léglise

À la croisée des langues

Annotation et fouille de corpus plurilingues

Pascal Vaillant* et Isabelle Léglise**

*Université Paris 13, Sorbonne Paris Cité, LIMICS, (UMRS 1142),
74 rue Marcel Cachin, 93017, Bobigny cedex, France
INSERM, U1142, LIMICS, 75006, Paris, France
Sorbonne Universités, UPMC Univ Paris 06, UMRS 1142, LIMICS, 75006, Paris, France
vaillant@univ-paris13.fr
**CNRS, Structure et Dynamique des Langues (SeDyL), (UMR 8202),
7 rue Guy Môquet, 94800, Villejuif, France
leglise@vjf.cnrs.fr

Résumé. Un programme de recherche en cours sur l'étude des phénomènes de contact de langues et de leur rôle dans le changement linguistique s'attache à recueillir des corpus plurilingues, témoignant d'une grande variété de phénomènes de contact sur un échantillon suffisamment varié de langues génétiquement et typologiquement distinctes. Cet effort a impliqué le développement d'une chaîne de traitement des corpus numériques qui tienne compte des spécificités des corpus plurilingues, pour la représentation des données linguistiques, leur stockage, leur annotation, leur visualisation, et les traitements de recherche d'information. Les normes existantes ont dû être étendues pour prendre en compte l'appartenance potentielle d'unités à plusieurs langues dans les pratiques langagiéres plurilingues. Dans cet article, nous décrivons la manière dont a été définie la structure de ces corpus plurilingues, et la conception technique de l'unité linguistique multilingue qui préside à la fouille de données dans ces corpus.

1 Introduction

Le contact de langues est l'une des forces motrices du changement linguistique. Cette assertion, évidente lorsque l'on pense au yiddish ou aux langues créoles, est également un postulat bien connu des historiens de la langue qui ont étudié, par exemple, le passage du latin aux langues romanes, ou l'émergence de l'anglais moderne. À l'origine de ces changements, il y a nécessairement l'interaction entre des individus aux répertoires linguistiques plurilingues (Gumperz, 1982) qui, en alternant et mélangeant les langues produisent toutes sortes de variations dans l'une ou l'autre des langues et des pratiques langagières plurilingues décrites dans la littérature comme *codeswitching*, *code-mixing* et *fused lects* (Auer, 1999), *polylanguaging* (Jørgensen et al., 2011), pratiques langagières hétérogènes (Léglise, 2012). Cette multitude d'actions individuelles (Matras, 2009) prend place dans des situations sociales multilingues dans lesquelles ces variations et innovations se propagent pour progressivement mener au changement (Léglise et Chamoreau, 2013). Ainsi, les situations de multilinguisme, impliquant des

individus plurilingues, sont ce qu'il y a de plus fréquent ; dans la plupart des régions et à la plupart des époques, elles constituent la norme plutôt que l'exception (Wurm, 1996).

Malgré cette continuité logiquement nécessaire entre le niveau microscopique et le niveau macroscopique, les sciences du langage ont jusqu'à présent le plus grand mal à se focaliser sur les étapes de transition et à tenter de les décrire. La linguistique descriptive tend à considérer chaque langue comme un système clos. La linguistique historique, qui ne peut ignorer les conséquences des phénomènes de contact, les a longtemps mentionnés sous des termes macroscopiques comme *substrat*, *superstrat* ou *adstrat* ; plus récemment, des chercheurs comme Thomason et Kaufmann (1988), Thomason (2001), Heine et Kuteva (2005, 2007), Aikhenvald et Dixon (2006), Peyraube (2002) développent des hypothèses fondées sur le contact de langues pour expliquer des évolutions touchant des sous-systèmes spécifiques de certaines langues ; cette prise en compte du contact reste toutefois inscrite dans la démarche portée par la linguistique historique, à savoir celle de l'hypothèse de reconstitution à partir du résultat du processus : les étapes intermédiaires ne sont en effet en général pas documentées (Winford, 2003; Léglise et Migge, 2006).

À l'autre bout de l'échelle, la linguistique interactionniste et la sociolinguistique enregistrent des pratiques langagières illustrant ces contacts linguistiques vivants, dans leurs manifestations concrètes, et s'intéressent aux paramètres liés directement à l'interaction : histoire individuelle ou collective des locuteurs, fonctions du mélange ou de l'alternance de langues dans l'interaction, regard de la société sur ces formes de parole etc. Dans le champ francophone, on peut notamment citer les travaux de Lüdi et Py (1986), Deprez (1994), Juillard (1995), Tabouret-Keller (2001).

Jusqu'aujourd'hui, il est difficile aux sciences du langage, faute de matériau, de tenir à la fois les deux extrémités de la corde, c'est-à-dire d'observer *in vivo* le changement linguistique en cours. C'est l'un des objectifs que s'est fixé le programme de recherche CLAPOTY [1] : recueillir, annoter et étudier des corpus numérisés dans lesquels se manifestent des phénomènes de contact de langues, sans se laisser dérouter à l'avance par ce que certains pourraient considérer comme un chaos résultant du mélange de plusieurs « systèmes » linguistiques. Ces corpus sont donc plurilingues et non multilingues — au sens de la linguistique de corpus [2].

Nos corpus plurilingues comprennent donc des interactions spontanées plurilingues illustrant des phénomènes de *codeswitching* ou de mélange entre plusieurs langues. Ces corpus plurilingues, sont encore peu nombreux, peu disponibles à la communauté des chercheurs, et peu « outillés » du point de vue des traitements informatisés disponibles. On peut citer la base ICOR de la plateforme CLAPI [3] qui comporte quelques données plurilingues, le projet LIPPS/LIDES [4] dont l'objectif était de développer des standards de transcription pour les langues mixtes et le *codeswitching* ou la base Bilingbank accesssible sous Talkbank [5].

1. Projet CLAPOTY (*Contacts de Langues : Analyses Plurifactorielles assistées par Ordinateur et conséquences Typologiques*, projet ANR-09-JCJC-0121-01) dirigé par I. Léglise (http://clapoty.vjf.cnrs.fr). P. Vaillant est responsable des tâches informatiques et créateur du schéma de documents au sein du projet.

2. On entend généralement par *corpus multilingues* des corpus comprenant des textes dans différentes langues, ces textes étant *a priori* chacun monolingue (voir (Schmidt et Wörner, 2012) pour un état des lieux des travaux actuels).

3. Cf. http://clapi.univ-lyon2.fr et http://icar.univ-lyon2.fr/projets/corinte/.

4. Cf. http://www.ling.lancs.ac.uk/staff/ruthanna/lipps/lipps.htm.

5. Cf. http://talkbank.org.

2 Une position de neutralité idéologique

Les phénomènes par lesquels se manifeste le contact de langues sont connus, et font l'objet d'une littérature relativement abondante (un panorama en est fourni par (Thomason, 2001) ou (Winford, 2003)). L'un de ceux dont l'observation est la plus banale est l'utilisation d'un mot d'une langue B au sein d'un discours dans une langue A (nous le décrivons ici dans des termes volontairement non scientifiques). Ce phénomène ne se limite bien entendu pas nécessairement à un seul mot : il peut impliquer des expressions, un énoncé ou une prise de parole entière au sein d'une interaction. Il peut impliquer une expression discontinue. Il concerne parfois des unités linguistiques inférieures au niveau du mot, comme des morphèmes grammaticaux (ex. marques de conjugaison d'une langue A affixées à des mots d'une langue B). Par ailleurs, on observe parfois, sans que soit utilisé de matériau phonétique propre à une langue A, qu'une intonation, des valeurs, ou des procédés de formation de termes ou de syntagmes typiques de la langue A sont utilisés dans la langue B.

