

A multivariate analysis of canonical and non-canonical uses of Switch-Reference markers in Mbyá narratives

Guillaume Thomas

Germino Duarte

Abstract: Switch-reference is a family of grammatical devices whose primary function is to indicate whether two linked clauses have coreferential pivots, where the pivot is a prominent argument in each clause. In some languages, in addition to their function of reference tracking, switch-reference markers can be used to indicate whether the events or situations described by the two linked clauses differ with respect to some parameter, such as time, place or actuality. This phenomenon is known as non-canonical switch-reference. Whether canonical and non-canonical switch-reference marking are distinct grammatical phenomena is still an open question. In this paper, we investigate uses of switch-reference markers in a corpus of Mbyá Guaraní (Tupian) narratives, and we argue that the alternation between canonical and non-canonical uses is an epiphenomenon of the multifactorial and probabilistic nature of switch-reference marker choice. In this perspective, there is only one grammatical process of switch-reference marking and the distinction between canonical and non-canonical switch-reference marking is matter of language use.

Keywords: Switch-reference, Mbyá Guaraní, lasso regression.

1 Introduction

This paper discusses the relation between canonical and non-canonical uses of switch-reference markers based on a case study of Mbyá Guaraní. Switch-reference (henceforth: SR) has been defined as “a morpheme associated with clause junctures that indicates whether a prominent argument in each clause co-refers” (McKenzie 2014; cf. Haiman and Munro 1983). We refer to the clause that contains the SR marker as the *marked clause*, and to the other one as the *reference clause*. In the following examples from Amele, the SAME (SM) suffix *-me* in (a) indicates coreference between the subject of the reference clause *na* \emptyset -*i-te-i-a* (‘he gave me the stick’) and the subject of the marked clause *uqa q-it-i-me-i* (‘he hit me’). The DIFFERENT (DF) suffix *-co* in (b) indicates disjoint reference between the subjects of the marked and reference clauses. Following Stirling (1993), we use the term *pivots* to refer to the two arguments that are related by SR marking.

(1) Amele (Stirling, 1993: 184)

- a. *Uqa q-it-i-me-i* *na* \emptyset -*i-te-i-a*.
3SG hit-1SG-PRED-SM-3SG stick give-PRED-1SG-3SG-TODP
‘He hit me and then gave me the stick.’

- b. *Hina ho-co-m sab je-i-a.*
 2SG come-DF-2SG food eat-3SG-TODP
 ‘You came and he ate the food.’

It is well known that in some languages, SR markers can be used to indicate whether the situations described by two clauses are similar or not, independently of coreference between pivots. This phenomenon is known as non-canonical switch-reference and is illustrated in (2), where a DIFFERENT marker is used despite pivot coreference, in order to indicate a change of place:

(2) **Amele** (Stirling, 1993: 216)

- Age ceta gul-do-co-bil I-i bahim na tac-ein*
 3PL yam carry-3SG-DF-3PL go-PRED floor on fill-3PL.REMP
 ‘They carried the yams on their shoulders and went and filled up the yam store.’

Given the existence of non-canonical uses of SR marking, some scholars have argued that the function of SR constructions is to mark thematic continuity or discontinuity across clauses, referential continuity being only one aspect thereof (see a.o. Mithun, 1993, 2021; Stirling, 1993; Watkins, 1993; van Gijn, 2012, 2016a,b; Pustet, 2013). From this perspective, canonical and non-canonical SR marking are understood as different uses of the same construction, which emphasize referential or non-referential dimensions of thematic continuity. Here, we adopt Givón’s (2001) definition of thematic coherence or continuity:

(3) **Coherence as continuity** (Givón, 2001: 328-329):

Coherence is the continuity or recurrence of some element(s) across a contiguous span of multi-propositional discourse.

Givón proceeds to list seven elements that can contribute to thematic continuity: referents (‘participants’), location, temporality, aspectuality, modality, perspective (‘narrative voice’) and action/events.¹ Under the view of SR constructions as indicators of thematic continuity, canonical uses of SR markers would attend to the first of these elements, while non-canonical uses would attend to others.

