# Grammaires factorisées pour des dialectes apparentés

Pascal Vaillant

*TALN 2008, Avignon, 9–13 juin 2008*

# Grammaires factorisées pour des dialectes apparentés

Pascal Vaillant
GRIMAAG, Université des Antilles-Guyane, B.P. 792, 97337 Cayenne cedex
pascal.vaillant@guyane.univ-ag.fr

**Résumé.** Pour la formalisation du lexique et de la grammaire de dialectes étroitement apparentés, il peut se révéler utile de factoriser une partie du travail de modélisation. Les sous-systèmes linguistiques isomorphes dans les différents dialectes peuvent alors faire l'objet d'une description commune, les différences étant spécifiées par ailleurs. Cette démarche aboutit à un modèle de grammaire à couches : le noyau est commun à la famille de dialectes, et une couche superficielle détermine les caractéristiques de chacun. Nous appliquons ce procédé à la famille des langues créoles à base lexicale française de l'aire américano-caraïbe.

**Abstract.** The task of writing formal lexicons and grammars for closely related dialects can benefit from factoring part of the modelling. Isomorphic linguistic subsystems from the different dialects may have a common description, while the differences are specified aside. This process leads to a layered grammar model: a kernel common to the whole family of dialects, and a superficial skin specifying the particular properties of each one of them. We apply this principle to the family of French-lexifier creole languages of the American-Caribbean area.

**Mots-clés :** TAG, modélisation, grammaire, variation dialectale.

**Keywords:** TAG, grammar, modelling, dialectal variation.

## 1 Introduction

Certains groupes de langues présentent des similitudes importantes quant au fonctionnement de leur système grammatical. Bien que cela ne soit pas logiquement nécessaire, ces similitudes sont le plus souvent liées à une origine historique commune, c'est-à-dire qu'elles se manifestent dans des groupes de dialectes génétiquement apparentés. Pour mémoire, nous prenons ici le mot « dialecte » dans le sens technique qu'il a en linguistique, où il désigne un système linguistique considéré non pas en lui-même, mais dans son appartenance à une famille de systèmes apparentés. La question de la *délimitation* du seuil de ressemblance qui définit une famille de dialectes, et du seuil de différence qui définit la frontière entre deux dialectes distincts, ne sera pas abordée dans cette discussion[1]. La réflexion que nous présentons peut donc s'appliquer indifféremment à ce que l'on choisirait d'appeler, selon le contexte, l'usage, ou les pré-supposés du locuteur, un *ensemble de variantes dialectales d'une même langue* ou une *famille de langues étroitement apparentées*.

[1] Dans le cas qui nous a occupé, nous avons affaire à des langues insulaires, pour lesquelles, malgré une variabilité géographique interne, la question de la délimitation des variétés principales au sein d'une famille ne fait pas l'objet de polémique. Celle des familles elle-même est plus discutée (cf. plus loin, § 3).

Entre les différents dialectes appartenant à une famille de ce type, les similitudes entre systèmes ne sont pas des événements isolés, mais se constituent elles-mêmes en système : on peut parler d'un isomorphisme (partiel) de systèmes. C'est notamment le cas, pour des raisons évidentes, quand il y a parenté génétique : c'est un système commun qui est hérité, et ce sont les divergences qui s'accumulent au fil du temps. L'argument que nous souhaitons avancer ici est que lorsque cet isomorphisme couvre une part suffisamment importante du système morpho-syntaxique, il est intéressant du point de vue de la description formelle de *factoriser*, c'est-à-dire de concevoir une modélisation en plusieurs couches : un noyau de grammaire commun, et une couche superficielle ne spécifiant que les phénomènes particuliers à chaque dialecte.

L'idée que nous défendons semble intuitive, et s'inscrit dans la cohérence logique d'approches comme celle du programme minimaliste (Chomsky, 1995). La visée de notre modélisation, cependant, reste très pratique, et centrée sur le cas d'une famille très resserrée de dialectes : ainsi, l'application que nous faisons du concept de « paramètres » se restreint à la paramétrisation d'un dialecte au sein de son taxon linguistique, et nous n'avons pas à examiner les niveaux de généralisation allant, en amont, de ce taxon à des principes universels.

Les facteurs contextuels qui rendent cet effort de factorisation particulièrement utile sont au nombre de deux. Le premier est la proximité des dialectes : le travail de factorisation est facilité si une grande partie de la grammaire des différents dialectes se retrouve dans le noyau commun, et que les couches superficielles sont minces. Le second est la rareté des moyens humains disponibles pour le travail de modélisation : s'il n'y a qu'une petite équipe de spécialistes pour modéliser plusieurs dialectes, alors elle a évidemment intérêt à concentrer ses efforts dans une optique de réutilisabilité, même si elle y perd potentiellement par rapport à la finesse de description qu'elle pourrait atteindre en se consacrant exclusivement à un système linguistique unique. Ces deux facteurs se trouvent réunis dans les situations que l'on pourrait grossièrement définir comme étant celles des « petites » langues dialectalisées[2]. Le cas d'espèce qui nous occupe, celui des créoles français de l'aire amérique-caraïbe, illustre bien cette conjonction de facteurs (cf. plus bas, § 3).

