# Décrire informatiquement une langue naturelle : application à quelques langues d'Afrique

Denys Duchier, Nicola Lampitelli, Brunelle Magnana Ekoukou, Yannick Parmentier, Simon Petitjean, Emmanuel Schang

# Décrire informatiquement une langue naturelle : application à quelques langues d'Afrique

Denys Duchier[(1)], Nicola Lampitelli[(2)], Brunelle Magnana-Ekoukou[(2)], Yannick Parmentier[(1)], Simon Petitjean[(1)], Emmanuel Schang[(2)]

(1) Laboratoire d'Informatique Fondamentale d'Orléans (LIFO)

(2) Laboratoire Ligérien de Linguistique (LLL)

**Résumé**

Cet article présente une description informatisée de problèmes linguistiques dans quelques langues d'Afrique (ikota et somali principalement). Après avoir exposé un rapide état de l'art sur la question et après avoir discuté les enjeux de l'usage des nouvelles technologies sur les langues disposant de peu de ressources numériques, nous décrivons l'usage que nous faisons de l'outil XMG (eXtensible MetaGrammar) pour la génération des formes verbales de l'ikota, pour le traitement du groupe nominal en somali, et pour la construction d'une grammaire électronique du créole de São-Tomense. Nous fournissons des éléments de comparaison avec le français lorsque cela paraît nécessaire. Nous concluons en présentant comment notre approche permet d'amorcer des travaux de constitution de ressources pour des langues peu dotées.

## 1 Introduction

Le travail présenté ici se place dans une longue lignée de travaux en Traitement Automatique des Langues (TAL) visant à permettre à l'ordinateur de manipuler les langues produites par l'être humain. Ici, le terme *manipuler* est à interpréter dans un sens large, dans la mesure où il englobe un grand nombre d'applications informatiques allant de tâches relativement simples telles que la correction orthographique à des tâches complexes comme la traduction automatique. Suivant la tâche considérée, une description plus ou moins fine de la langue est nécessaire, afin de fournir à l'ordinateur l'ensemble des informations utiles. Par exemple, si

l'on considère la correction orthographique, l'ordinateur doit avoir accès à un lexique [1] le plus exhaustif possible, et à un ensemble de règles morphologiques et grammaticales représentant des phénomènes linguistiques contraignant l'orthographe des mots (par exemple le pluriel ou l'accord sujet-verbe pour le français).

En plus de permettre le développement d'applications spécifiques, disposer d'une description (appelée également représentation) informatique [2] de la langue présente de nombreux avantages :

1. Tout d'abord, une telle représentation permet d'**exprimer des généralisations** sur les structures linguistiques (par exemple, sous forme de règles morphologiques ou syntaxiques abstraites) ;
2. elle permet de **vérifier une théorie linguistique** en la mettant en œuvre sur un grand nombre de structures linguistiques ;
3. elle permet de plus aisément **confronter une théorie linguistique aux données de terrain** (en vérifiant automatiquement si une description associée à une théorie linguistique donnée, est compatible avec la structure des données de terrain) ;
4. enfin, une telle représentation permet (comme nous allons le voir dans cet article) de **créer à moindre coût** [3] **des ressources linguistiques**, et est de ce fait particulièrement utile pour les langues peu dotées (c'est-à-dire pour lesquelles il n'existe pas ou peu de ressources telles que lexiques, grammaires, corpus, etc. disponibles).

Dans ce contexte, cet article vise (i) à présenter un cadre informatique permettant la description de ressources langagières pour différents niveaux linguistiques (en l'occurrence morphologie et syntaxe), mais aussi pour différentes théories formelles (morphologie dérivationnelle ou flexionnelle par exemple), ainsi que (ii) l'application de ce cadre à diverses langues d'Afrique.

Notre présentation repose sur le plan suivant. En section 2, nous présentons ce que l'on entend habituellement par *décrire une langue* dans le cadre des théories formelles. Nous porterons une attention particulière à la syntaxe et à la morphologie, car ces deux niveaux sont au cœur de

1. En première approximation, ce lexique correspond à une liste de mots.
2. On parle souvent de description (grammaire, lexique) électronique.
3. En effet, développer des ressources linguistiques manuellement représente un coût énorme, à titre d'exemple, on parle de dizaines d'hommes-années pour la grammaire d'arbres adjoints de l'anglais (XTAG Research Group, 2001).

nombreuses applications de TAL. En section 3, nous montrerons comment un cadre formel tel que celui offert par le langage informatique *eXtensible MetaGrammar* (XMG) permet de décrire des ressources linguistiques variées (syntaxiques ou morphologiques, pour diverses langues d'Afrique) de manière uniforme. Enfin, en section 4, nous concluons et présentons quelques perspectives pour des travaux futurs.

# 2 Décrire une langue dans le contexte des théories formelles

En TAL, les théories linguistiques formelles (c'est-à-dire reposant sur un modèle mathématique) ont la primauté, car elles se prêtent naturellement à une mise en œuvre informatique. C'est ainsi que des travaux tels que ceux de Chomsky (1957) en syntaxe générative ou Montague (1974) sur l'interface syntaxe / sémantique ont été à l'origine des premières réalisations logicielles de TAL permettant l'analyse automatique de phrases.

Dans cette section, nous allons dans un premier temps préciser ce que nous entendons généralement par *décrire formellement la syntaxe de la langue* (§ 2.1). Nous ferons ensuite de même avec la morphologie (§ 2.2), avant de présenter les avantages et limites des descriptions formelles (§ 2.3).