This literature raises the question of how to model the process of SR marker choice in such a way as to account for both canonical and non-canonical uses. The present manuscript addresses this question through a case study of SR marking in a corpus of narratives produced in Mbyá Guaraní, a Tupian language spoken by about 30,000 speakers in Argentina, Brazil and Paraguay (Ladeira, 2018), in which canonical uses of SR markers track subject coreference (Dooley, 1989, 1992, 1999).

In the first part of the manuscript, we explore the distribution of canonical and non-canonical uses of SR markers in our corpus. We evaluate the contribution of various factors to predicting the use of SAME vs. DIFFERENT markers using monofactorial tests and we motivate an analysis of SR marker choice in Mbyá as a probabilistic process that encompasses both canonical and non-canonical uses, eschewing the reduction of non-canonical uses to non-referential dimensions of thematic continuity. In the second part of the manuscript, we show that a multifactorial model of

¹Following Givón, we understand referential continuity to be restricted to reference to participants of events described in discourse. I.e., continuous reference to non-participants such as times does not fall under the label ‘referential continuity.’ This is merely a terminological point and not a denial that one may refer to times, events and such.

SR marker choice that is blind to the distinction between canonical and non-canonical uses will still predict non-canonical uses of SR markers with adequate frequency and in adequate contexts. We conclude from these results that the distinction between canonical and non-canonical uses of SR markers in Mbyá can be adequately modelled as a side effect of a single multifactorial process of SR marker choice.

The paper is structured as follows. In section 2, we give an overview of previous studies of SR in Mbyá. In section 3, we introduce our corpus and the variables to be used for the analysis of SR. In section 4, we explore the corpus and motivate a probabilistic analysis of SR marker choice. In section 5, we fit a model of SR marker choice to the corpus. We argue that this model provides an adequate analysis of both canonical and non-canonical uses of SR markers. Section 6 concludes.

2 Overview of Switch-Reference in Mbyá Guaraní

The SR system of Mbyá was studied notably by Dooley (1989, 1992, 1999). Switch-reference is expressed by the particles *vy* (SS) and *ramo* or its reduced form *rã* (DS), as illustrated in examples (4) and (5).²

(4) *Ava o-o vy mboi o-exa.*
 man A3-go SM snake A3-see
 ‘When the man_i went, he_{i/*j} saw the snake.’ (Dooley, 1989)

(5) *Ava o-o ramo mboi o-exa.*
 man A3-go DF snake A3-see
 ‘When the man went, the snake saw him.’ (Dooley, 1989)

In such constructions, SR markers track the reference of subjects (Dooley, 1989). Dooley (2015: 25) identifies the subject with the nominative argument a verb, i.e., the unique argument of intransitive verbs (S) or the (proto-)agent argument of (di)transitive verbs (A), and he discusses three constructions that are sensitive to this grammatical function, in addition to SR marking: (i) reflexive possessive markers must be bound by subjects, (ii) impersonal voice eliminates reference to the subject argument and (iii) the subject of post-verbal converbs must be coreferent with the subject of the verb they modify.

Dooley (1989) argues that SR in Mbyá tracks the reference of grammatical subjects rather than agents or topics. In example (6), the SAME marker *vy* indicates that the grammatical subjects of *okaru* (‘eating’) and *tove tomano* (‘let him die’) corefer, although the subject of *okaru*, which is also its agent, does not corefer with the addressee, which is the notional agent of the optative predication *tove tomano*:³

²A list of glosses is provided in appendix. Note that Mbyá Guaraní agreement follows an active-inactive alignment pattern. Intransitive verbs agree in person and number with their subject, different agreement markers being used for active verbs and inactive verbs. Transitive verbs agree either with their highest argument on the person hierarchy 1 > 2 > 3, or with their subject in case of a tie. Agreement with transitive subjects is expressed with active agreement markers, and agreement with objects is expressed with inactive agreement markers. In glosses, ‘A3’ stands for ‘third person agreement marker, active class’ and ‘B3’ stands for ‘third person agreement marker, inactive class.’

