Describing São Tomense Using a Tree-Adjoining Meta-Grammar

Emmanuel Schang[1], Denys Duchier[2], Brunelle Magnana Ekoukou[1],
Yannick Parmentier[2], Simon Petitjean[2]

(1) LLL, Université d'Orléans - 10, rue de Tours 45067 Orléans Cedex 2 – France
(2) LIFO, Université d'Orléans - 6, rue Léonard de Vinci 45067 Orléans Cedex 2 – France

firstname.lastname@univ-orleans.fr

Abstract

In this paper, we show how the interactions between the tense, aspect and mood preverbal markers in São Tomense can be formally and concisely described at an abstract level, using the concept of projection. More precisely, we show how to encode the different valid orders of preverbal markers in an abstract description of a Tree-Adjoining Grammar of São Tomense. This description is written using the XMG meta-grammar language (Crabbé and Duchier, 2004).

1 Introduction

São Tomense[1] is a Portuguese-based Creole language spoken on São Tomé Island (RDSTP). Like many (if not all) Creole Languages, it has preverbal markers expressing Tense and Aspect (TMA markers in the classical literature on Creole languages, see (Holm, 1989)), as shown in (1):

(1) a. tataluga xiga.
turtle come
'The turtle came.'
b. tataluga ka xiga.
turtle IMPF *come*
'The turtle is coming.'
c. tataluga tava ka xiga
turtle Anterior IMPF *come*
'The turtle was coming.'

Several approaches have been proposed to formally describe the combinations of TMA markers in São Tomense, including tree-based descriptions such as Tree-Adjoining Grammar (TAG) (Schang, 2000). Schang's TAG uses adjunction (*i.e.*, auxiliary trees) to encode the ordering of the TMA markers. As we shall see in Section 5, this is not satisfactory for several reasons. In this paper, we propose to shift the description of TMA markers to a meta-level, using the XMG language (Crabbé and Duchier, 2004). The paper is structured as follows. In section 2, we describe São Tomense TMA system. In Section 3, we introduce the XMG language. Section 4 focuses on the syntactic properties of the TMA markers. In Section 5, we then show how to control the TMA markers' combinations in an XMG meta-grammar. This meta-grammar is then compiled in order to produce a TAG where verbal elementary trees only contain correctly ordered TMA markers (realised as lexical nodes).

2 TMA system

Before describing the TMA markers and their combination, let us first look at the bare verbs.

2.1 Bare Verbs

As in many languages (and as in most of the Creoles), bare verbs are used to express the past perfective (or preterite) with dynamic processes (as in (1-a)) and express the present tense with stative verbs, as in (2).

(2) n konse mana bo.
1sg know sister your
'I know your sister.' (*'I knew your sister')

Stative verbs are often considered to collide with the TMA markers (Ferraz, 1979), but several uses of both have been noticed in ST spoken corpora

[1]The abbreviations are: ST (São Tomense) ; IMPF (imperfective) ; Asp (Aspect).

Proceedings of the 11th International Workshop on Tree Adjoining Grammars and Related Formalisms (TAG+11), pages 82–89,
Paris, September 2012.

(Schang, 2000), triggering an inchoative meaning (3).

(3) e ka sa yo godu.
3sg IMPF *be very fat*
'He is going to be very fat."

(Schang, 2000, p. 193) shows that bare verbs in São Tomense are literally "bare" and that no information on Tense or Aspect is attached to them, and that no functional projection (containing a zero morphem) is needed to account for the various uses of bare verbs. By contrast, the preverbal markers bear such temporal and aspectual features.

2.2 Aspect

ka is the most-used aspectual marker in São Tomense[2]. (Hagemeijer, 2007) and (Ferraz, 1979) provide several examples of its uses in various contexts, triggering habitual reading (4-a) , future tense (4-b) and conditionality (4-c):

(4) a. Zon ka kanta ni gleza.
John Asp sing in church
'John uses to sing at church.'
b. Zoze ka xiga amanha.
José Asp come tomorrow
'José will come tomorrow.'
c. xi bo **ka** bi amanha, bo ka be mu.
if 2sg Asp come tomorrow 2sg Asp see me
'If you come tomorrow, you'll see me.'