De nombreux termes techniques ont été créés au cours de ces deux derniers siècles. Par exemple, on fait classiquement référence aux phénomènes mentionnés ci-dessus, respectivement, par les termes d'*emprunt*, d'*alternance de langues* (renvoyant à des phénomènes de *codeswitching* ou *code-mixing*), d'*intégration morphologique*, de *calque*. Le problème que nous rencontrons, si nous essayons d'utiliser ces termes, est que leur définition se fonde plus fréquemment sur des exemples stéréotypiques que sur un faisceau de critères définitoires qui permette une caractérisation sans ambiguïté. Ainsi, on utilise généralement *emprunt* pour un mot étranger d'usage courant, et la plupart des auteurs qui parlent d'*alternance* donnent des exemples qui portent sur plusieurs éléments ou sur un énoncé. La question de choisir de parler d'*alternance limitée à un seul mot* ou d'*emprunt à usage unique* fournit par exemple matière à de nombreux débats entre spécialistes (Winford, 2003).

Un problème plus général se pose avec l'utilisation de termes techniques pré-existants — même ceux à qui certains auteurs ont tenté de donner une caractérisation rigoureuse : leur définition fait souvent appel à des notions de base postulées (comme celle de la *langue matrice* d'une prise de parole, qu'utilise Myers-Scotton (2002), qui est discutée dans la littérature — cf. notamment (Auer et Muhamedova, 2005)), que nous ne souhaitons pas adopter sans examen.

Notre projet a en effet nécessité de construire un corpus représentatif d'une variété suffisante de phénomènes de contact, avec une diversité suffisante de langues représentées pour ne pas courir le risque d'un biais de représentation, et de ne tirer d'éventuelles conclusions sur la structure de ces phénomènes que des observations que nous ferons sur le corpus (cf. (Léglise et Alby, 2013) à propos du corpus recueilli). C'est donc un projet empirique : en cela, il exclut d'utiliser, dans les pré-traitements que nous faisons subir aux données (c'est-à-dire lors de l'annotation des corpus), des éléments de description qui drainent avec eux des postulats théoriques que nous cherchons justement à remettre en question ou à redémontrer — sous peine de retrouver, en sortie de l'analyse, ce que nous avons introduit en entrée.

Notre problématique est donc la suivante : nous devons tout à la fois (1) noter et annoter les phénomènes dans les corpus, pour pouvoir les soumettre à l'analyse ; et (2) ne pas présupposer leur définition. Ceci nous a conduit à inventer un schéma d'annotation qui cherche à la fois à être exhaustif quant à la variété des phénomènes observés, et extrêmement terre-à-terre quant à la manière de les décrire. Les choix de description effectués sont décrits ci-dessous (sections 4 et 5).

3 Difficultés d'annotation de pratiques langagières plurilingues

Pour illustrer le type de manifestation linguistique que nous avons à annoter, prenons l'extrait suivant de l'un des textes qui composent le corpus CLAPOTY [6] :

(1) Corpus Clapoty — Nelson / Léglise : *EDF*

(1.1)	**Yèr**	**mo**	**té**	**pasé**	**la**				
	hier	1SG	PST	passer	là				
	Hier je suis passé ici								
(1.2)	**i**	**té**	**gen**	***an***	madame	un peu	costaud	à côté	là
	3SG	PST	avoir	INDF	dame	un peu	costaud	à côté	là
	il y avait une dame un peu forte, à côté, là								
(1.3)	**i**	m'	a	donné	[. . .]	comme	***té***	***ni***	***problem***
	3SG	1SG	avoir	donner		comme	PST	avoir	problème
	elle m'a donné [. . .] *comme il y avait un problème*								

Cette prise de parole a été enregistrée à Cayenne, en Guyane Française — région multilingue, dans un contexte où prédominent le français et le créole guyanais de Cayenne (langue créole à base lexicale française), mais où se fait également sentir une forte influence du créole à base française des Petites Antilles (notamment la variante martiniquaise / saint-lucienne). Les conventions de transcription utilisées sont les suivantes : on visualise trois lignes : une ligne de transcription, une ligne de traduction morphème par morphème, et une ligne de traduction libre. Dans la transcription, en première approximation, les unités dont on est sûr qu'elles peuvent être identifiées comme du français sont notées en romain maigre, les passages en créole guyanais sont en romain gras, et les passages en créole antillais en italique gras. Dans la traduction morphème par morphème, des équivalents français sont donnés pour les morphèmes à classe ouverte, et des abréviations de catégories grammaticales sont données pour les morphèmes à classe fermée [7].

3.1. La première question à régler est celle de l'appartenance d'un mot à une langue. L'exemple donné ici illustre une situation où les langues en contact ont déjà un grand stock de vocabulaire commun : c'est le cas, par définition, entre une langue créole et sa langue lexificatrice ; c'est vrai également entre deux langues de la même famille et *a fortiori* entre variantes stylistiques ou dialectales. Dans l'exemple ci-dessus, cela signifie qu'il existe un grand nombre d'unités pour lesquelles il est difficile de déterminer, à défaut d'indices externes clairs (forme phonétique, contexte syntagmatique), s'il s'agit d'unités françaises ou créoles. Dans la transcription ci-dessus par exemple, **yèr** (1.1) a été catégorisé comme créole, mais rien ne distingue extérieurement (à l'oral) ce mot créole du mot français « hier ». Le créole guyanais et le créole des Petites Antilles ont un stock de vocabulaire commun encore plus important, car ils font partie d'un continuum de langues créoles à base française apparentées (Pfänder, 2000, p. 192–199). Par exemple, la marque pré-verbale de passé **té** est identique dans les deux variétés de créole. Le mot **problem**, quant à lui, est identique dans les trois langues. En vérité, seul le mot

6. L'enregistrement et la première transcription du corpus a été réalisé par L. Nelson dans le cadre d'un mémoire de recherche (Nelson, 2008). L'annotation a été réalisée dans le cadre du projet Clapoty.

7. Abréviations utilisées ici : 1SG : première personne du singulier ; INDF : indéfini ; PST : passé.

ni (1.3) connote immédiatement le créole antillais, car l'équivalent normal en créole guyanais est **gen** (1.1)[8]. Pour un nombre non négligeable de mots de cet exemple, donc, la décision de l'attribution à une langue ou à une autre n'est pas évidente.

L'exemple illustre ici un contact entre langues apparentées, et l'on pourrait objecter qu'il n'est peut-être pas représentatif ; or le même type de questions se posent quelles que soient les langues en présence : il n'est jamais trivial de décider si un segment (d'un ou plusieurs mots) fourni par le lexique d'une autre langue doit être catégorisé comme « étranger » ou non. Quels critères doit-on utiliser pour trancher : l'assimilation phonologique, morphologique, syntaxique ? La fréquence ? L'ancienneté attestée de l'emprunt postulé ? Des expressions d'usage courant dans certaines situations professionnelles fournissent des illustrations quotidiennes de ce phénomène : ainsi, le premier élément de « *design* graphique » est un mot anglais. Faut-il, pour chacune de ses occurrences, l'étiqueter « langue anglaise » ? Outre le fait qu'il est bien souvent prononcé *disagne* par le locuteur francophone, sa subordination au contexte d'un syntagme français, où l'adjectif déterminant est d'une part français, et d'autre part postposé au nom déterminé, incitent à voir dans ce dernier un mot français d'origine étrangère plutôt qu'un basculement temporaire à l'anglais. La question se modifiera sensiblement dans le cas de « *webdesign* », où l'on pourra exposer l'argument inverse ; mais que dire dans ce cas de « *design web* », tout aussi fréquent dans l'usage des agences de communication en France ? Le filon des questions discriminantes est inépuisable. On l'a dit dans l'introduction : des débats théoriques entre experts sont sans fin sur la question de l'opposition définitoire entre *alternance limitée* et *emprunt à usage unique* ; notre position est tout simplement que le fait de vouloir trancher en attribuant un mot à une langue et une seule est non seulement difficile, mais, c'est plus grave, réducteur.

3.2. Une deuxième question non triviale, qui découle de la première, est celle de la transcription. Le morphème « là », que l'on trouve à la fin de (1.1) et de (1.2), est fondamentalement le même, tant sur le plan de la forme phonétique que sur le plan du sens, en français et en créole. Les normes de transcription orthographique usuelles prescrivent qu'on l'écrit « là » en français, et **la** en créole. Ici c'est essentiellement l'environnement syntagmatique qui a présidé au choix d'attribuer à la première occurrence la langue créole, et à la seconde la langue française.