³Note that in this example, unlike in previous ones, the marked clause follows the reference clause. The preferred position of the marked clause varies with the semantic relation between marked and reference clauses, see Dooley (2016: §21.3.1).

- (6) *Pe-juka e'ỹ teĩ tove t-o-mano ha'e ae o-karu e'ỹ vy.*
 A2PL-kill NEG CONC OPT OPT-A3-die 3 INT A3-eat NEG SM
 ‘Without your killing him, let him die all by himself from not eating.’
 (Dooley, 1989: 7)

In example (7), the DIFFERENT marker *ramo* indicates that the grammatical subjects of *omombe'u* (‘he talked’) and *aexa* (‘I saw’) do not corefer, although both clauses share the same topic, *compadre Galdino*, which is the object of the marked clause and the subject of the reference clause:

- (7) *Compadre Galdino ma a-exa Roberto r-o py ramo ma, gu-a'y-i*
 godfather Galdino BDY A1SG-see Roberto R-house in DF BDY REFL-son-DIM
o-mombe'u.
 A3-talk
 ‘Compadre Galdino, when I saw him at Roberto’s house, he talked about his little son.’ (Dooley, 1989: 10)

With respect to their external syntax, Dooley (2015: 119) argues that marked clauses in SR constructions are adverbial subordinate clauses. Note that SR markers underspecify the semantic relation between the marked and reference clauses. To illustrate, while the marked clauses in examples (4) and (5) are interpreted as temporal adverbial clauses, causal or conditional interpretations of SR constructions are also attested, among other interpretations.

The SR constructions we discussed up to this point relate a marked clause to a reference clause by subordination. SR markers are also attested sentence initially with the anaphoric pronoun *ha'e* in place of a marked clause, as illustrated by examples (8) and (9), where *ha'e* is anaphoric to the previous sentence. We call these construction *reduced switch-reference*. By contrast, we refer to SR constructions like those illustrated by (4) and (5) as *full switch-reference*.

- (8) *Peteĩ-gue je ava o-o o-i-ny t-ape r-upi.*
 one-time HSY man A3-go A3-be-CONV NPOSSD-road R-LOC
Ha'e vy je o-exa apere'a.
 ANA SM HSY A3-see preá
 ‘Once, a man was going on a road. He saw a preá.’ (Verissimo, 2002a)

- (9) *Guaxu je o-po-opo o-iko-vy, nd-o-guapy-i, nd-o-pyta-i guive.*
 deer HSY A3-jump-RED A3-be-CONV NEG-A3-sit-NEG, NEG-A3-stop-NEG ADD
Ha'e rã je irũ kuery o-porandu: “Mba'e tu r-endu” he'i.
 ANA DF HSY friend PL A3-ask what MIR A2SG-feel A3.say
 ‘A deer was jumping around, it couldn’t sit, it couldn’t stand still either. So his friends asked – What is wrong with you?’ (Verissimo, 2002a)

Note that in this pair of examples, the SR markers *vy* and *rã* occur at the beginning of a new sentence, and they indicate whether the subject of that sentence corefers with the subject of the previous sentence. In example (8), the subjects of both sentences refer to the man introduced in the first sentence, and a SAME marker is used. In example (9), the subject of the first sentence introduces a deer, and the subject of the second sentence refers to his friends. A DIFFERENT marker is used.