(Schang, 2000, p. 193) shows that all the various interpretations of *ka* boil down to an imperfective reading, which is the core meaning of this marker.

2.3 Tense

Two Tense markers occupy the same position: *tava* (anterior) and *sa* (present). Both markers derive from the Portuguese verb *estar* 'to be', in its 3sg imperfect indicative tense form and 3sg present tense form respectively. They inherit from the temporal value of the etymon.

While *tava* can freely combine with the verb, *sa* goes together with *ka*, often pronounced *xka*,[3] see (5).

(5) a. e tava bi.
3sg Tense come
'He had come.'
b. e sa ka bi.
3sg Tense Asp come
'He is coming.'
c. e tava ka bi.
3sg Tense Asp come
'He was coming.'
d. *e sa bi.

(5-b) illustrates the *sa ka* or *xka* (its short form) combination which triggers the progressive reading. Any other combination is blocked,[4] see (6).

(6) a. *e ka tava bi.
b. *e sa tava bi.
c. *e ka sa bi.

To summarize, São Tomense combines a few preverbal markers in order to derive a rich range of semantic interpretations.

3 eXtensible Meta-Grammar

As mentioned above, in this paper, we show how to move the description of TMA markers in a São Tomense TAG from the syntactic level (*i.e.*, the TAG elementary trees) to a meta-level, using the *eXtensible Meta-Grammar* (XMG) framework. This move makes it possible for the linguist to concisely describe the valid TMA orders.[5]

XMG is a declarative language for specifying tree-based grammars at a meta-level (Crabbé and Duchier, 2004). Basically, XMG allows to abstract over tree structures (*i.e.*, to capture generalizations) by defining (i) elementary tree fragments and (ii) conjunctive / disjunctive combinations of these fragments. Such an abstraction over a (tree) grammar is generally called a *meta-grammar*. It is compiled in order to automatically produce the underlying grammar.[6]

[2] Leaving aside its allomorph *ga*.

[3] It can also be pronouced 'e ska bi'.

[4] (Hagemeijer, 2007) reports some other combinations (*sa xka*, *ka ka*, *tava sa xka*) which are firmly rejected by our informants and absent from the fieldwork recordings we have. It suggests that some variation exists. But as we focus on standard ST we don't take it into account. Note however that these combinations can be seen as relaxed constraints on the system, and do not invalidate our analyses.

[5] This move presupposes that TMA markers should rather be treated as co-anchors of verbal elementary trees than anchors of auxiliary trees. This is motivated in Section 4.

[6] The compiler for the XMG language is also called XMG, and is freely available at `https://launchpad.net/xmg`.

The elementary tree fragments of the XMG language correspond to tree descriptions and are encapsulated within *classes*. Such a class provides the linguist with a mean to refer to a given tree description, *e.g.*, in order to reuse it in distinct contextes. These tree descriptions can contain (node or feature) variables, dominance and precedence constraints on nodes, and labelling constraints (association of a node with some feature structure). Note that the combinations of these tree descriptions are also encapsulated within classes, and that the default scope of a variable is the class. XMG is also equipped with an inheritance mechanism, which allows to import the content of a class and access directly its variables.

The compilation of an XMG specification amounts to (i) accumulating tree descriptions and then (ii) solving accumulated tree descriptions. As a result, a fully redundant grammar is generated (*i.e.*, TAG trees grouped into tree families).

The XMG language reveals expressive enough to describe a large amount of syntactic structures in a compact way, as shown by the various tree grammars designed with XMG for French (Crabbé, 2005; Perrier, 2007; Gardent, 2008), English (Alahverdzhieva, 2008) and German (Kallmeyer et al., 2008).

A particularly interesting feature of the XMG language is that it comes with a set of built-in *linguistic principles* that the linguist can activate in order to ensure the validity of the output structures (Crabbé et al., To appear). These principles not only guaranty the well-formedness of the grammar with respect to linguistic invariants, but also help the linguist to highly factorise her/his meta-grammar. Indeed, principles allow the linguist to avoid defining numerous alternative descriptions for exceptions, but to rather catch them during the compilation of the meta-grammar.