## 2.1 Décrire la syntaxe de la langue

Décrire la syntaxe d'une langue répond à deux objectifs distincts et complémentaires. Le premier consiste à établir un système de règles permettant de décider, au bout d'un temps fini, si une phrase S testée est ou n'est pas une phrase grammaticale de la langue L. Le second objectif consiste à fournir une représentation de cette phrase S explicitant sa structure et à partir de laquelle on peut interpréter le sens de S. Un arbre est une façon de représenter cette structure, mais d'autres formes sont possibles (les graphes notamment).

Ces deux objectifs sont toutefois liés et le linguiste de terrain opère les deux tâches simultanément.

## 2.2 Décrire la morphologie de la langue

A la suite de Hockett (1954), trois modèles d'analyse morphologique sont distingués. Le premier, appelé *Item-and-Arrangement* (IA), dérive di-

rectement de l'approche de la linguistique structuraliste consistant à analyser les mots en tant que séquences linéarisées de morphèmes lexicaux et fonctionnels. Les morphèmes (ou *items*) sont donc assemblés (cf. anglais *arranged*) les uns aux autres afin d'obtenir la forme de surface d'un mot donné. Par exemple, le mot anglais *dogs* 'chiens' est le résultat de l'enchaînement du morphème lexical DOG, signifiant 'chien', et le morphème flexionnel -Z, exprimant le trait [pluriel]. Le deuxième modèle, appelé *Item-and-Process* (IP), conçoit le mot comme le résultat d'une opération morphologique s'applicant à une forme de base (le lexème). Ainsi, le mot *dogs* est le résultat d'une opération telle que /X/ → /Xz/ (X = le lexème DOG ). Le pluriel *dogs* est donc formé d'un seul élément morphologique. Dans son article, Hockett mentionne brièvement un troisième modèle, *Word-and-Paradigm* (WP), qui consiste, comme le nom l'indique, en une approche favorisant les relations paradigmatiques entre les formes fléchies d'un mot donné. L'idée centrale du modèle WP est le refus de toute tentative de construction de la forme et du sens des mots de manière cumulative. Autrement dit, *dogs* a du sens en tant que mot occupant la case 'pluriel' dans le paradigme du mot 'chien'.

Au sein des premiers travaux en grammaire générative développés autour de (Chomsky, 1957), la morphologie n'occupe pas une place centrale. En effet, dans le modèle de la Grammaire Universelle proposé par Chomsky, la syntaxe est la seule composante générative et transformationnelle. Ce n'est qu'à partir de (Chomsky et Halle, 1968) que la grammaire générative se penche sur les questions liées à la forme des mots et au rapport entre le signifiant et le signifié lors du processus de formation des mots. Bien que développant principalement une théorie phonologique, Chomsky et Halle posent donc les bases d'une approche formelle à la formation des mots en adoptant implicitement un modèle IA. Nous retrouvons cette approche basée sur un modèle IA dans la théorie de la Morphologie Distribuée (MD) (Halle et Marantz, 1993; Embick, 2010) dans laquelle on fait l'hypothèse que les mots sont construits en syntaxe.

Dans l'analyse de l'ikota et du somali que nous proposons plus bas, cf. 3.2 et 3.3, nous utilisons une approche à la morphologie inspirée d'un modèle IA. En effet, nous postulons l'existence de morphèmes discrets et isolables les uns des autres à un certain niveau de l'analyse. Cependant, nous ne pouvons pas, à l'aide de notre outil informatique, prévoir une opération morpho-phonologique transformant une séquence sousjacente de morphèmes abstraits en un mot bien formé. Nous devons donc postuler

l'existence d'un système de règles ad-hoc renvoyant à une forme donnée du paradigme. [4]

Dans la section suivante, nous pointons les avantages ainsi que les limites des approches formelles.

### 2.3 Vers des représentations formelles : avantages et limites

Comme nous l'avons vu précédemment, les représentations formelles offrent l'avantage de se prêter naturellement à une mise en œuvre informatique. Suivant le formalisme utilisé, on cherche à obtenir deux bonnes propriétés : tout d'abord avoir une **expressivité** suffisante (c'est-à-dire pouvoir exprimer le plus directement possible, des phénomènes linguistiques relativement complexes, tels que l'accord sujet-verbe ou l'ordre des clitiques en français), et ensuite avoir une **complexité** aussi faible que possible (c'est-à-dire avoir des algorithmes de traitement qui se terminent dans un temps relativement court par rapport à la taille des données manipulées). [5]

Si on prend le cas de la syntaxe, à ce jour on ne connaît pas l'expressivité nécessaire pour représenter la syntaxe de l'ensemble des langues. Divers formalismes syntaxiques ont été proposés, à commencer par les grammaires hors-contextes (ou grammaires algébriques). [6] Celles-ci ne permettent pas de représenter des phénomènes de dépendances croisées (par exemple, ordre entre les arguments des prédicats dans une subordonnée en suisse allémanique (Shieber, 1985)). Par la suite, d'autres formalismes plus complexes ont été proposés, sans parvenir à un consensus sur la complexité nécessaire pour pouvoir décrire la syntaxe des langues naturelles.