## 2 Choix de modélisation formelle

Candito (1998) a montré l'intérêt, et la faisabilité, de l'idée consistant à factoriser une partie de la modélisation grammaticale pour des langues apparentées : son travail porte sur le développement parallèle de grammaires pour le français et l'italien. Ces deux langues sont suffisamment proches pour avoir un noyau de système grammatical conséquent en commun (ce qui rend l'approche « rentable » au regard de notre premier facteur mentionné au paragraphe précédent), tout en ayant des spécificités notables. Son approche se fonde sur le développement d'une *méta-grammaire*, description-source à partir de laquelle des grammaires TAG lexicalisées (LTAG) sont engendrées automatiquement.

Notre approche est sensiblement différente dans la mesure où nous souhaitons factoriser non seulement une description en amont, mais également les grammaires elles-mêmes. Une application « chargeant » la grammaire du martiniquais, par exemple, pour effectuer des tâches de génération, commence concrètement par charger les fichiers de grammaire du « créole commun », avant de charger la couche spécifiquement martiniquaise. Cette économie supplémentaire en termes de quantité de structures stockées est rendue possible par le degré de correspondance

[2] On parle aujourd'hui de « langues peu dotées » pour ne pas mêler le facteur du nombre de locuteurs d'une langue à celui du nombre de linguistes payés pour la décrire ; notons que ce qui nous intéresse en particulier ici sont les « *familles de langues* peu dotées ».

élevé entre les systèmes.

Il convient de mentionner en outre que les quatre créoles auxquels nous nous intéressons ont en commun non seulement de grands pans de leur système grammatical, mais également une partie importante de leur lexique (comme l'illustrent les tableaux de la partie § 3). Il est donc doublement pertinent de chercher à factoriser non pas une méta-structure, mais des structures du niveau de base (concrètement, des fichiers de grammaire contenant des schémas d'arbres élémentaires, mais également des fichiers de lexique).

Le travail que nous présentons ici s'inscrit dans le prolongement d'un effort de description formelle et de modélisation des langues régionales créoles de la zone des Antilles et de la Guyane, dont les premières étapes avaient déjà été exposées dans une communication (Vaillant, 2003) qui ne portait encore que sur le créole martiniquais. Ces travaux s'appuient sur le modèle des grammaires à adjonction d'arbres, à unification de structures de traits (FS-TAG) (Vijay-Shanker & Joshi, 1988; Abeillé, 1993), qui présentent l'avantage de permettre une modélisation de phénomènes linguistiques de portées diverses ; et d'offrir un ensemble d'outils et de ressources pour la représentation, l'analyse, et la génération. Pour nos propres tests, nous utilisons jusqu'à présent une implantation *ad hoc* en Prolog (cf. § 4).

La question centrale impliquée par le présent travail est la suivante : à quel niveau doivent s'articuler les parties communes et les parties spécifiques de la description linguistique, à partir du moment où l'on décide de factoriser les langues ? Afin d'obtenir une réelle factorisation des structures du niveau de base de la description, nous avons fait le choix de modéliser le paramètre de la langue sous la forme d'un trait descriptif inclus dans les structures de traits de la grammaire, au même titre que le nombre ou le degré de détermination (cf. plus bas, fig. 1 à 3). Ceci implique que les structures communes ne sont pas des « méta-arbres » virtuels, mais des éléments du modèle TAG exactement au même niveau que les autres — à ceci près que le paramètre de langue n'y est pas instancié.

## 3 Application à la famille des créoles français de l'aire américano-caraïbe

La famille de dialectes à laquelle nous appliquons la méthode décrite ici est la famille des créoles à base française de la zone américano-caraïbe. Nous nous focalisons ici sur quatre dialectes principaux dans le monde contemporain : le créole haïtien, langue nationale de la République d'Haïti, dans les Grandes Antilles ; le créole guadeloupéen, et le créole martiniquais, parlés respectivement dans les îles de Guadeloupe et de Martinique, dans les Petites Antilles ; enfin, le créole guyanais, dont la variante centrale décrite ici est parlée dans l'Est de la Guyane Française, sur la côte nord du continent sud-américain. Dans chacun de ces territoires, il existe des variantes, que nous négligeons pour nous concentrer sur la variété de la capitale[3].