Reduced SR is part of a broader class of sentence initial connectives formed by combining the anaphoric pronoun *ha'e* with a subordinating conjunction or postposition, as illustrated by the connective *ha'e gui* in example (10):

- (10) *Peteĩ ára je Vera o-o yakã py pira o-jopoi vy.*
 One day HSY Vera A3-GO river to fish A3-feed SM
Ha'e gui je pira o-gueno-ẽ ma.
 ANA SRC HSY fish A3-COMIT-leave ASP
 'One day, Vera went to the river to fish, and he caught a fish.' (Veríssimo, 2002a)

Dooley (1992, 2015) calls these constructions *reduced subordinate clauses* and argues that the pronoun *ha'e* refers to the content of a preceding discourse unit. In example (10), *ha'e* is anaphoric to the first sentence, and the postposition *gui* indicates that the event described in the second sentence (Vera caught a fish) is contingent on the event described in the first one (Vera went to the river). In other words, reduced subordinate clauses appear to be a form of tail-head linkage, a phenomenon that has been documented in multiple language families and areas (Guérin, 2019) and that has been independently observed to interact with SR (Guillaume, 2011).

In all previous examples, SR markers are used canonically: the choice of marker tracks subject reference. With reduced SR markers in examples (6) and (7), the subjects whose reference is being tracked are those of the reference clause and of the preceding sentence, which serves as the antecedent of the anaphoric pronoun *ha'e*. Non-canonical uses of SR markers are also attested in both types of SR constructions, as illustrated by examples (11) and (12). For ease of reading, we gloss non-canonical uses of *vy* and *rã~ramo* as SM_{NC} and DF_{NC} respectively:

- (11) *Nda-xe-ayvu kuaa-i r-e ramo, a-iko tema.*
 NEG-B1.SG-speak know-NEG R-ABL DF_{NC} A1.SG-live CONT
 'Even though I didn't know how to speak, I got along' (Dooley, 2011).
- (12) *Ha'e kuery ma je o-ma'ẽ guaxu 'rã o-ke vy.*
 ANA PL BDY HSY A3-look big HAB 3-sleep SM
Ha'e rã nd-o-ke-i vy ma je o-ma'ẽ r-a'y-'i.
 ANA DF_{NC} NEG-A3-sleep-NEG SM BDY HSY A3-look R-small-DIM
 'When they [owls] sleep, their eyes are wide open. But when they are awake, their eyes are narrow.' (Veríssimo, 2002b)

In example (11), the DIFFERENT marker *ramo* is used despite the coreference of the subjects of the marked and reference clauses. In example (12), *ha'e rã* relates two sentences with coreferential subjects. Non-canonical uses of SR markers are also attested with the same marker *vy*.

3 Corpus and variables

Our corpus consists of 81 narratives from the state of Parana and São Paulo in Brazil. It includes 1,313 sentences (14,575 tokens). The narratives were produced by seven adult Mbyá speakers from the state of Paraná. A first part of the corpus consists of 33 narratives composed in various workshops organized by Robert Dooley and the Summer Institute of Linguistics between 1976 and

1990. An interlinearized version of the corpus with a translation into English is available on the Archive of the Indigenous Languages of the America (Dooley, 2011). The second part consists of 48 narratives collected in Veríssimo (2002a) and Veríssimo (2002b) together with their translation into Brazilian Portuguese. Narratives in both part of the corpus have been used as educational material. This is especially true of narratives in the second part, which were created for literacy training. This may explain the short length of most narratives in the corpus, which averages 16 sentences per narrative.

Several layers of annotations were added by the first author and collaborators: interlinear glosses when missing, syntactic structure in dependency grammar, coreference relation and animacy (see Thomas et al., 2021). For this study, we added a layer of rhetorical relation annotations between discourse units that are related by a switch-reference marker. Rhetorical relation annotation was carried out collaboratively by the two authors, using consensus judgment. In doing so, we relied on the second author’s interpretation of the narratives as a native speaker of Mbyá as well as annotation guidelines described in section 3.2. We extracted 762 occurrences of SR constructions from the corpus: 392 full SR constructions, 370 reduced SR constructions. Of these, 368 use SAME markers and 394 use DIFFERENT markers. 16 occurrences of SR markers are used non-canonically. We say that a DIFFERENT marker is used non-canonically when the subjects of the SR construction corefer, and a SAME marker is used non-canonically when they don’t.