Dans les travaux sur la syntaxe qui seront présentés en section 3, le formalisme qui a été utilisé correspond aux grammaires d'arbres adjoints (*Tree-Adjoining Grammar*, TAG) (Joshi et Schabes, 1997). Ces grammaires offrent une expressivité intéressante de par leurs règles élémentaires, qui

4. Cf. (Bonami et Boyé, 2006) pour une discussion sur le statut des irrégularités flexionnelles dans les paradigmes. Les modèles IA traitent les irrégularités comme des accidents du système génératif, alors que les modèle WP leur donnent un statut linguistique à part entière du fait de postuler deux ou plusieurs formes pour un morphème donné.

5. Ces deux propriétés ne sont pas indépendantes l'une de l'autre, en effet plus un formalisme est expressif, plus il est complexe à traiter.

6. Pour une présentation détaillée de ce formalisme et de ses propriétés mathématiques, voir (Autebert *et al.*, 1997).

sont des arbres permettant de décrire des contraintes entre constituents éloignés dans la phrase [7], tout en ayant une complexité raisonnable (une phrase de longueur $n$ est analysable par un programme en un temps proportionnel à $n^6$, à un facteur constant près) [8]. [9]

Si on prend le cas de la morphologie, les travaux les plus courants pour l'analyse automatique de la structure des mots, reposent sur une représentation à base de machines à états finis (Beesley et Karttunen, 2003). Dans cette approche, les règles de (trans)formation des mots sont définies au moyen de grammaires algébriques reconnues par un automate. Cette approche est satisfaisante en terme de complexité (celle-ci est polynomiale), mais souffre parfois d'un manque d'expressivité (voir § 3.2).

# 3 Application au français, à l'ikota, au somali et à certains créoles

Nous allons à présent introduire le langage formel *eXtensible MetaGrammar* (XMG) qui permet de décrire de manière factorisée et modulaire des ressources linguistiques diverses. Ce langage sera appliqué dans un premier temps à la syntaxe du français (§ 3.1), puis à la morphologie de l'ikota (§ 3.2), celle du somali (§ 3.3), et enfin à la syntaxe du créole portugais de São Tomé (sãotomense) (§ 3.4).

## 3.1 Décrire les cadres de sous-catégorisation du verbe en français

Le but de ce paragraphe est de présenter brièvement le langage XMG, en l'appliquant à la description de la syntaxe du français au moyen des grammaires TAG. [10]

Une grammaire TAG décrivant la syntaxe du français est constituée de plusieurs dizaines de milliers d'arbres élémentaires. Chacun de ces arbres représente les relations entre un prédicat et ses arguments (s'il y en a).

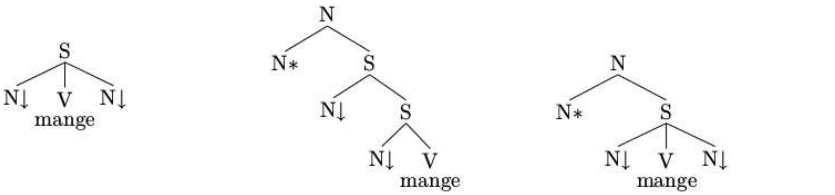
7. On parle de **domaine de localité étendu**.

8. On parle de complexité **polynomiale**.

9. Il est important de noter que ce choix de TAG est discutable, dans la mesure où ce formalisme ne permet pas de décrire l'ensemble des phénomènes syntaxiques des langues (il est par exemple difficile de décrire la négation en français avec TAG).

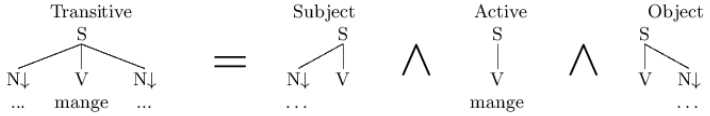
10. Pour une présentation plus détaillée de la description de la syntaxe du français en XMG, voir (Crabbé, 2005).

Ainsi, à un même prédicat, on associe un ensemble d'arbres élémentaires (appelé *famille* d'arbres) décrivant les divers usages du prédicat dans une phrase, comme illustré ci-dessous pour le prédicat *manger*.

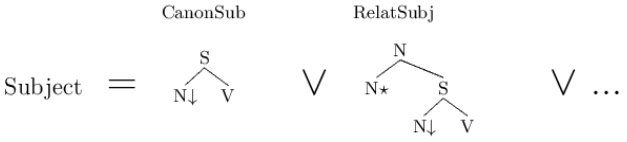



*Jean mange une pomme* *La pomme que Jean mange* *Jean qui mange une pomme* ...

L'idée sous-jacente à la création du langage XMG est de fournir des moyens de décrire des structures linguistiques par composition (conjonctive et disjonctive) d'informations. Ainsi un arbre TAG pour un verbe transitif pourrait être décrit comme la composition conjonctive de 3 unités d'information (appelées blocs) : un bloc pour le sujet, un pour la structure verbale et un pour l'objet.




De même, un bloc tel que le sujet pourrait être le résultat de la composition disjonctive de plusieurs blocs d'information (par exemple sujet canonique, sujet sous forme de relative, etc), exprimant ainsi la notion de structures alternatives à une structure canonique.