La question de la parenté génétique des créoles est une question théorique non-triviale, qui fait intervenir des considérations non seulement de linguistique comparative, mais également de sociolinguistique historique. Un premier type d'argument consisterait à poser que les créoles sont nécessairement apparentés, puisqu'ils dérivent tous du français ; quoiqu'ayant des fondements historiques, cet argument est contestable dans la mesure où si les créoles dérivent effectivement du français, ils en dérivent d'une manière atypique par rapport aux critères habituels de la linguistique historique. En effet, à la source de leur évolution se trouve, à un point dans le temps,

[3] Cette famille comprend aussi les créoles français parlés sur d'autres îles des Petites Antilles, ainsi qu'un créole résiduel parlé par quelques milliers de familles en Louisiane, que nous n'incluons pas dans notre étude faute de ressources descriptives suffisantes à l'heure actuelle.

une masse parlante constituée en majorité de locuteurs dont la langue cible (ici le français) n'est pas la langue maternelle, et par qui elle a été imparfaitement apprise[4]. Un second débat oppose des théories de monogenèse et de polygenèse des différentes langues créoles : y a-t-il pu y avoir un « proto-créole », à l'origine des langues créoles contemporaines ? Nous ne pouvons entrer dans le détail de ces discussions, qui sont périphériques au thème de cette communication (pour une présentation des discussions sur les créoles, leurs origines et leur rapport au français, nous renvoyons à (Hazaël-Massieux, 2002)). Disons simplement si les différentes théories postulant une origine commune pour *tous* les créoles (ceux des Antilles comme ceux de l'Océan Indien) ne trouvent aujourd'hui que peu de défenseurs, il est en revanche très généralement acquis, parmi les linguistes spécialistes du champ, que les créoles français de l'aire américano-caraïbe forment une famille génétique, à cause de la quantité considérable de convergences dans leurs systèmes grammaticaux, qui ne sont justement en aucune manière explicables par une parenté commune au français. Pfänder (2000, p.192–209) propose une analyse de cette famille en termes d'aire dialectale, opposant centre (Antilles) et périphérie (Louisiane et Guyane), et en donne des tableaux comparatifs pour le système d'expression du temps et de l'aspect.

Nous n'entrerons pas ici dans une description détaillée de la structure du système grammatical de chaque créole. Une présentation synthétique des principales caractéristiques structurales de la famille a été donnée dans (Vaillant, 2003), et illustrée par des exemples en créole martiniquais (mais totalement généralisables aux trois autres dialectes qui font l'objet du présent article).

Dans les tableaux qui suivent, compilés à partir de diverses sources (notamment (Pfänder, 2000) et (Damoiseau, 2007) pour la perspective comparative, mais également plusieurs autres ouvrages pour les points précis de description de chaque dialecte pris individuellement) ; puis complétés et modifiés par nous après des observations sur des corpus récents ; nous nous contentons de montrer des exemples de comparaison de sous-systèmes grammaticaux. Nous présentons ici, pour leur exemplarité, le système des degrés de détermination dans le groupe nominal, ainsi que le système noyau d'expression du temps, du mode et de l'aspect (par des particules) dans le groupe prédicatif. Les deux tableaux présentés (1 et 2) suffiront à donner une idée du degré de similitude lexicale et syntaxique de ces dialectes (N.B. il n'y a pas de morphologie flexionnelle, car ces langues sont isolantes).

**Détermination dans le groupe nominal**

Les créoles considérés possèdent tous quatre degrés de détermination systématiques du nom : un générique, un indéfini, un spécifique, et un démonstratif. Le degré générique exprime le concept pris dans ses caractéristiques générales de catégorie ; en français, il pourrait se traduire indifféremment par un singulier ou un pluriel : *zwazo gen dè zèl* (haït.), l'oiseau a deux ailes / un oiseau a deux ailes / les oiseaux ont deux ailes. Pour des raisons d'économie descriptive, dans la formalisation, nous traitons ce degré générique comme une valeur sémantique possible de l'expression du degré indéfini au pluriel (qui s'exprime lui aussi par le nom seul sans article). Le degré indéfini s'exprime comme en français à l'aide d'un adjectif numéral, qui a gardé une valeur plus spécialisée, plus proche de son origine numérale, qu'en français contemporain. Le degré spécifique (plus fortement déterminé que le défini français) s'exprime par un article postposé, qui dérive historiquement d'un adverbe déictique. Enfin, le degré démonstratif dérive d'une combinaison impliquant un ancien pronom démonstratif, qui se retrouve tantôt préposé (guya.), tantôt postposé au nom (autres créoles), et auquel se rajoute la marque du défini spécifique (qui s'y est amalgamée dans le cas des créoles guadeloupéen et martiniquais). Le

[4]Cette considération conduit certains linguistes à aller jusqu'à refuser de parler de transmission génétique entre langue source et créole, p.ex. (Thomason & Kaufman, 1988, p. 152).

pluriel s'exprime soit par un marqueur préposé dérivé d'un ancien démonstratif pluriel (mart., guad.), soit par un pronom personnel pluriel postposé (haït., guya.), qui en guyanais s'est amalgamé à la marque du défini (*yé la* [forme historique, décrite en 1872] ¿ *ya*).

Dans la formalisation, nous ne retenons donc plus que trois degrés de détermination (indéfini, spécifique et démonstratif), qui se croisent aux deux valeurs de nombre (singulier et pluriel). Par ailleurs, étant donné que la marque du degré indéfini ne se combine avec aucune autre, alors qu'il y a par ailleurs une combinaison entre les marques du démonstratif et du spécifique (avec démonstratif $\Rightarrow$ spécifique), nous modélisons le degré indéfini par une absence de trait, alors que le spécifique se caractérise par le trait $\langle \text{spe} = + \rangle$, et le démonstratif par la combinaison de traits $\langle \text{spe} = + \rangle, \langle \text{dem} = + \rangle$ .