Each observation in our study corresponds to a single occurrence of an SR construction. Relevant properties of these observations are coded as values of a series of variables. A first variable, `marker_type`, encodes the form of the SR marker in an observation. It has two levels: SAME (if the SR marker is *vy*) or DIFFERENT (if the SR marker in the observation is *ramo* or *rã*). Note that these two levels only indicate the form of the SR marker (*vy* versus *ramo* or *rã*). Whether the pivots in a given SR construction are coreferential or not is coded by a different set of variables described in section 3.1.

We define 7 variables that are potential predictors of SR marker choice. These variables can be grouped into 5 classes: referential continuity, rhetorical relations, mirativity, spatio-temporal continuity and construction type. We discuss these five classes of variables in the following subsections.

3.1 Referential continuity

We hypothesize that SR marking in Mbyá may be sensitive to referential continuity beyond pivot coreference. We define a variable `referential_continuity`, which captures a subset of coreference relations between the marked clause and the reference clause of an SR construction. However, since the marked clause can either precede or follow the reference clause, it is more convenient to define this variable using the concepts of *antecedent discourse unit* and *current discourse unit*:

(13) Antecedent and current discourse units:

Given a switch-reference marker that relates two discourse units, the current discourse unit is the one that was uttered last, and the antecedent discourse unit is the one that was uttered first.

Discourse Units (DUs) are defined in Segmented Discourse Representation Theory, the theory of discourse structure that underlies our treatment of rhetorical relations (see subsection 3.2). With full SR, the discourse units related by SR markers are the marked clause and the reference clause.

With reduced SR, the discourse units related by SR are the reference clause and the propositional antecedent of the anaphoric pronoun *ha'e*.

In order to measure referential continuity beyond subject coreference, we borrow the notion of Backward Looking Centre from Centering Theory (Grosz and Sidner, 1986; Grosz et al., 1995), redefining it for the purposes of the present study. In the following definition, it is assumed that referential expressions are ranked by prominence in a clause according to their grammatical function, where subjects are more prominent than objects, which in turn are more prominent than obliques and adjuncts:

(14) Backward Looking Centre:

The Backward Looking Centre (BC) of a switch-reference construction is the referential expression in the current DU that has the most prominent antecedent in the antecedent DU.

With these definitions in place, we can define the variable `referential_continuity`, which has 4 levels ordered by decreasing continuity (SS > S0.OS > 00 > none):

(15) `referential_continuity`

- a. SS: the BC is the subject of the current DU, and its antecedent is the subject of the antecedent DU.
- b. S0.OS:
either the BC is the subject of the current DU, but its antecedent is not the subject of the antecedent DU,
or the BC is not the subject of the current DU, but its antecedent is the subject of the antecedent DU.
- c. 00: the BC is not the subject of the current DU, and its antecedent is not the subject of the antecedent DU.
- d. none: no referential expression in the current DU has an antecedent in the antecedent DU.

Borrowing again from Centering Theory, we hypothesize that coreference between more prominent arguments generates more discourse coherence, which is captured in the ordering of the levels of this variable. The highest degree of referential continuity, SS, corresponds to coreference between the canonical pivots of the SR construction. We expect that higher levels of referential continuity will be associated with a higher proportion of SAME marker use.

Example (16) illustrates our use of `referential_continuity`. The SR marker *ramo* relates the antecedent DU labelled π_a to the current DU labelled π_b . The BC of this SR construction is the dative argument of π_b (*xevy*, ‘to me’), and its antecedent is the subject of π_a . Its reference is the narrator. The subject of π_b (*mamaẽ*, ‘mother’) is not the BC, since its antecedent in π_a is less prominent than the antecedent of *xevy*. The level of `referential_continuity` for the SR construction (π_a, π_b) is S0.OS:

- (16) *Ha'e rire ma xee ma a-ju vy* [π_a *mamaẽ pe a-porandu .*] *Ha'e_a ramo*
 ANA SEQ BDY B1.SG BDY A1.SG-come SM mother DAT A1.SG-ask ANA DF
ma [π_b *mamaẽ aipo-e'i xe-vy*] : “ *O-o guyra-'i avy vy* ” *he'i*
 BDY mother ATTN-say B1.SG-DAT A3-go bird-DIM hunt SM A3.SAY
 ‘After a while, I went and asked my mother. She told me: “He went bird hunting”, she said.’
 (Dooley, 2011)

3.2 Rhetorical relations

Several prominent approaches to discourse interpretation argue that a discourse is perceived as coherent only if its utterances are connected by rhetorical relations (also known as coherence relations) in a tree-like or graph-like discourse structure (Hobbs, 1985; Mann and Thompson, 1988; Asher and Lascarides, 2003). In his study of reduced SR in Mbyá, Dooley (1989) noted that the use of DIFFERENT SR markers appears to be favoured by relations of contrast and counter-expectation between the discourse units related by SR markers. This was illustrated with examples (11) and (12) in section 2. In the present study, we hypothesize more generally that SR markers relate discourse units that are connected by a rhetorical relation, and that the nature of this relation may affect the speaker’s choice of SR marker.

We analyze rhetorical relations using Segmented Discourse Representation Theory (SDRT; Asher and Lascarides, 2003). SDRT is built on top of a dynamic semantics (Discourse Representation Theory, see Kamp and Reyle, 1993), to which it adds speech act discourse referents and rhetorical relations between them. Speech act discourse referents (SA-drefs) label the content of clauses and other discourse units (DUs).

Since they have propositional content, the marked and reference clauses related by SR markers are DUs that introduce SA-drefs. The same goes for the antecedent of the anaphoric pronoun *ha'e* in reduced SR. We can therefore situate the DUs related by SR markers on an SDRT graph. To illustrate, consider discourse (17) and its simplified discourse structure in figure 1. The first occurrence of the SAME marker *vy* relates the marked clause *eixu rugue oẽmba* (‘the wasps got out’) to the reference clause *xepipa rive* (‘they stung me’). These clauses introduce the SA-drefs π_a and π_b respectively, which are related by *Narration*. The second occurrence of *vy* relates the pronoun *ha'e* to its reference clause. *Ha'e* is anaphoric to *xepipa rive*, and the reference clause is the matrix clause of the second sentence, *xero katy ajevvy* (‘I came back home’), which introduces the SA-dref π_d . This SA-dref relates to π_b by *Narration*. The SA-dref π_c is introduced by the clause *aja'eo reve* (‘crying’), which is adjoined to the reference clause of the SR construction. We relate π_c to π_d with a relation of *Elaboration*, π_c being subordinate to π_d in the discourse structure.⁴

- (17) *Ha'e ramo-ve* [π_a *eixu r-ugue o-ẽ-mba vy*] [π_b *xe-pi-pa rive*]
 ANA DS-INT wasp R-swarm A3-get.out SS B1.SG-sting-COMP EXCL
Ha'e_b vy [π_c *a-jae'o r-eve*] [π_d *xe-r-o katy a-jevvy*] ...
 ANA SS A1.SG-cry R-MAN B1.SG-R-house DIR A1.SG-come.back
 ‘The wasps got out and stung me all over. I came back home crying (...)’ (Verissimo, 2002b)

⁴Note that π_c could also be analyzed as relating to π_d by the relation *accompanying circumstance*, which has been proposed for the analysis of participial adjuncts in English. See Behrens and Fabricius-Hansen (2010) and Behrens et al. (2012) for a discussion of this relation and the analysis of participial adjuncts in SDRT.