Dans ce contexte, décrire une grammaire du français revient à décrire (i) une hiérarchie de blocs élémentaires (voir Figure 1 contenant la hiérarchie des arguments verbaux de Crabbé (2005)), puis (ii) des règles de composition de ces blocs pour en décrire d'autres (voir Figure 2 pour un exemple d'une telle règle). Les blocs intégrant toute l'information nécessaire à la

description des arbres élémentaires de la grammaire TAG visée sont appelés blocs axiomes. [11]

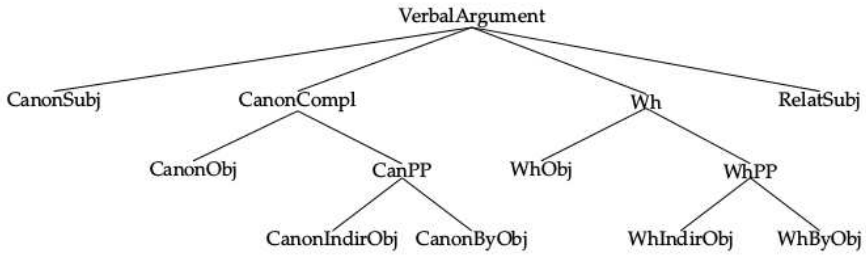



FIGURE 1 – Hiérarchie de blocs dans une métagrammaire du français

$$
\begin{array}{lcl}
Subject & \rightarrow & CanonSubj \vee RelatSubj \\
Object & \rightarrow & CanonObj \vee WhObj \\
ByObject & \rightarrow & CanonByObj \vee WhByObj \\
IndirectObject & \rightarrow & CanonIndirObj \vee WhIndirObj
\end{array}
$$

$$
\begin{array}{lcl}
Transitive & \rightarrow & (\ Subject \wedge Active \wedge Object\ ) \\
 & \vee & (\ Subject \wedge Passive \wedge ByObject\ ) \\
 & \vee & (\ Subject \wedge Passive\ )
\end{array}
$$

FIGURE 2 – Composition de blocs dans une métagrammaire du français

Le langage XMG permet donc de décrire des ressources linguistiques au moyen de deux mécanismes principaux : la possibilité de définir des *abstractions* sur des blocs d'information, et de composer ces abstractions *conjonctivement* ou *disjonctivement*, ce qui peut être représenté formellement comme suit :

$$Bloc \quad := \quad Description \mid Bloc \wedge Bloc \mid Bloc \vee Bloc$$

On remarque qu'un bloc, en plus de pouvoir correspondre à la composition de blocs, peut également correspondre à une abstraction sur une description, qui elle-même peut correspondre à un fragment d'arbre ou à d'autres types d'information (par exemple, structures de traits), comme nous allons le voir dans les sections 3.2 à 3.4.

11. Ces blocs axiomes sont passés au compilateur XMG afin de produire la grammaire TAG visée (Le Roux et Parmentier, 2005).

## 3.2 Décrire la morphologie verbale en ikota

L'ikota, encore appelé kota fait partie de la cinquantaine de langues bantoues répertoriée au Gabon. C'est une langue tonale et à classes nominales. Pour ce qui est de la morphologie verbale, nous répartissons les verbes de l'ikota dans trois groupes en fonction des suffixes verbaux car ils sont à l'origine d'un phénomène d'harmonisation vocalique. Les tableaux (1), (2) et (3) montrent la conjugaison de trois verbes à la première personne.

TABLE 1 – Conjugaison de bòʤákà ≪ manger ≫ (groupe 1)

| 1 | 2 | 3 | 4 | 5 | 6 | 7 | Valeur |
|---|---|---|---|---|---|---|---|
| m- | à- | ʤ | | | -á | | présent |
| m- | à- | ʤ | | | -á | -ná | passé d'hier |
| m- | à- | ʤ | | | -á | -sá | passé lointain |
| m- | é- | ʤ | | | -à | | passé récent |
| m- | àmò- | ʤ | | | -á | | passé moyen |
| m- | é- | ʤ | | -ák | -à | | futur moyen |
| m- | é- | ʤ | | -ák | -à | -ná | futur de demain |
| m- | é- | ʤ | | -ák | -à | -sá | futur lointain |
| m- | ábí- | ʤ | | -ák | -à | | futur imminent |

TABLE 2 – Conjugaison de bòwɛ́ʧɛ̀ ≪ donner ≫ (groupe 2)

| 1 | 2 | 3 | 4 | 5 | 6 | 7 | Valeur |
|---|---|---|---|---|---|---|---|
| m- | à- | w | | | -ɛ́ | | présent |
| m- | à- | w | | | -ɛ́ | -nɛ́ | passé d'hier |
| m- | à- | w | | | -ɛ́ | -sɛ́ | passé lointain |
| m- | é- | w | | | -ɛ̀ | | passé récent |
| m- | àmò- | w | | | -ɛ́ | | passé moyen |
| m- | é- | w | | -ɛ́ʧ | -ɛ̀ | | futur moyen |
| m- | é- | w | | -ɛ́ʧ- | ɛ̀ | -nɛ́ | futur de demain |
| m- | é- | w- | | -ɛ́ʧ | -ɛ̀ | -sɛ́ | futur lointain |
| m- | ábí- | w | | -ɛ́ʧ | -ɛ̀ | | futur imminent |

De ces tableaux, il ressort que les formes verbales de l'ikota se composent de sept classes de positions (CP) selon le modèle PFM (Stump, 2001). Une CP peut être pleine ou vide :
- La CP 1 marque l'indice sujet ;
- La CP 2 est occupée par un exposant ayant rapport au temps ;
- La CP 3 marque le stem ;