Certains dialectes présentent un phénomène d'assimilation nasale qui change la forme de surface de l'article spécifique postposé (haït., mart., guya.) ; d'autres, en outre, changent aussi la forme de l'article selon que le mot précédent se termine par une consonne ou par une voyelle (haït., mart.). Les quatre combinaisons possibles sont indiquées dans le tableau 1.

| | | **haït.** | **guad.** | **mart.** | **guya.** | **français** |
|---|---|---|---|---|---|---|
| Degré générique | | *moun* | *moun* | *moun* | *moun* | personne (humaine) |
| Singulier | indéfini | *yon moun* | *on moun* | *an moun* | *roun moun* | une personne |
| | spécifique | *moun nan* | *moun la* | *moun lan* | *moun an* | la personne |
| | | *tab la* | *tab la* | *tab la* | *tab a* | la table |
| | | *chyen an* | *chyen la* | *chyen an* | *chyen an* | le chien |
| | | *zwazo a* | *zozyo la* | *zwézo a* | *zozo a* | l'oiseau |
| | démonstratif | *moun sa a* | *moun lasa* | *moun tala* | *sa moun an* | cette personne |
| | | *tab sa a* | *tab lasa* | *tab tala* | *sa tab a* | cette table |
| Pluriel | indéfini | *moun* | *moun* | *moun* | *moun* | des personnes |
| | spécifique | *moun yo* | *sé moun la* | *sé moun lan* | *moun yan* | les personnes |
| | | *tab yo* | *sé tab la* | *sé tab la* | *tab ya* | les tables |
| | | *chyen yo* | *sé chyen la* | *sé chyen an* | *chyen yan* | les chiens |
| | | *zwazo yo* | *sé zozyo la* | *sé zwézo a* | *zozo ya* | les oiseaux |
| | démonstratif | *moun sa yo* | *sé moun lasa* | *sé moun tala* | *sa moun yan* | ces personnes |
| | | *tab sa yo* | *sé tab lasa* | *sé tab tala* | *sa tab ya* | ces tables |

TAB. 1 – Détermination dans le groupe nominal

Notons que certaines portions du système présenté ici connaissent en pratique des variations dans l'usage oral, notamment en ce qui concerne le degré de détermination spécifique, dans les régions où le bilinguisme créole / français est la règle. Nous n'avons pas tenté de prendre en compte ces variations dans le présent état de notre modélisation.

**Temps et aspect dans le groupe prédicatif**

Une description classique donnée, en créolistique, du système TMA (Temps, Mode, Aspect) des langues créoles « atlantiques »[5] fait état d'un système à trois composantes s'ordonnant suivant un ordre strict : marquage (ou non) d'un temps passé ; marquage (ou non) d'un mode pouvant avoir des emplois de futur ou d'irréel, selon les contextes ; marquage (ou non) d'un aspect imperfectif. Une version canonique de ce système est décrite en français, par exemple, par Valdman (1978), qui définit ces trois catégories comme une catégorie de temps (*passé* ), et deux catégories d'aspect (*prospectif* et *continuatif* ). La marque du milieu (celle que Valdman appelle prospectif) prend une valeur d'irréel lorsqu'elle est combinée avec un passé.

On a donc une combinatoire (*té* / $\varnothing$) $\times$ (*ké* / $\varnothing$) $\times$ (*ka* / $\varnothing$) (en désignant ces marques par les

[5] On retrouve ce schéma dans d'autres langues créoles à base anglaise.

formes sous lesquelles elles se manifestent dans les trois créoles guadeloupéen, martiniquais et guyanais), qui engendre théoriquement les huit combinaisons possibles : ∅, *ka*, *ké*, *ké ka*, *té*, *té ka*, *té ké*, *té ké ka*. Ces huit combinaisons sont attestées à différents degrés, avec les valeurs précisées dans le tableau 2. En créole haïtien, les formes correspondantes sont *te*, *va* et *ap*, et les combinaisons présentent des formes contractées[6].

Ce schéma de base connaît des variations. Ainsi le terme d'aspect « imperfectif » recouvre-t-il un complexe de valeurs diverses (progressif, fréquentatif, ou simplement inaccompli) qui se manifestent différemment dans les différents dialectes considérés. Ainsi, si la marque *ka* recouvre tous ces aspects en créole guadeloupéen ou martiniquais (jusqu'à pouvoir prendre la valeur temporelle générale correspondant au présent français), il n'en va pas nécessairement de même en créole guyanais, et encore moins en haïtien (ou l'inaccompli est non-marqué, et où la valeur aspectuelle la plus remarquable de la particule *ap* est le progressif — qui prend fréquemment une valeur temporelle de futur). Le tableau 2 prend en compte ces différences.