In the meta-grammar for ST described in Section 5, we use the unifications over feature structures labelling nodes, which are triggered during tree description solving, to rule out invalid TMA orders. In a future work, we plan to rather describe valid TMA orders via a dedicated linguistic principle.

4 Projecting Aspect and Tense

Prior to describing our meta-grammar of ST, let us describe interesting properties of TMA markers, which will motivate our formal description of ST.

(Schang, 2000) and (Hagemeijer, 2007) propose a description of the properties of the TMA markers that we complete below. Contrary to the full verb *sa* and *tava* ('be'), which can be used as copula, as in (7), *sa* and *tava* as TMA do not have the properties of the verbs they originate from, a fact we will show below.

(7) a. kafe sa kentxi.
coffee be hot
'The coffee is hot.'
b. kafe tava kentxi.
coffee be.Anterior hot
'The coffee was hot.'

The question we address here is the nature of *ka*, *sa* and *tava*. We present a series of tests which shows that TMA markers behave differently from verbs (auxiliaries included), adverbs and adjectives (note that hereafter we use the reduced form *xka* instead of the full form *sa ka*).

- Coordination

  Contrary to lexical items, TMA markers cannot be coordinated (neither overtly nor covertly):

  (8) *Zon sa i/o tava ka kume.
  John Tense and/or Tense Asp eat
  'John is and/or was eating.'

  Note that the TMA markers don't show the properties of French and English auxiliaries with regard to coordination.

- Reiteration

  TMA markers cannot be reiterated on the same verb (9), contrary to adverbs for instance (see (Schang, 2012) for a study of lexical reiteration in ST).

  (9) *Zon sa sa ka kume / *Zon ka ka kume / *John Tense Tense Asp eat / *John Asp Asp eat

- Negation

  Sentential negation in ST is double-headed. The first particle comes to the immediate left of the TMA markers and the second one

comes in sentence-final position (see (Hagemeijer, 2007) and (Schang, 2000) for a description).

(10) Zon **na** xka (*na) kume loso
John Neg1 TMA (Neg1) eat rice
fa.
Neg2
'John doesn't eat the rice.'

However, *fa* is used without *na* in partial negation (contrastive negation):

(11) a. ami fa!
me Neg2
'Not me!'
b. karu fa!
car Neg2
'Not the car!'
c. kume fa!
eat Neg2
'Not eating!'
d. glavi fa!
beautiful Neg2
'Not beautiful!'
e. leve-leve fa!
slowly Neg2
'Not slowly'
f. isa fa!
this Neg2
'Not this one!'

(12) *{ka/xka/tava/tava ka} fa!
[Tense and Asp markers negated]

The TMA markers cannot be negated (12) while pronouns, nouns, verbs, adjective, adverbs and strong demonstratives can, as in (11-a-f).
While English auxiliaries for instance can be negated, TMA markers cannot (13):

(13) a. Zon tava ka kume?
John Tense Asp eat
'Was he eating?'
b. *Inon, e na tava ka
no 3sg Neg1 Tense Asp
fa.
Neg2
'No, he wasn't.'

- Participle-like constructions

  Some verbs of Portuguese origin have been incorporated in ST lexicon with their past participle form (ex. Port.: chegadu > ST: xigadu). While they can be complement of a full verb (*fika* 'to stay', or sa 'to be' (the full verb used as copula), they cannot appear with TMA markers, as shown in (14) (adapted from (Hagemeijer, 2007, p.132)) :

  (14) a. *kinte ka/xka balidu.
  garden TMA swept
  b. kinte sa/fika balidu.
  garden is/stays swept
  'The garden has been/remains swept'

- Question-answer pairs

  TMA markers cannot form a minimal answer:

  (15) a. Zon ka/xka bali kinte?
  John TMA sweep garden
  'Does John sweep/is sweeping the garden?'
  b. efan, e ka/xka *(bali).
  yes he TMA sweep
  'Yes, he does.'

- VP-fronting:

  (16) a. bo ka/xka bali kinte.
  you TMA sweep garden
  'you (sweep/are sweeping) the garden.'
  b. bali kinte so bo
  sweep garden FOCUS you
  ka/xka *(bali)
  TMA sweep
  'SWEEP THE GARDEN is what he does/is doing.'