Alors, question accessoire ? Non, car le choix de transcription implique implicitement un choix d'attribution de la langue, et les valeurs grammaticales, sémantiques et pragmatiques associées à une même réalisation phonétique ne sont pas toujours identiques dans une langue et dans une autre. Un simple phonème, selon la manière dont il est réalisé (ou non-réalisé) peut impliquer la manifestation d'une valeur ou d'une autre dans une catégorie grammaticale ou sémantique (genre, nombre ou personne, par exemple), et/ou connoter un positionnement dans un registre discursif (forme d'adresse, registre de langue ...), et ces choix de valeurs ne sont pas identiques selon la langue dans laquelle on les considère.

Ainsi, le mot « madame » peut être, de façon neutre, utilisé avec un article indéfini en créole — alors qu'en français cet usage serait familier ou ironique. De même, le pronom « i(l) » (pronom sujet de troisième personne du singulier masculin) est fréquemment prononcé sans « l » final en français oral : le premier morphème de la ligne (1.3) pourrait donc aussi bien, par hypothèse, être le **i** créole que le « i(l) » français — à ceci près que l'identifier comme « i(l) »

8. Étymologiquement, du créole contemporain aux sources françaises, en passant par les traces attestées en créole archaïque : *ni* < *tini* < *tenir* (Antilles) ; *gen* < *genyen* < *gagner* (Guyane).

suppose en même temps non seulement une élision de la consonne, mais une neutralisation du genre — car on s'attendrait en français écrit à la forme anaphorique au féminin (« elle ») ; cette neutralisation est bien documentée dans les corpus oraux en français (dans l'hexagone comme dans différentes zones géographiques) mais encore non décrite en français parlé en Guyane ; en créole, la distinction de genre n'existe pas et **i** est la forme unique de troisième personne. Faire un choix de transcription implique donc en réalité de forcer l'incorporation d'un certain nombre d'éléments annexes de jugement linguistique non démontrés. Cette question surgit bien plus souvent qu'on ne pourrait l'imaginer.

Face à ce dilemme, la solution de tout transcrire en alphabet phonétique international ne fait que repousser le problème et en créer d'autres : (1) le choix de chaque symbole peut devenir un problème, car la perception auditive elle-même est active (il arrive fréquemment que deux transcripteurs, en toute bonne foi, *entendent* deux sons différents) ; (2) la faisabilité technique de la transcription, déjà fastidieuse, est multipliée ; (3) de même de la lisibilité du résultat ; (4) enfin, plus grave, si cette transcription phonétique ne sert qu'à ne pas décider de l'attribution d'une langue, alors on ne fait que balayer sous le tapis les éléments d'information corrélés aux choix que l'on n'a pas voulu faire. Ces éléments sont pourtant potentiellement importants, et la solution n'est pas de s'en débarrasser.

3.3. La troisième question à trancher est celle de la frontière — sur l'axe syntagmatique — des passages relevant d'une langue et des passages relevant d'une autre. À partir du moment où l'on admet que certains mots ne peuvent être attribués que de manière indécise à une langue ou à une autre, alors il devient impossible de délimiter de manière certaine la frontière d'un segment dans la langue A et d'un segment dans la langue B. Ainsi, en (1.2), le créole cède-t-il la place au français avant le mot « madame », au niveau de ce terme qui servirait ainsi de pivot commun, ou après ?

L'absence de réponse tranchée aux trois questions mentionnées ci-dessus implique qu'il existe des *passages multilingues* — que ce soit au sens paradigmatique, c'est-à-dire des « segments flottants » entre langues (Ledegen, 2012) ; ou au sens syntagmatique, c'est-à-dire des zones de transition — dont il faut tenir compte en tant que tels pour représenter convenablement la réalité des corpus de contacts de langue. Or les schémas de documents existants, notamment ceux inspirés par les directives de la TEI (ci-dessous, § 4.4), ne comportent pas la possibilité de rattacher un segment linguistique à plusieurs langues à la fois. Nous avons donc été amenés à inventer une manière de structurer les données qui le permette. Le formalisme technique de représentation de ces passages multilingues sera décrit plus bas (notamment § 4.4.2).

4 Le schéma de documents « Corpus-Contacts »

4.1 Choix techniques

Les textes recueillis doivent être normalisés pour permettre des traitements communs en termes de fouille de données — indépendamment de la personne qui s'est chargée de la transcription, ou des langues manifestées. Il faut donc recueillir l'ensemble des informations (la transcription elle-même, et l'ensemble des couches d'annotation) dans un format homogène qui permette de structurer et d'étiqueter les annotations.

C'est la problématique usuelle des corpus numériques, au sujet de laquelle se sont accumulées trois décennies d'expérience. Nos choix techniques sont donc en grande partie conformes aux normes actuelles, qui ont cherché à apporter une réponse à ces questions :

– le système d'encodage de caractères *Unicode* (Allen, 2012) permet de représenter les caractères de tous les systèmes d'écriture actuellement en usage, l'A.P.I. inclus ;
– le méta-langage d'annotation *XML* (Bray et al., 2008) fournit un cadre général de description de l'information structurée, adaptable aux besoins de notre tâche de structuration de corpus ;
– les normes proposées par le consortium *TEI* pour l'encodage et l'annotation des corpus numériques (Burnard et Bauman, 2008), qui comportent un volet plus spécifique pour l'annotation des informations linguistiques (chap. 15 : « Language corpora »).

Pour structurer les documents du corpus, nous avons créé un schéma de documents XML, c'est-à-dire une description générique de leur syntaxe interne (Fallside et Walmsley, 2004) ; (cf. § 4.3). Ce schéma de documents est nommé **Corpus-Contacts**.

4.2 Objectifs généraux

Le but de ce schéma de documents est de fournir un schéma de structure de documents contenant des corpus linguistiquement hétérogènes. Son utilisation a trois objectifs pratiques.

En premier lieu, le schéma de documents normalise la représentation des corpus. Au lieu de laisser chaque chercheur inventer ses propres conventions, et utiliser son propre vocabulaire, pour représenter les phénomènes qu'il observe, il établit une structure d'annotation commune. Ainsi, il permet à un groupe de chercheurs intéressés aux phénomènes de contacts de langues de disposer d'un système de représentation homogène. Cette normalisation permet à chacun d'entre eux de profiter, dans le cadre d'accords de mutualisation, de l'ensemble du corpus commun.

Deuxièmement, les corpus enregistrés contiennent une représentation de la structure et non de la forme. Un fichier XML au format **Corpus-Contacts** contient directement, par exemple, une indication que telle unité lexicale relève de telle langue. Ceci n'est plus indiqué (comme dans l'exemple 1) par une mise en forme superficielle comme la mise du mot en corps gras : celle-ci oblige en effet à avoir recours à une convention extérieure pour savoir ce qu'elle représente (ce qui la rend peu généralisable), et interdit en outre de cumuler des annotations lorsqu'un mot est potentiellement attribuable à plusieurs langues à la fois. Ceci n'empêche pas la mise en forme de ces indications de structure pour les rendre plus lisibles, et conformes aux conventions habituelles de représentation des linguistes. Mais cette mise en forme ne nécessite pas de travail supplémentaire de la part de l'utilisateur : elle est le résultat d'une conversion automatique réalisée par l'utilisation d'une feuille de style XSLT (Clark, 1999).

Enfin, la représentation structurée des corpus permet ensuite de définir des fonctions de recherche d'information et de classification sur des documents structurés.