| | **haït.** | **guad.** | **mart.** | **guya.** |
|---|---|---|---|---|
| Accompli / Aoriste | *danse* | *dansé* | *dansé* | *dansé* |
| Inaccompli / Présent | *danse* | *ka dansé* | *ka dansé* | *(ka) dansé* |
| Fréquentatif | *danse* | *ka dansé* | *ka dansé* | *ka dansé* |
| Progressif | *ap danse* | *ka dansé* | *ka dansé* | *ka dansé* |
| Futur proche | *pral danse* | *kay dansé* | *kay dansé* | *k'alé / kay dansé* |
| Futur | *va danse* | *ké dansé* | *ké dansé* | *ké dansé* |
| Futur inaccompli *(très peu fréq.)* | *vap danse* | *ké ka dansé* | *ké ka dansé* | *ké ka dansé* |
| Passé accompli (plus-que-parfait) | *te danse* | *té dansé* | *té dansé* | *té dansé* |
| Passé inaccompli | *tap danse* | *té ka dansé* | *té ka dansé* | *té ka dansé* |
| Irréel | *ta danse* | *té ké dansé* | *té ké dansé* | *té ké dansé* |
| Irréel inaccompli | *ta vap danse* | *té ké ka dansé* | *té ké ka dansé* | *té ké ka dansé* |
| Conditionnel / Optatif | *ta danse* | *té ké dansé* | *sé dansé* | *té ké dansé* |

TAB. 2 – Marquage du temps et de l'aspect dans le groupe prédicatif

Dans les figures 1 et 2, nous montrons les éléments principaux de la modélisation des données linguistiques du groupe nominal, exposées dans le tableau 1, sous forme de TAGs intégrant un paramètre de langue $l$ [7].

On notera que les arbres *Det Dem (gp,mq)* et *Plur (gp,mq)*, qui concernent deux dialectes parmi les quatre (guadeloupéen et martiniquais), figurent dans la couche générale sans risque d'interférer avec la construction du démonstratif ou du pluriel en haïtien ou en guyanais (les contraintes d'unification ne permettent pas l'adjonction d'un démonstratif antillais sur un démonstratif haïtien ou guyanais, ni celle d'un pluriel antillais sur un pluriel haïtien ou guyanais).

L'adjonction du démonstratif en haïtien ou en guyanais se fait bien au-dessus du niveau des compléments de nom (attention au paramètre *bar* dans les arbres *Dem (gf)* et *Dem (ht)*), mais en-dessous de l'article spécifique ; ex. *moun Sentoma sa yo* (haït.) : les gens de Saint-Thomas ; *sa moun Senloran yan* (guya.) : les gens de Saint-Laurent.

Le système TMA, quant à lui, possède des mécanismes en grande partie communs : les arbres auxiliaires qui modélisent l'adjonction de valeurs aspectuelles ou temporelles sont donc tous

[6] va ap ¿ vap ; te ap ¿ tap ; te va ¿ ta ; te va ap ¿ ta vap.

[7] Les abréviations suivantes sont utilisées pour les attributs : *bar* = niveau de barre (1 = nom avec compléments, mais sans détermination ; 2 = groupe nominal) ; *nbr* = nombre ; *spe* = déterminant spécifique ; *dem* = déterminant démonstratif ; *cns* = le constituant se termine par une consonne ; *nas* = le constituant se termine par une syllabe nasale ; *lan* = langue. Les valeurs utilisés pour identifier les quatre créoles sont basées sur les codes bilittéraux de la norme ISO-3166 pour les noms de pays : HT pour *Haïti*, GP pour *Guadeloupe*, MQ pour *Martinique*, et GF pour *Guyane Française*. Les variables non-instanciées sont notées en italique.

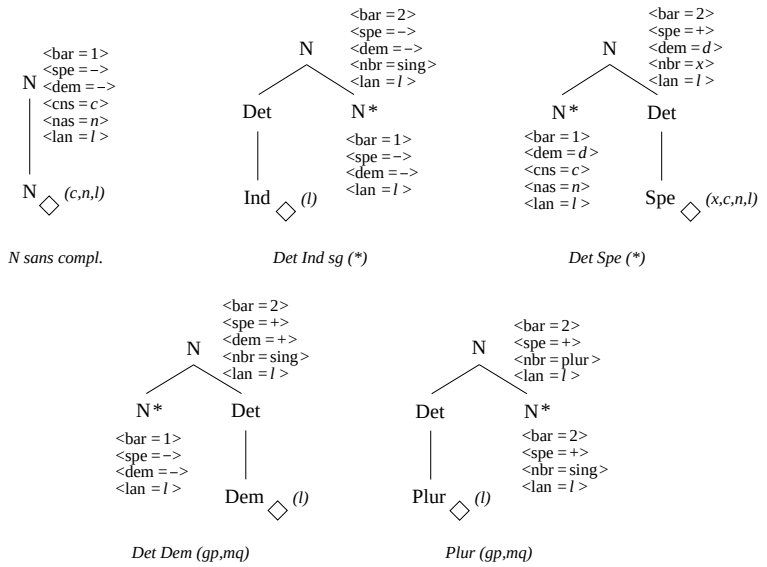


FIG. 1 – Éléments communs de modélisation du groupe nominal.