- Pseudo-cleft

  (17) a. kume/dansa/kanta sa kwa
  eat/dance/sing is thing
  ku e ka/xka fe.
  that he TMA do
  'Eating/dancing/singing is what he does/is doing.'

b. *ka kume/dansa/kanta sa
Asp eat/danse/sing is
kwa ku e ka fe.
thing that 3sg Asp do

In the fronted position where only the lexical verb (without its functional projections) is allowed, the TMA are excluded. No ellipsis is allowed for the inflected verb. To describe it in classic words, it shows that the material copied to the focus position originates below INFL.

We conclude from these tests that the TMA markers are clearly functional elements, as inflectional affixes in English and French are.

The reason why TMA markers are not represented as prefixes in the relevant litterature comes from adverb placement. The adverb *kwaji* can be inserted between Tense and Aspect, as in:

(18) Tataluga sa kwaji ka koda.
turtle Tense almost Asp wake-up
'The turtle is about to wake up.'

(19) Tataluga (??kwaji) xka (*kwaji) koda.

(19) shows that when *kwaji* is inserted, *sa* and *ka* cannot freely agglutinate as *xka/ska*. Note incidentally that the agglutinated form *xka* is thus built post-syntactically in phonology.

5 Describing Tense and Aspect in São Tomense using a Meta-Grammar

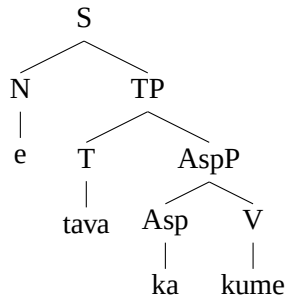
(Schang, 2000) proposes an analysis in the TAG framework which treats the TMA markers as adjuncts to V and uses Tense and Aspect features on the foot node of the adjunct tree to reject invalid combinations. However, a description based on the concept of Extended Projections (Grimshaw, 1991) (see also (Frank, 2004) for a similar approach) better reflects the fact that TMA markers are not adjuncts such as adjectives or adverbs are. Consequently, we treat here TMA markers as extended projections of V, which can remain bare or be stretched with Tense and Aspect projections.

Thus, Tense and Aspect markers are not stored in the Lexicon (they don't anchor any tree) but are co-anchors of the elementary tree associated with verbs.

Let us consider (21), which illustrates the structure of (20).

(20) e tava ka kume.
3sg Tense Asp eat
'He was eating.'

(21)

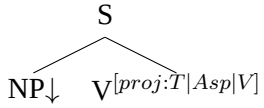


In (21), S is a projection of V, the maximal functional stretching of the verb.

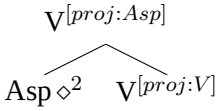
These facts can easily be recast in XMG's framework. To this aim, the structure (20) is broken down into four pieces (*i.e.* classes) each containing minimal information. These Classes are listed below.

- $CanSubject$: to express what is usually called the External Argument of the verb. It is described in (22-a) .
- $Intransitive$ verb: the minimal projection of V. It is described in (22-e) .
- $Aspect$: as a projection of the aspectual marker. It is described in (22-b) .
- $Tensed$: as a projection of Tense. Note that $Tensed$ refers to a disjunction of two tree fragments, which differ according to the *past* feature labelling the Tense marker. This distinction allows us to treat the case where a non-past Tense marker must precede an Aspect marker. The corresponding two tree fragments are described in (22-c) and (22-d).

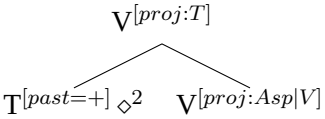
(22) a.



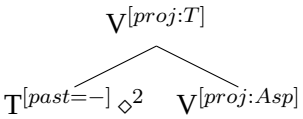
b.



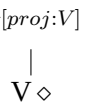
c.



d.



e.