Dans la conception d'un schéma de documents répondant aux besoins ainsi définis, on est soumis à deux contraintes :

– une contrainte de **normalisation** d'une part, nécessaire dans la perspective de maximiser la réutilisabilité et le partage des corpus (transposition aisée dans un autre format, portage aisé sur un autre site, ouverture de certaines parties du corpus à des communautés d'utilisateurs plus vastes, utilisation d'outils standards ...) ;
– une contrainte de **simplicité** d'autre part, absolument vitale pour l'acceptation de la ressource par les utilisateurs (si un utilisateur doit investir un temps important pour se former à un langage documentaire complexe, puis à nouveau du temps, lors de chaque saisie ou modification de corpus, pour remplir des dizaines de champs d'information obliga-

toires et la plupart du temps vides, il risque tout simplement de renoncer aux avantages de la normalisation).

Le choix qui a été fait ici a été de donner la priorité à la simplicité, sans sacrifier la normalisation. En l'occurrence, ceci signifie que :

- Seules les informations nécessaires, au stade actuel du programme de recherche, ont été incluses dans le schéma de documents **Corpus-Contacts**. Des pans entiers de normes existantes (notamment TEI) n'y ont pas été intégrées car elles n'ont pas d'utilité pour les chercheurs impliqués dans le programme de recherche sur les contacts de langue. Le fichier XSD contenant la description du schéma est donc « aussi petit que possible ».
- Le minimum concevable d'informations a été défini comme étant obligatoire. L'utilisateur peut presque commencer à utiliser le schéma de document en tapant du texte au kilomètre dans un éditeur XML [9], et ne commencer à utiliser les possibilités d'enrichissement de l'information que lorsque le besoin s'en fait sentir pour lui.
- Pour autant, les informations qui — parmi toutes celles dont l'encodage est prévu dans la norme TEI — sont utiles pour les besoins de **Corpus-Contacts**, respectent l'organisation et la nomenclature de la norme TEI, qui s'est imposée dans l'usage international comme la norme générale de référence pour la représentation des textes.

4.3 Structure globale

Corpus-Contacts est un *schéma de documents* XML au sens donné à ce terme par le W3C (Fallside et Walmsley, 2004). Un *document* XML est un document structuré contenant des informations stockées avec le texte sous forme de « balises ». Il se compose d'une hiérarchie d'*éléments* (par exemple : un livre se définit comme un ensemble de chapitres, eux-mêmes constitués d'un ensemble de paragraphes), eux-mêmes caractérisables par des *attributs* (par exemple : tel paragraphe est en français, tel paragraphe est en créole).

XML n'est pas une norme qui définit à l'avance, et dans le détail, tous les éléments et tous les attributs utilisables dans tous les types de documents. C'est une norme paramétrable, qui offre la possibilité de définir des types de documents en fonction des applications. Ainsi, on peut créer une famille de documents XML correspondant à des fiches bibliographiques, une autre contenant des textes littéraires, etc. Le concept qui permet cette polyvalence est le *schéma de documents*. Un schéma de documents est la définition de la structure commune que doivent avoir plusieurs documents XML de la même famille et destinés aux mêmes usages [10].

Le schéma de documents **Corpus-Contacts** est donc la description générique de ce en quoi consiste un corpus de données linguistiques utile pour la recherche sur le contact des langues. Il contient un squelette de document minimal (description des informations obligatoires), et

9. Dans le cadre du projet CLAPOTY, nous avons utilisé JAXE (éditeur XML développé en Java par Damien Guillaume, de l'Observatoire de Paris), logiciel facilement portable et extensible, qui a pu être configuré pour les besoins de notre projet, et installé sur plusieurs ordinateurs de systèmes d'exploitation différents. URL : http://jaxe.sourceforge.net/fr/. Cela étant, le schéma de documents XML se prête en principe à la manipulation par n'importe quel agent logiciel, et un script a récemment été développé par Sarra El-Ayari (Labex EFL) pour permettre l'importation et l'exportation du schéma de documents **Corpus-Contacts** à partir de la plateforme d'édition de corpus ELAN. URL : http://www.mpi.nl/corpus/html/elan/

10. Il existe un autre mécanisme de définition de famille de documents, plus ancien, hérité de SGML : la DTD *(Document Type Definition)*. Le *schéma XML* offre des possibilités supplémentaires, comme par exemple celle de définir des contraintes d'intégrité (ex. un locuteur mentionné dans un texte doit figurer dans l'inventaire des locuteurs défini dans l'en-tête du corpus).

définit des types d'éléments et d'attributs à utiliser pour tout un ensemble d'informations supplémentaires possibles.

L'élément racine d'un document tel que défini par notre schéma **Corpus-Contacts** est le *corpus* : un document contient un corpus et un seul.

Ce *corpus* est constitué d'un *en-tête* global, suivi de plusieurs *textes*. L'en-tête contient des informations valables pour l'ensemble du corpus : titre, éditeur, description, inventaire des locuteurs, inventaire des langues, et caractérisation dans des typologies définies par des auteurs ayant cherché à catégoriser les phénomènes de contact de langues sous plusieurs aspects : interactionnel (Lüdi, 1987), systémique (Auer, 1999), ou socio-historique (Winford, 2003) — voir (Léglise et Alby, 2013) pour une présentation de ces typologies [11].

Chaque *texte* est constitué d'un *en-tête* de texte, suivi d'une séquence d'un ou plusieurs *événements*. Un *événement* peut être soit une *indication paraverbale*, soit une *prise de parole*.

Les *indications paraverbales* explicitent des éléments de corpus liés à des événements situationnels, sans que ces éléments soient des fragments de langue (ex. « (rires) », « (A siffle) », etc.) Elles recouvrent les éléments dénommés *incident*, *kinesic*, et *vocal* dans les propositions du guide TEI, chap. 8 : « Transcriptions of speech » (Burnard et Bauman, 2008, p. 231–233). Elles peuvent constituer des événements à part entière, mais elles peuvent aussi, dans d'autres contextes, s'insérer dans des *prises de parole*.

Les *prises de parole* sont les éléments de base de la manifestation linguistique dans l'interaction. Chaque *prise de parole* est attribuable à un locuteur [12]. Dans la structure du document, la *prise de parole* est décomposée en quatre lignes d'information (« *tiers* ») : la *transcription* ; la *traduction interlinéaire morphème par morphème* ; la *liste des catégories morphosyntaxiques* de chaque morphème (« *POS-tags* ») ; et la *traduction libre*. La *transcription* est jalonnée d'indications de frontière de morphème (les *points de tabulation*) qui permettent d'aligner les trois premières lignes d'information.

C'est dans la ligne de *transcription* des prises de parole que se trouvent les annotations. Une partie de ces annotations sont classiques dans les corpus d'oral transcrit (indications paraverbales ou linguistiques ; indications des pauses ou des chevauchements ...) et recoupent le chap. 8 de la TEI ; certaines en revanche sont spécifiques à la description de l'hétérogénéité linguistique.

4.4 Éléments de description de l'hétérogénéité linguistique

Si la TEI a déjà prévu une structuration détaillée des annotations sur un grand nombre de plans possibles, l'hétérogénéité linguistique, dans sa complexité (cf. plus haut, § 3), reste peu prise en compte. Sur les 1600 pages de la dernière version de ses « *Guidelines* », elle n'est mentionnée que sur une seule page, et le cas est réglé par la recommandation d'utiliser une balise *foreign* :

11. La caractérisation des corpus en fonction de ces typologies « reçues » est importante pour notre projet de recherche, dont l'un des objectifs est justement d'interroger leur pouvoir explicatif dans le domaine des contacts de langue. Nous notons donc d'une part, au niveau global de chaque corpus, les classifications *a priori* faites selon ces typologies sur la base de critères observables externes ; et d'autre part, au niveau des transcriptions, tous les critères observables relevés sur la manifestation linguistique elle-même (annotés avec le moins d'*a priori* méthodologiques possibles). Le but des opérations de fouille de données qui seront faites sur l'ensemble de ces corpus est de déterminer si les catégorisations émergeant de la classification automatique des données recoupent celles proposées par la littérature linguistique, et si elles révèlent un rôle prédictif des paramètres identifiés par ces auteurs pour identifier certains phénomènes de contact.