TABLE 3 – Conjugaison de bòbɔ́nɔ́kɔ̀ ≪ choisir ≫ (groupe 3)

| 1 | 2 | 3 | 4 | 5 | 6 | 7 | Valeur |
|---|---|---|---|---|---|---|---|
| m- | à- | bɔ́n | | | -ɔ́ | | présent |
| m- | à- | bɔ́n | | | -ɔ́ | -nɔ́ | passé d'hier |
| m- | à- | bɔ́n | | | -ɔ́ | -sɔ́ | passé lointain |
| m- | é- | bɔ́n | | | -ɔ̀ | | passé récent |
| m- | àmò- | bɔ́n | | | -ɔ́ | | passé moyen |
| m- | é- | bɔ́n | | -ɔ́k | -ɔ̀ | | futur moyen |
| m- | é- | bɔ́n | | -ɔ́k | -ɔ̀ | -nɔ́ | futur de demain |
| m- | é- | bɔ́n | | -ɔ́k | -ɔ̀ | -sɔ́ | futur lointain |
| m- | ábí- | bɔ́n | | -ɔ́k | -ɔ̀ | | futur imminent |

- La CP 4 marque la voix. Cette position est vide à la voix active et pleine à la voix passive ;
- La CP 5 marque l'aspect ;
- La CP 6 marque la voyelle thématique ;
- La CP 7 marque l'éloignement temporel.

La structure du verbe en ikota est la suivante :

| Indice sujet- | Indice temporel- | Stem | -(Voix) | -(Aspect) | -Voyelle thématique | -(Éloignement) |
|---|---|---|---|---|---|---|

La formalisation dans XMG s'inspire du concept de classes de positions de Stump (2001). Elle utilise la notion de *domaine topologique* (Bech, 1955) qui consiste en une séquence linéaire de champs organisée dans des blocs élémentaires. Un bloc élémentaire va fournir deux types d'informations : la forme des items lexicaux et les propriétés morphosyntaxiques propres à chaque item. Dans l'environnement XMG, à un champ doit correspondre un item et un seul qui représente la forme phonologique lexicale d'un exposant. Aux sept CP relevées dans le tableau de la structure du verbe ci-dessus va correspondre sept blocs élémentaires dans XMG. *méʤákànáá* ≪ je mangerai (futur de demain) ≫ par exemple est décrit dans la métagrammaire comme la concaténation simultanée de sept blocs élémentaires.

TABLE 4 – Formalisation de *méʤákànáá* ≪ je mangerai ≫

| 1 ⟵ m | 2⟵é | 3⟵ʤ | 4⟵nul | 5⟵Ák | 6⟵À | 7⟵ nÁ |
|---|---|---|---|---|---|---|
| p = 1<br>n = sg | tense = futur | g1 | active = + | tense = futur<br>prog = - | theme = g1 | proxi = day |

La description dans XMG équivaut à une phonologie à deux niveaux (Koskenniemi, 2013), car la méta-grammaire modélise uniquement le niveau lexical de la phonologie. Le niveau de surface est dérivé par post-traitement (environnement qui décrit les règles morphophonologiques). Nous avons par exemple la forme *òéʤákàná* au niveau lexical. Cette forme n'est pas prononçable dans la langue. Pour palier à ce problème, nous posons une règle selon laquelle *ò* + *é* = *ó* et nous obtenons *óʤákàná* ≪ tu mangeras ≫. La formalisation de la conjugaison de trois verbes à l'actif et au passif (en tenant compte des personnes et des classes nominales), en incluant la négation, a permis d'obtenir environ 600 formes verbales fléchies. celles-ci peuvent être exportées au format XML pour une éventuelle réutilisation.

### 3.3 Décrire la morphologie nominale en somali

Le somali, ou somali commun, est une langue afroasiatique appartenant au groupe couchitique de l'est. La morphologie de la langue étant fort complexe, nous nous concentrons ici sur quelques détails cruciaux pour la compréhension de notre analyse.

Les noms du somali sont organisés en groupes flexionnels, distingués sur la base des trois critès suivants : [12]

(1) a. position de l'accent tonal
b. forme du pluriel
c. genre du pluriel par rapport au genre du singulier

En ce qui concerne (1-a), nous suivons Hyman (1981) et considérons que chaque nom du somali porte un accent tonal (AT). L'accent tonal consiste en une variation mélodique accompagnée d'une variation de l'intensité relative sur une syllabe, cf. (Le Gac, 2001) pour plus de détails. L'AT s'associe soit à la dernière voyelle, soit à la pénultième (une voyelle longue compte comme deux unités). Au singulier, l'AT est sur la dernière voyelle lorsqu'un nom est fémin et sur la pénultième lorsque le nom est masculin : *ínan* M sg 'garçon' et *inán* F sg 'fille'. [13]

Les noms du somali peuvent être pluralisés à l'aide de trois stratégies

12. Nous ne tiendrons pas compte de la flexion casuelle, cf. (Saeed, 1993).

13. Nous adoptons l'orthographe officielle somalie. Dans les cas suivants, des conventions spécifiques s'appliquent : <sh> /ʃ/, <kh> /x/, <dh> /ɖ/, <x> /ħ/, <c> /ʕ/, <'> /ʔ/, <j> /d͡ʒ/ and <y> /j/.

différentes. La première est la suffixation de la voyelle *-o*. Cette opération peut s'accompagner de la gémination de la dernière consonne du radical du singulier : *albáab* M sg 'porte' → *albaabbó* F pl 'portes'. La deuxième est l'ajout de la voyelle *-a* et la copie de la dernière voyelle du radical : *míis* M sg 'table' → *miisás* M pl 'tables'. La troisième stratégie implique le seul changement de la position de l'AT : *mádax* M sg 'tête' → *madáx* F pl 'têtes'.