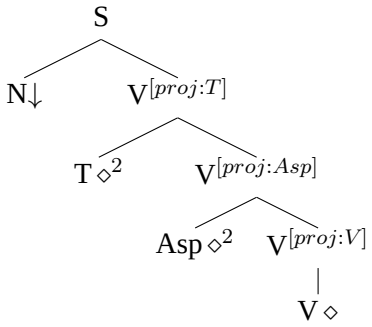
Thus, (21) is built up from the following conjunction of Classes:

$$CanSubject \wedge Intransitive \wedge Aspect \wedge Tensed$$

The feature ***proj(ection)*** is used here to rule out invalid combinations in the output elementary tree.[7] As mentioned in Section 3, during the compilation of the meta-grammar, the accumulated tree descriptions are solved in order to produce minimal tree models (which correspond to the elementary TAG trees of the grammar being described). In the present case of TMA markers, the tree description solver will compute verbal elementary trees by identifying nodes belonging to the tree fragments introduced in (22). For such a node identification to succeed, the nodes need to be labelled with feature structures, which unify. While giving a linguistically motivated account of the properties of TMA markers, the *proj* feature will help the meta-grammar compiler to only produce valid elementary trees (recall that Tense must dominate/precede Aspect and V).

From the conjunction of classes given above, the result of the meta-grammar compilation are elementary trees for intransitive verbs, including the tree associated with *kume* 'to eat' depicted in (23).

(23)

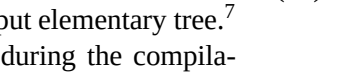


To fill the Tense and Aspect slots, this verb appears in the Lexicon as associated with two co-anchor equations (*cf*, ◇-nodes refer to anchors and $\diamond^2$-nodes to co-anchors in (23)):[8]

- T$\rightarrow tava\ [past = +]$
- Asp $\rightarrow ka$

A felicitous side-effect of incorporating the TMA markers in the elementary tree of the verb appears when looking at the derivation tree of the sentence "e tava ka kume" (24) where functional information such as Tense and Aspect do not appear as adjuncts but are hold by the verb (tree α1, tree α2 being the elementary tree of the pronoun *e* '3sg').

(24)



It is interesting to notice that, in this context, TMA markers can be treated similarly to Tense and Aspect affixes in some agglutinative languages (see (Duchier et al., 2012) for an analysis of Ikota – Bantu B25 – with XMG), diverging only in the way they combine.

Of course, treating TMA markers as co-anchors raises the question of the production of the numerous elementary trees and the computational efficiency of parsing with these.

Regarding the production of elementary trees, the use of the XMG framework makes it possible to concisely describe elementary trees (including TMA markers), the XMG compiler being in charge of producing the redundant elementary trees.[9]

Regarding the computational efficiency of parsing with TAG grammars having TMA markers embedded in verbal elementary trees, it may not be a problem for the following reasons. While this treatment of TMA markers causes the grammar to have a much higher number of elementary trees (TMA markers are no longer factored out, as it is the case when using auxiliary trees), it is worth considering two points.

First, Creoles are known to have little morphology (McWhorter, 2001) and ST does not allow many transformations (no voice and no argumental affixation). The extra cost of enlarging the

[7] In the values associated with feature proj, "|" refers to disjunction.

[8] Here, we adopt a grammar-lexicon interface comparable to that of the XTAG project (XTAG Research Group, 2001), where the grammar is made of unanchored trees, anchoring being realized at parsing.

[9] The question on how to produce the large lexicon used to anchor the grammar (that is, containing the co-anchor equations) remains to be answered, nonetheless one option would be to use techniques for automatic lexicon acquisition such as that of (Sagot, 2005).

grammar size is thus low (and make ST grammar size still reasonable).

Second, when replacing auxiliary trees with co-anchoring equations, the parsing complexity is somehow moved from the actual parsing step (where adjunction is processed) to the lexical selection and anchoring step (which is done prior to actual parsing, see *e.g.* (Gardent et al., 2011)). In other words, the complexity here raises when selecting the right lexical entries, and anchoring the many trees associated with these entries. But, once the elementary trees are anchored, it will be possible to select a pertinent subgrammar (that is, to remove useless trees with respect to the sentence to parse) using techniques such as polarity-based filtering (Gardent et al., 2011).

6 Conclusion

In this paper, we have shown how to implement the concept of projection at an abstract level (the meta-grammar) in order to describe a crucial domain of the syntax of São Tomense, namely the TMA markers. We claim that the TMA markers have to be integrated in the TAG elementary trees of verbs instead of anchoring auxiliary trees, as it was done before (Schang, 2000). This comes from the fact these markers can be considered as functional elements.