12. Il est par ailleurs possible de noter les chevauchements, lorsque plusieurs locuteurs s'expriment en même temps.

> « Words or phrases which are not in the main language of the text should be tagged as such (...) :
> “`John eats a <foreign xml:lang="fr">croissant</foreign> every morning.`” (...) » (Burnard et Bauman, 2008, p. 65).

> N.B. On peut noter que les plates-formes d'annotation existantes qui imposent un format de document, qu'elles s'inspirent plus ou moins directement de la TEI, n'ont pas été plus loin que ce point dans la représentation des données hétérogènes d'un point de vue linguistique. La plate-forme d'annotation de l'oral *Transcriber* [13], par exemple, dispose d'une « balise de changement de langue ». D'autres plate-formes d'annotation de corpus, telles ELAN [14], GATE [15], Glozz [16], ou XSTANDOFF [17], sont plus génériques, en ce sens qu'elles n'imposent pas un schéma d'annotation prédéfini, mais permettent à l'utilisateur de définir lui-même son modèle d'annotation, en même temps qu'il édite le corpus. De telles plate-formes pourraient tout à fait être utilisées comme outils d'édition de corpus multilingues, une fois paramétrées pour faire usage du schéma de documents proposé ici (mais sans que chaque utilisateur s'autorise à modifier le schéma de son côté une fois un travail commun commencé — sous peine de perdre le bénéfice de la normalisation).

La spécificité du schéma d'encodage **Corpus-Contacts** réside donc en grande partie dans ce domaine. Comme nous l'avons vu (§ 3), dans les interactions quotidiennes, il n'est pas aisé, voire il est contre-indiqué, d'identifier de manière univoque *la* langue d'un énoncé ou d'un segment d'énoncé. Nous proposons donc ici un système qui permet d'attribuer *plusieurs* langues à un passage (avec un ordre de dominance / vraisemblance).

Le système proposé, avec cette extension, est rétro-compatible avec la méthode « traditionnelle » d'identification de la langue des textes (utilisation de l'attribut *lang*, recommandée par l'IEFC dans le monde XML/HTML, ainsi que par la TEI). Il permet de préciser plusieurs langues lorsque cela est nécessaire, ou de s'en tenir à l'attribution d'une seule langue si cela suffit.

Par ailleurs, dans le cas d'énoncés à rattachement multiple, un agent logiciel qui ne reconnaît pas cette extension pourra se rabattre sur la première langue, qui est celle qui sera indiquée dans l'attribut *lang* (méthode classique) — cette solution, qui ne nous convient pas méthodologiquement, assure toutefois la compatibilité de notre système d'annotation avec les systèmes antérieurs.

Nous allons commencer ci-dessous par exposer brièvement la méthode générale d'identification d'une langue, ainsi que les normes de représentation utilisées (§ 4.4.1). Puis nous expliquerons la structure de l'élément d'information utilisé pour coder le rattachement simultané d'un passage à plusieurs langues (§ 4.4.2).

4.4.1 L'attribut *xml:lang*

Un attribut XML, *lang*, sert à identifier la langue d'un segment linguistique, soit au niveau d'un énoncé entier, soit au niveau d'un fragment d'énoncé (repéré par l'élément *segment*).

13. URL : http://trans.sourceforge.net/
14. URL : http://www.mpi.nl/corpus/html/elan/
15. URL : http://gate.ac.uk
16. URL : http://www.glozz.org/
17. URL : http://www.xstandoff.net

Suivant les recommandations de la TEI, la valeur de l'attribut *lang* est déterminée selon la norme en usage sur internet [18], et codifiée par l'« Internet Society » sous la référence RFC-5646 [19].

La norme RFC-5646 prévoit que la langue proprement dite est codée par une étiquette tirée de l'une des variantes de la norme ISO-639. La variante 1 de cette norme (ISO-639-1) comprend des codes à deux lettres utilisés pour les langues les plus courantes (*fr* pour le français, *en* pour l'anglais ...) La variante 2 comprend des codes à deux et trois lettres, mais son répertoire est assez limité [20], pour l'usage de linguistes. Nous utilisons donc concrètement la variante la plus étendue, centralisée par le SIL : l'ISO 639-3 [21], qui a pour vocation d'attribuer un code à trois lettres à toutes les langues connues. La plupart des langues y ont donc un codage sous la forme d'une étiquette à trois lettres, comme *fra* pour le français ou *eng* pour l'anglais.

Il existe par ailleurs trois étiquettes spéciales : *mul (« multiple languages »)* pour les passages contenant plusieurs langues à la fois ; *und (« undetermined »)* pour les passages dont on n'a pas réussi à identifier la langue ; et *zxx (« non-linguistic content »)* pour les passages de contenu articulé mais non-linguistique (ex. « chouba douba douwa »).

Les recommandations du RFC-5646 prévoient en outre la possibilité d'ajouter des précisions à l'étiquette de langue. Les précisions (facultatives) peuvent concerner :

1. une indication de variante, par exemple de variante dialectale, codée par une étiquette de 5 à 8 lettres (par exemple *djk-aluku* désigne la variante aluku du businenge tongo, langue créole à base anglaise parlée dans la région du Maroni en Guyane française, et *djk-ndyuka* la variante ndyuka) [22] ;
2. une indication de système d'écriture ;
3. une indication d'aire géographique, servant à préciser qu'on souhaite identifier une variante régionale d'une langue de grande extension (par exemple, *eng-US* pour l'anglais des États-Unis, et *eng-GB* pour l'anglais de Grande-Bretagne [23]).

On peut souhaiter dénoter une aire géographique ne correspondant pas à un pays, et il est alors possible d'utiliser les codes numériques désignant des zones géographiques du monde utilisés par la division des statistiques de l'ONU [24]. Par exemple, *spa-419* peut être utilisé pour désigner globalement l'espagnol d'Amérique latine et des Caraïbes.

18. Cette norme est employée par exemple pour caractériser la langue utilisée par un site web.

19. *Internet Engineering Task Force*, RFC *(Request For Comments)* 5646 : *Tags for Identifying Languages*. URL : http://tools.ietf.org/html/rfc5646.

20. Sa communauté d'utilisateurs est constituée essentiellement de documentalistes, et son usage est donc orienté vers les langues de l'édition.

21. ISO *(International Standards Organization)* standard 639-3 : *Codes for the representation of names of languages – Part 3*. C'est le SIL *(Summer Institute of Linguistics)* qui a été désigné comme organisme centralisateur de cette norme. URL : http://www.sil.org/iso639-3/.

22. Pour définir le nom des langues et leurs différentes subdivisions, nous nous appuyons ici sur les travaux réalisés par les linguistes de la région (cf. notamment Goury et Migge (2003) dans l'exemple cité) et sur le point de vue et les idéologies linguistiques des acteurs sociaux (cf. Léglise et Migge (2006) pour ce même exemple).

23. La norme RFC 4646 prévoit que l'étiquette utilisée pour dénoter une extension géographique soit, dans le cas typique, tirée de la liste des codes de noms de pays établie par la norme ISO 3166.
URL : http://www.iso.ch/iso/fr/country_codes/iso_3166_code_lists/french_country_names_and_code_elements.htm.

24. M 49 : *Codage statistique normalisé des pays et zone*. URL : http://unstats.un.org/unsd/methods/m49/m49regnf.htm.

L'institution chargée de coordonner les conventions techniques régissant le fonctionnement d'internet, l'IANA, a également pour objectif de maintenir à jour une table normalisée des étiquettes utilisables pour déterminer l'attribut *lang*[25].