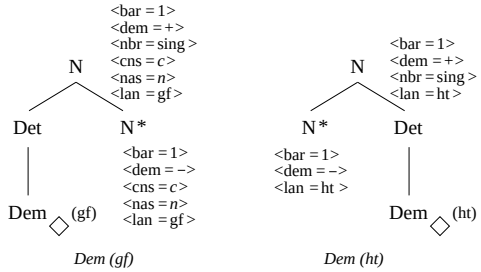


FIG. 2 – Éléments de modélisation du groupe nominal spécifiques au haïtien et au guyanais

communs (fig. 3), à la nuance près que celui qui rend compte de valeurs générales de l'aspect imperfectif (duratif, fréquentatif) ne s'unifie pas lorsque le paramètre de langue indique le créole haïtien. Ce sont ensuite les variables lexicales qui font la différence entre les dialectes[8].

## 4 Implantation informatique

L'approche proposée a été testée avec une petite implantation des FS-TAG en PROLOG[9], développée à l'origine pour une autre application (Vaillant, 1999) et pour la langue allemande, et adaptée ultérieurement au créole martiniquais (Vaillant, 2003). Cette grammaire est testée en génération.

Les arbres syntaxiques sont des structures récursives composées de triplets ⟨ cat, liste_ss_arbres, traits ⟩. L'élément liste_ss_arbres peut consister soit en une liste de triplets ayant à leur tour la même forme, soit en une étiquette lexicale (cas des nœuds feuilles). Les familles d'arbres sont gérées à la volée par l'unification d'étiquettes lexicales dans les arbres schéma.

Les opérations de substitution et d'adjonction sont définies comme des prédicats Prolog manipulant des arbres. Chaque nœud substituable, dans un arbre, est porteur d'une étiquette ; celle-ci doit être fournie en argument à l'opération de substitution pour qu'elle « sache » à quel endroit effectuer l'unification. La liste des correspondances entre nœuds à substituer et rôles dans la structure argumentale du lexème prédicatif est gérée par une liste qui est intégrée dans la

[8]Les abréviations suivantes sont utilisées pour les attributs dans la fig. 3 : Temps : *pas* = passé ; Aspects : *psp* = prospectif ; *prx* = prospectif proximal (aspect imminent ˜ valeur temporelle de futur proche) ; *imp* = imperfectif ; *prg* = progressif.

[9]Environnement utilisé : SWI-Prolog, développé par Jan Wielemaker, Université d'Amsterdam : `http://www.swi-prolog.org`.

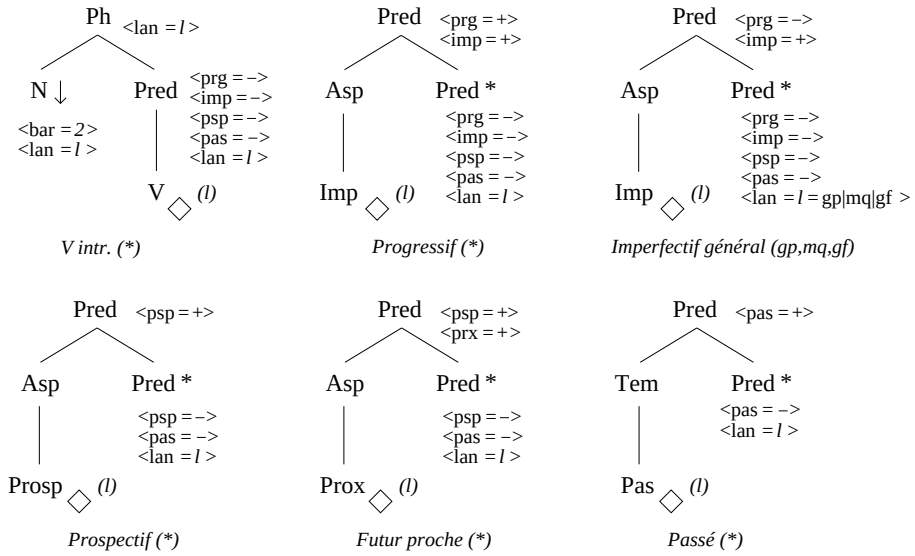


FIG. 3 – Éléments communs de modélisation du groupe prédicatif

description de ces lexèmes sous forme d'arbres élémentaires.

Pour l'adjonction en revanche, la structure argumentale n'est pas disponible dans la description de l'arbre initial (mais dans celle de l'arbre auxiliaire) : il n'existe pas de spécification précise *a priori* du nœud de l'arbre initial auquel l'adjonction doit être réalisée (puisque tout arbre initial se suffit à lui-même, et peut se passer d'adjonction). Afin de situer le lieu d'adjonction le plus adéquat, nous avons programmé une opération d'adjonction « à proximité de ». Cette opération repère, dans l'arbre initial sur lequel doit se faire l'adjonction, le nœud feuille correspondant au lexème argument, et tente d'adjoindre l'arbre auxiliaire « au plus bas », c'est-à-dire au nœud de catégorie compatible situé le plus bas possible sur le chemin allant du lexème feuille à la racine de l'arbre initial. Cette façon de procéder évite que des déterminants ne se branchent à un niveau trop élevé dans l'arbre.