In this context, we chose to use a meta-grammatical framework, namely the XMG system, in order to facilitate the description of verbal elementary trees equipped with nodes for TMA markers. By *facilitate*, we do not only mean that the meta-grammar compiler will take care of the tedious task of producing the numerous elementary trees concerned with TMA markers, but also (and mainly) that an abstract level may be the right place to implement a linguistic theory such as that of projection used here.

Acknowledgments

We are grateful to the three anonymous reviewers for their helpful comments.

References

Katya Alahverdzhieva. 2008. XTAG using XMG. Master Thesis, Nancy Université.

Benoît Crabbé and Denys Duchier. 2004. Metagrammar redux. In *Constraint Solving and Language Processing, First International Workshop (CSLP 2004)*, volume 3438 of *Lecture Notes in Computer Science*, pages 32–47, Roskilde, Denmark. Springer.

Benoît Crabbé, Denys Duchier, Claire Gardent, Joseph Le Roux, and Yannick Parmentier. To appear. XMG: eXtensible Meta-Grammar. *Computational Linguistics*.

Benoît Crabbé. 2005. *Représentation informatique de grammaires fortement lexicalisées : Application à la grammaire d'arbres adjoints*. Ph.D. thesis, Université Nancy 2.

D. Duchier, B.M. Ekoukou, Y. Parmentier, S. Petitjean, E. Schang, et al. 2012. Describing Morphologically-rich Languages using Metagrammars: a Look at Verbs in Ikota. In *Workshop on'Language technology for normalisation of less-resourced languages', 8th SALTMIL Workshop on Minority Languages and the 4th workshop on African Language Technology*, pages 55–60.

Luiz Ivens Ferraz. 1979. *The Creole of Sao Tome*. Witwatersrand University Press, Johannesburg.

Robert Frank. 2004. *Phrase structure composition and syntactic dependencies*, volume 38. Mit Press.

Claire Gardent, Yannick Parmentier, Guy Perrier, and Sylvain Schmitz. 2011. Lexical Disambiguation in LTAG using Left Context. In *5th Language & Technology Conference - LTC'11*, pages 395–399, Poznań, Poland.

Claire Gardent. 2008. Integrating a Unification-Based Semantics in a Large Scale Lexicalised Tree Adjoining Grammar for French. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 249–256, Manchester, UK.

Jane Grimshaw. 1991. Extended projection. Master thesis, Brandeis University.

Tjerk Hagemeijer. 2007. *Clause structure in Santome*. Ph.D. thesis, University of Lisbon.

J.A. Holm. 1989. *Pidgins and Creoles*, volume 1&2. Cambridge Univ Pr.

Laura Kallmeyer, Timm Lichte, Wolfgang Maier, Yannick Parmentier, and Johannes Dellert. 2008. Developing a TT-MCTAG for German with an RCG-based Parser. In *The sixth international conference on Language Resources and Evaluation (LREC 08)*, pages 782–789, Marrakech, Morocco.

J.H. McWhorter. 2001. The worlds simplest grammars are creole grammars. *Linguistic typology*, 5(2-3):125–166.

Guy Perrier. 2007. A French Interaction Grammar. In *proceedings of the 6th Conference on Recent Advances in Natural Language Processing (RANLP 2007)*, pages 463–467, Borovets, Bulgaria.

Benoît Sagot. 2005. Automatic acquisition of a Slovak Lexicon from a Raw Corpus. In *Proceedings of the 8th International Conference on Text, Speech and Dialogue (TSD'05)*, volume 3658 of *Lecture*

Notes in Artificial Intelligence, pages 156–163, Karlovy Vary, Czech Republic. Springer-Verlag.

Emmanuel Schang. 2000. *L'émergence des créoles portugais du golfe de Guinée*. Ph.D. thesis, Université Nancy 2.

Emmanuel Schang. 2012. Reduplication in São Tomense. *The Morphosyntax of Reiteration in Creole and Non-Creole Languages*, pages 235–250.

XTAG Research Group. 2001. A lexicalized tree adjoining grammar for english. Technical Report IRCS-01-03, IRCS, University of Pennsylvania.