Enfin, certains noms changent de genre au pluriel par rapport au singulier. C'est le cas de tous les noms pluralisés à l'aide du changement de la position de l'AT (M → F), ainsi que certains noms parmi ceux suffixés par *-o* (M → F aussi bien que F → M). En revanche, les noms pluralisés par réduplication de la dernière consonne du radical ne changent jamais de genre au pluriel.

Le tableau 5 ci-dessous illustre cette situation en présentant cinq classes flexionnelles distinctes. [14]

Table 5 – Noms du somali

| classe | singulier | genre | pluriel | genre | |
|---|---|---|---|---|---|
| 1.a | *naág* | F+ | *naagó* | M | 'femme' |
| .b | *galáb* | | *galbó* | | 'après-midi' |
| 2.a | *albáab* | M+ | *albaabbó* | F | 'porte' |
| .b | *daríiq* | | *dariiqyó* | | 'route' |
| 3 | *ílik* | M | *ilkó* | M | 'dent' |
| 4 | *míis* | M | *miisás* | M | 'table' |
| 5 | *baré* | M | *barayáal* | F | 'enseignant' |
| 6 | *sheekó* | F | *sheekoóyin* | M | 'compte' |

La description à l'aide de XMG que nous proposons doit relever un défi majeur. Ceci consiste à rendre compte du nombre élevé d'opérations (morpho)phonologiques ayant lieu à l'intérieur de chaque groupe. Par exemple, les noms appartenant au groupe 1.b sont caractérisés par la perte de la deuxième voyelle au pluriel. Ainsi, *galáb* F sg 'après-midi' devient, après suffixation du marqueur du pluriel *-o*, *galbó* M pl 'après-midis'. Ce phénomène phonologique a lieu dans la classe 3, aussi. Mais, dans ce cas, le genre du nom ne change pas au pluriel. Dans les deux cas, la deuxième voyelle est identique à la première. [15]

14. Cf. (Saeed, 1993) pour deux classes supplémentaires.
15. cf. (Barillot, 2002) pour une analyse exhaustive de ce phénomène en somali.

Dans un premier temps, nous donnons ici la structure générale de la métagrammaire XMG des noms en somali, puis nous présentons plus en détail l'un des défis majeurs de cette description, la représentation du changement de genre lors du passage du singulier au pluriel accompagnée d'une gémination (cas du nom *porte* par exemple).

La description XMG est construite autour d'un ensemble de blocs élémentaires correspondant aux noms de la langue associés à des traits morpho-syntaxiques (genre, classe, etc.). Ainsi le bloc suivant décrit le nom *femme* ('woman' en anglais) :

$$Woman[C,G] \quad \rightarrow \quad \boxed{\begin{array}{l} 1 \leftarrow \text{naag} \\ \hline \mathsf{C} = 1 \\ \mathsf{G} = \mathtt{f} \end{array}}$$

Ce bloc est paramétré par deux informations notées C et G représentant respectivement la classe et le genre du nom. On regroupe le lexique dans une abstraction appelée *Nom* (Noun) :

$$Noun[C,G] \quad \rightarrow \quad Woman[C,G] \vee Door[C,G] \vee Fire[C,G] \vee Skin[C,G] \vee \ldots$$

Enfin, les pluriels sont définis comme l'ajout aux noms d'un suffixe (noté 2 dans les blocs ci-dessous) dépendant de leur classe (et résultant parfois dans un nouveau genre) :

$$Plural \quad \rightarrow \quad Noun[C,G] \wedge \left\{ \begin{array}{l} \boxed{\begin{array}{l} 2 \leftarrow \text{o} \\ \hline \mathsf{C} = 1 \\ \mathsf{G} = \mathtt{f} \\ \text{gender} = \mathtt{m} \\ \text{class} = 1 \end{array}} \vee \boxed{\begin{array}{l} 2 \leftarrow +\text{o} \\ \hline \mathsf{C} = 2\text{a} \\ \mathsf{G} = \mathtt{m} \\ \text{gender} = \mathtt{f} \\ \text{class} = 2\text{a} \end{array}} \vee \boxed{\begin{array}{l} 2 \leftarrow +\text{o} \\ \hline \mathsf{C} = 2\text{b} \\ \mathsf{G} = \mathtt{m} \\ \text{gender} = \mathtt{f} \\ \text{class} = 2\text{b} \end{array}} \\ \vee \boxed{\begin{array}{l} 2 \leftarrow \text{o} \\ \hline \mathsf{C} = 3 \\ \mathsf{G} = \mathtt{m} \\ \text{gender} = \mathtt{m} \\ \text{class} = 3 \end{array}} \vee \boxed{\begin{array}{l} 2 \leftarrow \text{oC} \\ \hline \mathsf{C} = 4 \\ \mathsf{G} = \mathtt{m} \\ \text{gender} = \mathtt{m} \\ \text{class} = 4 \end{array}} \vee \boxed{\begin{array}{l} 2 \leftarrow \text{ayaal} \\ \hline \mathsf{C} = 5 \\ \mathsf{G} = \mathtt{m} \\ \text{gender} = \mathtt{f} \\ \text{class} = 5 \end{array}} \vee \boxed{\begin{array}{l} 2 \leftarrow \text{oyin} \\ \hline \mathsf{C} = 6 \\ \mathsf{G} = \mathtt{f} \\ \text{gender} = \mathtt{m} \\ \text{class} = 6 \end{array}} \end{array} \right\}$$

Si l'on prend le cas de *femme*, le pluriel sera formé en sélectionnant le bloc fournissant le suffixe *o* et produisant un genre (noté gender ci-dessus) masculin.