Il est possible de spécifier la valeur de l'attribut *lang* à différents niveaux de généricité. Par exemple, un énoncé en français prononcé par un francophone de Guyane peut également être étiqueté *fr* (français) ou *fr-GF* (français de Guyane) : les deux usages sont conformes à la norme. La décision de fixer le niveau approprié de généricité appartient entièrement à l'éditeur du corpus : il n'y a pas de règle générale, si ce n'est celle de la pertinence. Au moment de choisir le niveau de généricité, l'éditeur doit se rappeler que l'information concernant les niveaux supérieurs est automatiquement incluse dans l'information concernant les niveaux les plus spécifiques, alors que l'inverse n'est pas vrai. Il convient donc simplement de songer au degré de précision qui mérite d'être conservé. Ce principe de base étant énoncé, le choix de l'étiquetage dépend du cas de figure. On pourrait considérer, d'un point de vue privilégiant l'exactitude, qu'il est conseillé de donner toujours la précision maximale (les niveaux plus génériques pouvant de toute façon en être déduits), mais il existe en pratique des cas où il est justifié d'utiliser un niveau plus générique. Par exemple, à moins de vouloir dénoter explicitement l'usage de tournures régionales en français, il est probablement inutile — voire probablement erroné, car non démontré — de spécifier systématiquement *fr-GF* pour tout énoncé en français enregistré en Guyane (cf. Léglise (2012) pour une discussion sur ce point). De même, dans le cas de l'étiquetage de segments courts, où la forme manifestée est une forme générique qui, sans contexte supplémentaire, ne permet pas de distinguer entre deux variétés identiques dans deux variantes dialectales apparentées, il est pertinent de conserver une étiquette générique.

4.4.2 Rattachement d'un passage à plusieurs langues

Afin de représenter le fait qu'un énoncé ou segment d'énoncé peut être rattaché à plusieurs langues à la fois, on a introduit dans le schéma de documents **Corpus-Contacts** un élément (facultatif), appelé *langues*.

Cet élément *langues* est facultatif, au contraire de l'attribut principal *lang*, rattaché directement au niveau supérieur (au niveau de l'élément *transcription* ou *segment*). Dans le cas où l'attribution de la prise de parole ou du segment à une langue est univoque, seul l'attribut *lang* est requis.

Dans le cas, en revanche, où l'on souhaite rattacher un énoncé ou segment d'énoncé à plusieurs langues, on doit utiliser les deux :

- la liste de langues, contenue dans un élément *langues* immédiatement subordonné à l'élément *transcription* ou *segment* concerné ;
- *et* l'attribut *lang*, qui reste obligatoire, en tant qu'attribut, rattaché à l'élément *transcription* ou *segment*.

La liste de langues ne contient qu'une liste ordonnée d'étiquettes de codes de langues, conformes à la norme décrite ci-dessus (§ 4.4.1). Il n'y a pas d'attribut supplémentaire pour indiquer, par exemple, une probabilité de rattachement (impossible à quantifier). En revanche, on doit considérer que l'ordre dans lequel les langues sont mentionnées est potentiellement signifiant : il reflète un ordre de vraisemblance du rattachement. Si l'on a utilisé cette possibilité de rattachement multiple parce qu'on *hésite* entre deux langues, alors la première langue est plus « probable » que la deuxième ; et si on l'a utilisé parce qu'on pense que le segment relève

25. IANA *(Internet Assigned Numbers Authority) Language subtags regitry*.
URL : http://www.iana.org/assignments/language-subtag-registry.

de deux langues *à la fois*, alors la première langue citée est la plus fréquente dans l'alternance de langues considérée.

Lorsque l'on utilise cette possibilité, l'usage logique consiste à indiquer, au niveau supérieur (celui de la prise de parole ou du segment pour lequel on donne une liste de langues possibles), que l'attribut *lang* a la valeur *mul (« multiple languages »)*. Au final l'étiquette *mul* nous sert donc à coder des segments multilingues ; mais il est important de noter que ce « multilinguisme » peut avoir deux interprétations : une interprétation paradigmatique (P), et une interprétation syntagmatique (S).

(P) Lorsqu'un segment (généralement bref), possédant à peu près le même signifiant dans deux langues A et B, ne donne pas suffisamment de critères pour déterminer s'il doit être rattaché à l'une de ces langues plutôt qu'à une autre, l'étiquetage *mul* signifie : ce segment pourrait être aussi bien une unité de A ou une unité de B — voire : ce segment est peut-être une *unité flottante* (au sens de (Ledegen, 2012)) entre A et B dans le répertoire d'un locuteur bilingue. La liste des langues entre lesquelles il y a hésitation (ou flottement) est donnée par l'élément *langues*.

(S) Lorsqu'un énoncé ou une prise de parole comporte plusieurs segments en différentes langues, ou qui peuvent être catégorisés comme à rattachement multiple, sans qu'il soit évident qu'il s'agisse de brèves insertions dans une langue donnée, alors nous avons souhaité utiliser l'étiquetage *mul* pour signifier : cet énoncé doit tout entier être considéré comme multilingue (appartenant potentiellement à plusieurs langues possibles), il n'est pas possible de le considérer comme la manifestation d'une langue matrice et d'inserts — cf. (Auer et Muhamedova, 2005) pour une discussion par exemple.

Un exemple de l'interprétation (P) est donnée dans l'exemple (2)[26]. L'ensemble de la phrase est principalement en créole, mais dans le cas de « pour l'instant », on n'a pas affaire à un mot du vocabulaire créole fondamental. Il n'est pour autant pas possible d'affirmer qu'il s'agit d'une pure importation du français, non seulement à cause des indices phonologiques[27], mais aussi parce qu'il s'agit d'un connecteur de discours qui semble fréquemment accessible aux locuteurs, dans les deux langues (le reste du corpus en témoigne).

(2) Corpus Clapoty — Vaillant : *Lignes de vie*

	pour	l'instant					
Piské	***pou***	***lenstan***	***sé***	***journalis***	***ki***	***ni***	***la***
puisque	pour	l'instant	être.COP	journaliste	REL;SBJ	avoir	là

Puisque pour l'instant ce (ne) sont (que) des journalistes qui sont là

On note donc qu'il s'agit d'un passage multilingue, en mentionnant la liste de langues concernées — ici, le créole antillais (*acf*) et le français (*fra*). En XML, ce flottement est indiqué comme suit (lignes 3 à 7).

```
<transcription lang="acf">
 piskè <tab/>
 <segment lang="mul">
  <langues><langue lang="acf"/><langue lang="fra"/></langues>
```

26. Abréviations introduites ici : COP : copule ; REL : relatif ; SBJ : sujet.

27. Le /r/ est élidé — encore que ce phénomène puisse tout à fait s'observer également en français oral, surtout chez les locuteurs de cette région.

```
   <trans_alt lang="acf">pou</trans_alt> <trans_alt lang="fra">pour</trans_alt> <tab/>
   <trans_alt lang="acf">lenstan</trans_alt> <trans_alt lang="fra">l'instant</trans_alt>
  <tab/>
 sé <tab/> jounalis <tab/> ki <tab/> ni <tab/> la
```

(4) Corpus Clapoty — Léglise : *Cour de récréation*

Vini	**non**	bande	de	putes
venir	d'accord	bande	de.GEN	pute

Venez ici, bande de putes

Dans cet exemple, le premier mot est un impératif en créole guyanais. Le deuxième mot, « non », est une particule énonciative (un « ponctuant ») renforçant l'impératif, d'usage courant en Guyane dans cette fonction, aussi bien en français qu'en créole ; on ne sait pas très bien s'il doit être considéré ici comme un mot français, un mot créole, ou un mot typique du français parlé en Guyane. Le reste de la prise de parole consiste en un groupe nominal français en fonction d'apostrophe.

La représentation source (XML) de ce passage est la suivante.

```
<transcription lang="mul"><langues><langue lang="gcr"/><langue lang="fra"/></langues>
 vini
 <tab/>
 <langues><langue lang="gcr"/><langue lang="fra"/></langues>
  non
 <tab/>
 bande <tab/> de <tab/> putes
```

outre, l'examen systématique de ces zones de « frontière » pourrait permettre de tester l'hypothèse, qui a été formulée par certains linguistes, selon laquelle certains sous-systèmes grammaticaux seraient plus perméables que d'autres aux changements induits par le contact de langues (Matras, 2007; Léglise, 2012).

Cependant, ces analyses nécessitent de passer du niveau de l'occurrence au niveau du type, et ce passage présente des difficultés particulières dans le cas des unités rattachées à plusieurs langues à la fois.