Le principe de base du générateur est le suivant : le graphe sémantique est parcouru à partir d'un nœud, et le programme tente d'engendrer un arbre initial correspondant à l'expression du lexème correspondant et de sa structure argumentale. Si une partie du graphe sémantique reste non-générée après cette première passe directe, le programme essaye de l'exprimer sous forme d'arbre auxiliaire et de l'adjoindre à l'arbre déjà existant. Ce principe est appliqué récursivement.

En ce qui concerne l'application de notre approche factorisée : celle-ci est simplement réalisée par l'ajout de la contrainte d'une valeur précise du paramètre *langue* au niveau du résultat souhaité. La figure 4 illustre l'application de cette méthode.

Le gain en termes de quantité de représentations grammaticales peut ici être évalué de manière grossière par le rapport de volume entre l'ensemble des fichiers utilisés dans la version quadrilingue de la modélisation, et l'ensemble des fichiers utilisés pour la version martiniquaise seulement : ce rapport est d'environ 210 %[10]. En partant du postulat (pour simplifier) selon lequel chaque lexique-grammaire unilingue aurait environ la même taille, notre modélisation permet donc de passer d'un facteur 4 à un facteur 2,1, soit un gain de 48 %.

[10]Cette évaluation est biaisée par le fait qu'outre le passage à un travail de modélisation pluridialectal, nous avons également augmenté la quantité de phénomènes grammaticaux décrits ; elle est donc probablement encore surévaluée par rapport à ce qu'elle pourrait être dans le cas idéal. Afin de proposer une évaluation plus rigoureuse du gain, nous devrions « mettre à jour » la version martiniquaise seulement de façon à ce qu'elle ait exactement la même couverture que la version factorisée actuelle.

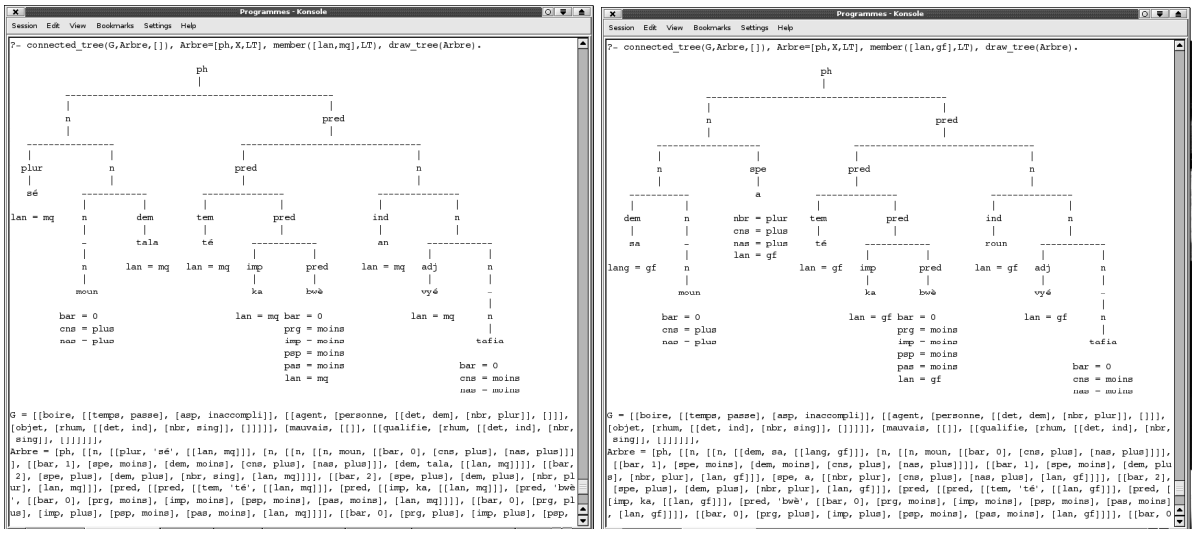

FIG. 4 – Deux exemples de phrases engendrées par le système à partir d'un graphe sémantique (G) : une phrase en créole martiniquais (« Sé moun-tala té ka bwè vyé tafia ») et la phrase correspondante en créole guyanais (« Sa moun-yan té ka bwè vyé tafia »). Sans contrainte sur le trait langue (appel au prédicat '`connected_tree(G,Arbre,[]).`'), quatre phrases sont générées (correspondant aux quatre dialectes représentés). Pour limiter la génération à une seule langue, on impose une contrainte sur le trait langue au niveau le plus élevé du graphe, qui empêche l'unification avec les autres possibilités (en appelant par exemple '`connected_tree(G,Arbre,[]), Arbre=[ph,X,LT], member([lan,mq],LT).`').