Pour le cas problématique de *porte*, le suffixe contribué par le passage au pluriel est noté *+o* (car ce nom appartient à la classe 2a). Ici le '+' dénote une opération de dédoublement de la syllabe consonnante finale, qui sera effectué a posteriori par une outil dédié (en l'occurence l'outil HFST (Koskenniemi et Yli-Jyrä, 2008).

### 3.4 Décrire la syntaxe de créoles

La majorité des descriptions des langues africaines (en particulier subsahariennes et dans un cadre francophone [16]) a été effectuée dans le cadre des théories de l'énonciation (modèle de (Culioli, 1999)) ou dans un cadre typologique-fonctionnel (dans le sillage de (Creissels, 1991)). Ces modèles ont l'avantage de fournir au linguiste de terrain des outils descriptifs pertinents mais ils ne sont pas entièrement formalisés. Ainsi, bien que la description initiale d'une langue puisse être correctement faite dans ces modèles, ils ne peuvent pas être immédiatement utilisés dans une grammaire informatisée et une étape de retranscription dans un cadre formel est nécessaire. Nous évoquerons ici des travaux sur des langues créoles qui s'inspirent fortement des travaux de (Creissels, 2006) et sont le fruit de recherches de terrain sur des langues encore peu décrites (notamment en ce qui concerne le sãotomense).

Deux langues créoles font l'objet de travaux dans le cadre de la grammaire d'arbres adjoints (TAG), pour lesquels XMG a été utilisé : le créole de l'île de Guadeloupe (Antilles françaises) et le sãotomense (désormais ST, qui est un créole portugais du Golfe de Guinée). Bien que les travaux sur le créole guadeloupéen soient les plus aboutis à l'heure actuelle, nous présentons ici un fragment de grammaire du ST qui illustre le fait qu'une langue typologiquement éloignée du français et de l'anglais peut assez aisément être décrite par une métagrammaire.

ST est une langue qui exprime les notions de temps de mode et d'aspect par des marqueurs préverbaux (TMA) qui sont à la fois dépendants sémantiquement du verbe (l'interprétation sémantique du marqueur ne peut se faire sans connaître la classe aspectuelle du verbe) et syntaxiquement indépendants (ils peuvent être séparés du verbe par des adverbes). Ils sont strictement ordonnés par catégorie : Temps < Aspect < V. On considère donc qu'ils occupent une place de co-tête au sein d'un arbre élémentaire contenant un verbe (voir (Schang *et al.*, 2012) pour une analyse détaillée). Les exemples en (2) en illustrent le fonctionnement :

(2) a. Zon kume.
Jean manger
Jean a mangé.

16. La tradition linguistique française et belge de description des langues a été poursuivie et enrichie dans les universités d'Afrique francophone mais est demeurée quasiment inconnue des universités anglosaxonnes.

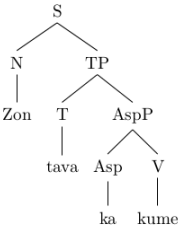
b. Zon ka kume.
Jean Imperf. manger
Jean mange.

c. Zon sa ka kume.
Jean Present Imperf. manger
Jean est en train de manger.

d. Zon tava ka kume.
Jean Antérieur Imperf. manger
Jean est en train de manger.

On peut représenter la structure de (2-d) par l'arbre suivant :

(3)

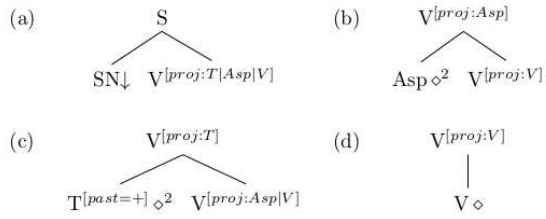



De façon plus abstraite, on peut dire que tous les arbres élémentaires des verbes du ST devront comprendre :

— une place pour le sujet (obligatoire en ST)
— une place pour chacun des TMA éventuellement réalisés
— une place pour l'élément lexical (le verbe)
— une place pour les complément éventuels de ce verbe

Pour exprimer cette généralisation avec XMG, nous avons décomposé les arbres élémentaires verbaux en fragments d'arbres qui se recomposent en fonction des combinaisons de TMA autorisées :

(4)



Ainsi, on peut dire que (3) est résultat de la conjonction des fragments suivants :

$$Sujet(\text{4-a}) \wedge Intransitif(\text{4-d}) \wedge Aspect(\text{4-b}) \wedge Temps(\text{4-c})$$

L'utilisation de XMG permet donc au linguiste de créer des généralisations sur les objets de sa grammaire, ce qui permet de créer un nombre important de structures complexes à partir de formes de base. A titre d'exemple, la grammaire du créole guadeloupéen comprend à l'heure actuelle plus de 500 formes d'arbres élémentaires à partir d'une trentaine de fragments de base.