Pour résoudre cette question, nous adoptons un modèle de positionnement des unités observables dans les corpus dans un espace multidimensionnel. On peut présenter cette modélisation schématiquement de la façon suivante : tout se passe comme si l'identité linguistique d'un type n'était pas une fonction dans un ensemble de langues (*gen* $\mapsto$ créole ; *avoir* $\mapsto$ français), mais une fonction dans un espace vectoriel dont les langues sont les axes (*gen* : 100% créole, 0% français ; *avoir* : 0% créole, 100% français ; *travail/travay* : 50% créole, 50% français).

Ce principe aboutit à créer, lors de l'indexation et de l'analyse du corpus, une matrice de dispersion des unités (types) dans l'espace des langues, comme illustré sur la figure 5. On note v le nombre de vocables (types d'unités distinctes) dans le corpus, et l le nombre de langues impliquées dans la situation de contact. Pour chaque vocable w_i, $l_{i,k}$ représente la « part » de la langue k dans l'identité linguistique de w_i.

$$L = \begin{pmatrix} l_{1,1} & l_{2,1} & \dots & l_{i,1} & \dots & l_{v,1} \\ l_{1,2} & l_{2,2} & \dots & l_{i,2} & \dots & l_{v,2} \\ \vdots & \vdots & & \vdots & \ddots & \vdots \\ l_{1,l} & l_{2,l} & \dots & l_{i,l} & \dots & l_{v,l} \end{pmatrix}$$

FIG. 2 – *matrice de dispersion des mots-types dans l'espace des langues. Le vecteur colonne* $\boldsymbol{l}_i = (l_{i,1}, l_{i,2}, \dots l_{i,l})$ *représente la « position » du vocable* w_i *dans l'espace des langues (NB. « espace des langues » a ici un sens purement mathématique (espace vectoriel) et ne présuppose aucune conception linguistique).*

Dans les travaux antérieurs (Nock et al., 2009), où l'on considère que les systèmes des différentes langues sont cloisonnés et où un mot (w_i) ne peut appartenir qu'à une seule langue à la fois (L_k), tous les coefficients $l_{i,j}$ valent 0 sauf un ($l_{i,k} = 1$). Dans le cadre des corpus plurilingues du schéma **Corpus-Contacts**, n'importe quel vecteur colonne $\boldsymbol{l}_i$ peut avoir plusieurs coefficients non-nuls, selon la « dispersion linguistique » de w_i.

Si chaque occurrence de w_i n'était attribuée qu'à une seule langue, $l_{i,k}$ représenterait la proportion des occurrences de w_i attribuée à la langue L_k. Étant donné que certaines occurrences peuvent elles-mêmes être multilingues, la détermination de la matrice de dispersion des types est un peu plus complexe.

On procède de la manière suivante : pour chaque vocable w_i :

1. On identifie les d_i (> 1) occurrences de w_i dans le corpus : $\{\mathrm{T}_{i1}, \mathrm{T}_{i2}, \dots, \mathrm{T}_{id_i}\}$;
2. pour chaque occurrence T_{ij}, on calcule la contribution de T_{ij} à la valeur de $\boldsymbol{l}_i$:

$$l_{i,j,k} = \begin{cases} 0 & \text{si } \mathrm{L}_k \notin \lambda(\mathrm{T}_{ij}) \\ \frac{1}{|\lambda(\mathrm{T}_{ij})|} & \text{si } \mathrm{L}_k \in \lambda(\mathrm{T}_{ij}) \end{cases}$$

(où $\lambda(\mathrm{T}_{ij})$ désigne l'ensemble des langues attribuées à l'occurrence T_{ij}) ;

3. on fait la somme des contributions des occurrences de w_i :
   $l_{i,k} = \sum_{j=1}^{n_i} l_{i,j,k}$

Puisque $\forall i, j,\ \sum_{k=0}^{l} l_{i,j,k} = 1$ (chaque occurrence a un poids de 1 quelle que soit sa « dispersion » linguistique), nous sommes assurés que le poids total de chaque vecteur $\boldsymbol{l}_i$ est d_i (le nombre d'occurrences du mot dans le corpus). Ainsi le poids relatif des mots dans le corpus (la distribution de leur fréquence) n'est pas distordu par cette méthode de prise en compte du plurilinguisme.

6 Considérations pratiques et conclusion

Nous avons présenté les aspects techniques d'une initiative d'encodage de corpus plurilingues qui prend en compte la complexité des phénomènes réels de contact de langues. L'objectif de l'encodage de ces phénomènes est de pouvoir les analyser par des méthodes automatiques, et d'en faire émerger des catégories déterminées uniquement sur une base empirique — dont nous pourrons évaluer *a posteriori* le recoupement avec certaines catégories définies dans la littérature spécialisée.

Afin de ne pas imposer aux données linguistiques une précatégorisation consciente ou inconsciente, nous avons dû développer un schéma d'annotation qui permette de noter toutes les informations disponibles sur tous les phénomènes observés, tout en évitant de qualifier ceux-ci par des termes renvoyant à des concepts *a priori* (emprunt, alternance, mélange ...) La structure de notre schéma est donc « agnostique » quant aux théories du contact linguistique, et présente un aspect volontairement simpliste (en termes de dimension de l'espace des variables décrites).

Le schéma de documents est concrètement utilisé dans le programme de recherche ANR Clapoty [30] depuis 2009, et actuellement dans le cadre du programme LC1 du Labex EFL [31]. À ce jour [32], il a permis de collecter 94 corpus multilingues recueillis par 10 linguistes sur des terrains et dans des aires linguistiques très variées, représentant 33 langues différentes dans des situations de contact très contrastées. Des outils de saisie [33], de validation [34], d'import [35], d'extraction automatique [36], de visualisation [37], et d'exploration des données [38], ont été mis à la disposition des membres du projet sur un serveur interne utilisant le système d'exploitation *Debian Linux 7* et le serveur HTTP *Apache 2* [39]. L'expérience a permis d'éprouver la solidité

30. URL : http://clapoty.vjf.cnrs.fr

31. Programme LC1 *(Multifactorial Analysis of Language Change)* dirigé par I. Léglise dans le cadre du Labex EFL (financé par ANR/CGI). URL : http://www.labex-efl.org

32. Décompte fait sur la base de données du projet à la date du 9 janvier 2014.

33. Éditeur XML JAXE, cf. note 9 *supra*.

34. Implantée par des fonctions de la librairie *libxml2* (URL : http://www.xmlsoft.org) appelées par l'interpréteur PHP intégré à Apache (URL : http://httpd.apache.org).

35. Réalisé par un script PHP d'Apache.

36. Réalisé par un script PHP, qui analyse et décortique les documents XML à l'aide de fonctions de la librairie *libxml2*, puis peuple, grâce aux données extraites, une base de données relationnelles implantée sur un serveur MySQL Community Server 5.1 (URL : http://dev.mysql.com/downloads/mysql/).

37. Réalisée par une application *à la volée*, par le serveur Apache, d'un filtre de transformation XSLT développé en parallèle au schéma de documents décrit dans cet article.

38. Un concordancier en PERL (module d'Apache) a été développé par Anne Garcia-Fernandez, dans le cadre du programme LC1 du Labex EFL.

39. Chaque utilisateur reste néanmoins libre du système client qu'il utilise, la plus grande partie de ces outils fonctionnant en mode serveur à travers une interface web. Le seul outil fonctionnant en mode client est l'éditeur XML

et la généricité des choix de représentation effectués, et décrits dans cet article. Des différences spontanées d'usage se sont manifestées dans les premières étapes du travail, lorsque les linguistes ont pris en main les outils d'annotation (éditeur XML, interface de validation, de téléchargement et de visualisation sur un site web) . Ces différences ont été peu à peu harmonisées à l'occasion de séminaires de réflexion collective débouchant sur des prescriptions d'utilisation des balises et des codes normalisés (langues, étiquettes morpho-syntaxiques).

Les outils mis en place ont d'ores et déjà permis que ces corpus soient utilisés dans des travaux de recherche linguistique. La phase de fouille de données automatique et semi-automatique est en cours de démarrage.