## 5 Discussion et Conclusion

Le choix de modélisation décrit plus haut (§ 2) peut sembler contre-intuitif si l'on raisonne suivant les habitudes de la linguistique structurale, pour laquelle la langue est par définition le système englobant les différentes catégories, et ne saurait donc elle-même constituer l'une des catégories. Il se justifie pourtant en pratique, par sa capacité à cristalliser les structures communes de différents dialectes proches : cela revient alors à considérer ces derniers comme des sous-systèmes d'un méta-système plus générique.

La modélisation du système d'un locuteur monolingue s'obtient alors, tout simplement, en instanciant le paramètre langue à l'une de ses valeurs possibles, en écrémant les arbres qui ne peuvent pas s'unifier avec cette valeur, puis en effaçant le paramètre — devenu redondant — de toutes les descriptions. On peut donc, par cette procédure, spécifier la description de la famille de dialectes pour retomber sur la description, plus classique, d'une langue unique et homogène.

Il est utile de remarquer, par ailleurs, que notre modèle de description « à couches » apporte potentiellement un aperçu intéressant de ce que pourrait être le système linguistique interne (*I-Language*) d'un locuteur multilingue alternant entre ces différents dialectes. De tels locuteurs existent, dans certaines situations sociolinguistiques : ce n'est pas généralement le cas des Antillais de Guadeloupe ou de Martinique (qui sont plutôt bilingues créole / français), mais c'est le cas des membres des communautés haïtiennes de Guadeloupe et de Guyane (plusieurs dizaines de milliers de personnes dans chaque cas), et sans doute de ceux des communautés antillo-guyanaises des grandes agglomérations de métropole (où une nouvelle *koinè* pan-créole peut se constituer). La théorie TAG a déjà été utilisée pour la modélisation de processus cognitifs (par exemple sur la question de l'acquisition de la langue maternelle, par Frank (2000)). Nous l'utilisons ici (bien que l'utilisation d'autres formalismes eût également été possible) dans un contexte qui rend possible d'aborder ce type de questions. La modélisation de la grammaire

du *I-Language* d'un sujet bilingue peut constituer un sujet d'étude passionnant, si l'on accepte de lever l'hypothèse simpliste des répertoires cloisonnés[11]. Cela ne constitue bien sûr pas notre propos ici, mais cela montre un intérêt théorique potentiel du type de travaux exposé, et un cas possible de retournement de point de vue — dans lequel le choix de considérer la langue comme un paramètre alternant dans un méta-système passerait du statut de *bug* à celui de *feature* .

## Références


ABEILLÉ A. (1993). *Les nouvelles syntaxes*. Paris: Armand Colin.

CANDITO M.-H. (1998). Building parallel LTAG for French and Italian. In *Proceedings of the 17$^{th}$ International Conference on Computational Linguistics (COLING 1998)*, p. 211–217, Montréal (Québec, Canada).

CHOMSKY N. (1995). *The Minimalist Program*. Current studies in linguistics. Cambridge (Massachussets, É.-U.-A.): MIT Press.

DAMOISEAU R. (2007). Le créole guyanais dans la famille des créoles à base lexicale française de la zone américano-caraïbe. In S. MAM-LAM-FOUCK, Ed., *Comprendre la Guyane d'aujourd'hui*, p. 501–514, Matoury (Guyane Française): Ibis Rouge.

FRANK R. (2000). From regular to context-free to mildly context-sensitive tree rewriting systems: The path of child language acquisition. In A. ABEILLÉ & O. RAMBOW, Eds., *Tree Adjoining Grammars: Formalisms, Linguistic Analysis and Processing*, p. 101–120, Stanford (Californie, É.-U.-A.): CSLI Publications.

HAZAËL-MASSIEUX M.-C. (2002). Les créoles à base française : une introduction. *TIPA (Travaux Interdisciplinaires du laboratoire Parole et Langage d'Aix-en-Provence)*, **21**, 63–86. Disponible en ligne : `http://aune.lpl.univ-aix.fr/lpl/tipa/21/tipa21-hazael.pdf`.

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

THOMASON S. & KAUFMAN T. (1988). *Language Contact, Creolization, and Genetic Linguistics*. Berkeley (Californie, É.-U.-A.): University of California Press.

VAILLANT P. (1999). *Learning to Communicate in German through an Iconic Input Interface: Presentation of the GLOTTAI Project*. Rapport interne, Humboldt Universität zu Berlin. Disponible en ligne : `http://www.vaillant.nom.fr/pascal/glottai/presentation/`.

VAILLANT P. (2003). Une grammaire formelle du créole martiniquais pour la génération automatique. In B. DAILLE, Ed., *Actes de TALN 2003 (Traitement automatique des langues naturelles)*, p. 255–264, Batz-sur-mer: ATALA IRIN.

VALDMAN A. (1978). *Le créole : structure, statut et origine*. Paris: Klincksieck.

VIJAY-SHANKER K. & JOSHI A. (1988). Feature-structure based Tree Adjoining Grammars. In *Proceedings of the 12$^{th}$ International Conference on Computational Linguistics (COLING 1988)*, p. 714–719, Budapest (Hongrie).


[11] Une thèse sur ce sujet est en cours à l'Université des Antilles-Guyane (Jocelyne Litou).