# 4 Conclusion

Dans cet article, nous avons montré comment décrire plusieurs dimensions de la langue (syntaxe, morphologie) formellement au moyen du langage informatique XMG. Ce langage offre un environnement expressif et modulaire permettant de décrire des ressources linguistiques variées et qui peuvent par la suite être compilées en ressources électroniques dans un format XML. Nous avons montré comment ce langage XMG a été utilisé pour décrire plusieurs langues d'Afrique (ikota, somali et sãotomense). Nous travaillons actuellement à l'enrichissement du langage XMG, afin de permettre au linguiste de définir son propre langage de description en fonction de la langue décrite. Il pourrait ainsi définir de quels types de structures de données il a besoin (arbres syntaxiques, structures de traits, etc.) et les appliquer à plusieurs niveaux d'une langue (ou à plusieurs langues dans le cas d'études multi-lingues). En parallèle à cela, nous continuons à travailler à l'extension de nos descriptions de l'ikota, du somali et du sãotomense, afin de parvenir à des ressources à couverture raisonnable pour pouvoir être utiles à la communauté (ces ressoures seront distribuées librement).

# Références

Autebert, J.-M., Berstel, J. et Boasson, L. (1997). Context-Free Languages and Push-Down Automata. *In* Rozenberg, G. et Salomaa, A., éditeurs : *HandBook of Formal Languages*, volume 1, pages 111–174. Springer Verlag.

Barillot, X. (2002). *Morphophonologie gabaritique et information consonantique latente en somali et dans les langues est-couchitiques*. Thèse de doctorat, Université Paris 7.

Bech, G. (1955). Studien über das deutsche Verbum infinitum, 1+ 2, København : Det Kgl. *Danske Videnskabernes Selskab (Historisk-filologiske Meddelelser, vol. 35/2 og 36/6)*.

Beesley, K. R. et Karttunen, L. (2003). *Finite State Morphology*. Center for the Study of Language and Information.

Bonami, O. et Boyé, G. (2006). Deriving Inflectional Irregularity. *In* Müller, S., éditeur : *Proceedings of the HPSG06 Conference*, pages 361–380. Stanford : CSLI Publications.

Chomsky, N. (1957). *Syntactic Structures*. Mouton, The Hague.

Chomsky, N. et Halle, M. (1968). *Sound Patterns of English*. Berlin and New York : Mouton de Gruyter.

Crabbé, B. (2005). *Représentation informatique de grammaires fortement lexicalisées : Application à la grammaire d'arbres adjoints*. Thèse de doctorat, Université Nancy 2.

Creissels, D. (1991). *Description des langues négro-africaines et théorie syntaxique*. Ellug.

Creissels, D. (2006). *Syntaxe générale : une introduction typologique*. Hermes sciences.

Culioli, A. (1999). *Pour une linguistique de l'énonciation : Formalisation et opérations de repérage*, volume 2. Editions Ophrys.

Embick, D. (2010). *Localism versus Globalism in Morphology and Phonology*. M.I.T. Press.

Halle, M. et Marantz, A. (1993). Distributed Morphology and the Pieces of Inflection. *In* Hale, K. et Keyser, S. J., éditeurs : *The View from Building 20*, pages 111–176. MIT Press, Cambridge, Massachusetts.

Hockett, C. F. (1954). Two models of grammatical description. *Word*, 10:210–234.

Hyman, L. (1981). Tonal Accent in Somali. *Studies in African Linguistics*, 12(2):169–203.

Joshi, A. K. et Schabes, Y. (1997). Tree Adjoining Grammars. *In* Rozenberg, G. et Salomaa, A., éditeurs : *Handbook of Formal Languages*, volume 3, pages 69–123. Springer Verlag, Berlin.

Koskenniemi, K. (2013). An informal discovery procedure for two-level rules. *Journal of Language Modelling*, 1(1):155–188.

Koskenniemi, K. et Yli-Jyrä, A. (2008). CLARIN and Free Open Source Finite-State Tools. *In Proceedings of the International Workshop on Finite State Methods for NLP (FSMNLP)*, pages 3–13, Ispra, Italy.

Le Gac, D. (2001). *Structure prosodique de la focalisation : le cas du somali et du français*. Thèse de doctorat, Université Paris VII.

Le Roux, J. et Parmentier, Y. (2005). *The XMG manual*. LORIA, Nancy.

Montague, R. (1974). English as a Formal Language. *In* Thomason, R. H., éditeur : *Formal Philosophy : Selected Papers of Richard Montague*, pages 188–222. Yale University Press, New Haven, London.

Saeed, J. I. (1993). *Somali Reference Grammar*. Kesington, Maryland : Dunwoody Press.

Schang, E., Duchier, D., Ekoukou, B. M., Parmentier, Y., Petitjean, S. *et al.* (2012). Describing Sao Tomense Using a Tree-Adjoining Meta-Grammar. *In 11th International Workshop on Tree Adjoining Grammars and Related Formalisms (TAG+ 11)*, pages 82–89.

Shieber, S. (1985). Evidence against the context-freeness of natural language. *Linguistics and Philosophy*, 8(3):333–343.

Stump, G. T. (2001). *Inflectional Morphology. A theory of Paradigm Structure*. New York : Cambridge University Press.

XTAG Research Group (2001). A Lexicalized Tree Adjoining Grammar for English. Rapport technique IRCS-01-03, IRCS, University of Pennsylvania.