Le programme de recherche qui a fait naître ce travail poursuit son initiative sur d'autres tâches, qui n'en sont encore qu'à leurs étapes initiales. D'une part, l'application de méthodes de catégorisation automatique de données pourrait permettre à l'avenir de faire émerger des régularités qui n'apparaissent que lorsque l'on considère le corpus au niveau global. On pourrait ainsi voir apparaître des configurations de variables de niveaux divers (catégories d'unités linguistiques, fonctions syntaxiques ou communicatives, rapports de force entre langues) ayant un comportement spécifique dans les phénomènes de contact de langues. D'autre part, une annotation qualitative systématique des phénomènes de contact observés dans ces corpus (au niveau de l'occurrence comme au niveau du type) est également en cours (cf. (Léglise et Alby, 2013)), dans le cadre de grilles d'explication plurifactorielles ; l'objectif de ce travail est de réunir les différents axes explicatifs restés jusqu'ici traditionnellement disjoints et de les confronter aux résultats de l'analyse automatique.

Références

Aikhenvald, A. Y. et R. M. W. Dixon (Eds.) (2006). *Grammars in Contact : A Cross-Linguistic Typology*. Oxford (Angleterre, Royaume-Uni) : Oxford University Press.

Allen, J. D. *et al.*. U. C. (Ed.) (2012). *The Unicode Standard, Version 6.2*. Mountain View (Californie, États-Unis) : The Unicode Consortium.

Auer, P. (1999). From code-switching via language mixing to fused lects : Toward a dynamic typology of bilingual speech. *International Journal of Bilingualism 3*(4), 309–332.

Auer, P. et R. Muhamedova (2005). 'embedded language' and 'matrix language' in insertional language mixing : some problematic cases. *Rivista di Linguistica 17*(1), 35–54.

Bray, T., J. Paoli, C. M. Sperberg-McQueen, E. Maler, et F. Yergeau (Eds.) (2008). *Extensible Markup Language (XML) 1.0 (Fifth Edition)*. World Wide Web Consortium (W3C).

Burnard, L. et S. Bauman (Eds.) (2008). *TEI P5 : Guidelines for Electronic Text Encoding and Interchange*. Oxford (Angleterre, Royaume-Uni) : Text Encoding Initiative.

Chamoreau, C. et L. Goury (Eds.) (2012). *Changement linguistique et langues en contact. Approches plurielles du domaine prédicatif*. Paris (France) : CNRS Éditions.

Clark, J. (Ed.) (1999). *XSL Transformations (XSLT) Version 1.0*. World Wide Web Consortium (W3C).

Deprez, C. (1994). *Les enfants bilingues : langues et familles*. Paris (France) : Didier.

JAXE ; mais comme il est développé en Java, il a pu être installé avec succès sur des postes clients utilisant les trois systèmes d'exploitation les plus courants (Windows, Linux et Macintosh).

Fallside, D. C. et P. Walmsley (Eds.) (2004). *XML Schema Part 0 : Primer (Second Edition)*. World Wide Web Consortium (W3C).

Goury, L. et B. Migge (2003). *Grammaire du Nengee. Introduction aux langues aluku, ndyuka et pamaka*. Paris (France) : IRD Éditions.

Gumperz, J. (1982). *Language and Social Identity*. Cambridge (Angleterre, Royaume-Uni) : Cambridge University Press.

Heine, B. et T. Kuteva (2005). *Language Contact and Grammatical Change*. Cambridge (Angleterre, Royaume-Uni) : Cambridge University Press.

Heine, B. et T. Kuteva (2007). *The Genesis of Grammar*. Oxford (Angleterre, Royaume-Uni) : Oxford University Press.

Jørgensen, J. N., M. S. Karrebæk, L. M. Madsen, et J. S. Møller (2011). Polylanguaging in superdiversity. *Diversities 13*(2), 23–37.

Juillard, C. (1995). *Sociolinguistique urbaine. La vie des langues à Ziguinchor, Sénégal*. Paris (France) : CNRS Éditions.

Ledegen, G. (2012). Prédicats "flottants" entre le créole acrolectal et le français à la réunion : exploration d'une zone ambiguë. In Chamoreau et Goury (2012), pp. 251–270.

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 Chamoreau et Goury (2012), pp. 203–230.

Léglise, I. et S. Alby (2013). Les corpus plurilingues, entre linguistique de corpus et linguistique de contact : réflexions et méthodes issues du projet CLAPOTY. *Faits de Langues* (41), 95–122.

Léglise, I. et C. Chamoreau (2013). Variation and change in contact settings. In I. Léglise et C. Chamoreau (Eds.), *The interplay of variation and change in contact settings*, pp. 1–20. Amsterdam (Pays-Bas) : John Benjamins.

Léglise, I. et B. Migge (2006). Towards a comprehensive description of language varieties : A consideration of naming practices, ideologies and linguistic practices. *Language in Society 35*(3), 313–339.

Lüdi, G. (1987). Les marques transcodiques : regards nouveaux sur le bilinguisme. In G. Lüdi (Ed.), *Devenir bilingue, parler bilingue*, pp. 1–19. Tübingen (Allemagne) : Niemeyer.

Lüdi, G. et B. Py (1986). *Être bilingue*. Berne (Suisse) : Peter Lang.

Matras, Y. (2007). The borrowability of structural categories. In Y. Matras et J. Sakel (Eds.), *Grammatical Borrowing in Cross-Linguistic Perspective*, pp. 31–73. Berlin (Allemagne) : Walter de Gruyter.

Matras, Y. (2009). *Language Contact*. Cambridge (Angleterre, Royaume-Uni) : Cambridge University Press.

Myers-Scotton, C. (2002). *Contact Linguistics : Bilingual Encounters and Grammatical Outcomes*. Oxford (Angleterre, Royaume-Uni) : Oxford University Press.

Nelson, L. (2008). Le contact de langues au travail : Étude de l'alternance codique entre les langues français-créole dans les situations de service à l'accueil direct d'EDF guyane. Master's thesis, Université Lyon 2. Mémoire de Master 2.

Nock, R., P. Vaillant, C. Henry, et F. Nielsen (2009). Soft memberships for spectral clustering, with application to permeable language distinction. *Pattern Recognition 42*(1), 43–53.

Peyraube, A. (2002). L'évolution des structures grammaticales. *Langages 146*, 46–58.

Pfänder, S. (2000). *Aspekt und Tempus im Frankokreol*. ScriptOralia. Tübingen (Allemagne) : Günter Narr Verlag.

Schmidt, T. et K. Wörner (Eds.) (2012). *Multilingual Corpora and Multilingual Corpus Analysis*. Amsterdam (Pays-Bas) : John Benjamins.

Tabouret-Keller, A. (2001). Pour une vision dynamique des situations linguistiques complexes. *La linguistique 37*(1), 21–28.

Thomason, S. G. (2001). *Language Contact : An Introduction*. Edinburgh (Écosse, Royaume-Uni) : Edinburgh University Press.

Thomason, S. G. et T. Kaufmann (1988). *Language Contact, Creolization, and Genetic Linguistics*. Berkeley (Californie, États-Unis) : University of California Press.

Winford, D. (2003). *An Introduction to Contact Linguistics*. Oxford (Angleterre, Royaume-Uni) : Blackwell.

Wurm, S. (1996). *Atlas des langues en péril dans le monde*. Paris (France) : UNESCO.

Summary

In the frame of a research programme on the study of language contact phenomena and of their role in linguistic change, there currently is an effort to collect plurilingual corpora, exhibiting a great variety of contact phenomena on a sample of languages of various genetical and typological background. This has implied developing a specific document processing software for digital corpora with internal plurilingualism, in order to represent, store, annotate, and visualize their linguistic data, and to build data mining tools. Existing encoding standards have been extended to cope with such phenomena as speech segments "floating" between languages, occurring in plurilingual talk. In this article, we describe the structure that has been defined for the plurilingual corpora, and the background definition of plurilingual linguistic units that is used for statistical analysis in the corpora.