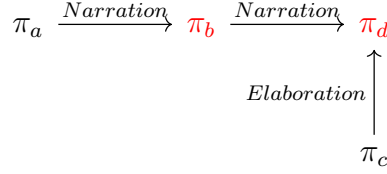


Figure 1: Simplified discourse structure of example (17)

Every occurrence of an SR construction in the corpus was annotated for its rhetorical structure by the two authors. This annotation was based on consensus judgments between the second author, who provided expertise in the interpretation of narratives as a native speaker of Mbyá, and the first author, who provided expertise in theories of discourse semantics. The annotation followed the guidelines for SDRT annotations of Reese et al. (2007). Our annotation was constrained by the requirement that any two discourse units connected by an SR marker should be related by a unique rhetorical relation. In cases where several rhetorical relations were considered to be plausible candidates, the relation that was judged to be more salient by the second author was chosen.

Because of the large number of rhetorical relations in our inventory (see table 1 below), some rhetorical relations are quite rare in the corpus. In order to mitigate this scarcity, rhetorical relations were grouped into coarser classes along the dimensions of additivity and polarity introduced by Sanders et al. (2021).⁵

The variable polarity has two levels, positive and negative. Negative rhetorical relations include contrastive, adversative and concession relations. In our inventory of rhetorical relations, only *Contrast* is unambiguously negative. Sanders et al. (2021) also discuss a negative use of the *Alternation* relation in SDRT, which is notably expressed by exclusive disjunction in English. All other relations are positive. The following examples illustrate negative polarity relations with a canonical and a non-canonical use of SR markers, respectively:

- (18) [π_a *A-j-exa-uka* *ta* *ra'aga*] π_a *vy rive ta'vy* [π_b *xee ae a-je-juka*
 A1.SG-REF-see-CAUS PROSP CF SM EXCL FRUST I INT A1.SG-REFL-kill
rai .] π_b

almost

‘I was just going to show off, but I almost got myself killed.’ (Veríssimo, 2002a)

⁵Sanders et al. present a classification of rhetorical or coherence relations that is applicable to annotation schemas used in major theories of discourse structure, including SDRT. The dimensions used in their classification are *polarity*, *basic operation*, *source of coherence*, *implication order* and *temporality*. Out of these five dimensions, we determined that only *polarity*, *basic operation* and *temporality* are relevant to our study of SR marking. In this process, we were guided by the hypothesis that the function of SR marking is to mark thematic continuity across discourse units. *Source of coherence* distinguishes between rhetorical relations that connect discourse units at the level of their propositional content, and rhetorical relations that express the speaker’s opinion, argument, claim or conclusion. This dimension appears to be orthogonal to the distinction between thematic continuity and discontinuity and therefore irrelevant to the present study. *Implication order* only applies to a subset of rhetorical relations, which raises issues for the statistical modelling of SR marker choice. In addition, it also appears to be orthogonal to thematic (dis)continuity, since it merely distinguishes causal relations that introduce the cause first from those that introduce the consequence first. We attend to *temporality* as part of a broader treatment of spatio-temporal continuity, see section 3.3. This leaves us with two dimensions, *polarity* and *basic operation*, the second of which we rename as *additivity* for conciseness and clarity.

- (19) [π_a *Nda-xe-ayvu* *kuaa-i* *r-e*] *ramo* [π_b *a-iko* *tema*] .
 NEG-B1.SG-speak know.how-NEG R-ABL DF_{NC} A1.SG-live CONT
 ‘I didn’t know how to speak, but I survived.’ (Dooley, 2011)

Positive rhetorical relations have already been illustrated in previous examples, see for instance *Narration* and *Elaboration* in example (17). Because negative relations express a form of thematic rupture (such as the frustration of an expected outcome), we expect that, everything else being equal, discourse units related by a negative relation will have a higher proportion of DIFFERENT markers than discourse units related by a positive relation.

The variable additivity has two levels, additive and causal. Rhetorical relations that are causal all involve an implicational relation, while additive ones do not. The SDRT relations *Result*, *Explanation* and *Consequence* are causal, and the relations *Alternation*, *Contrast* and *Elaboration* have causal and additive uses. All other relations are additive. Example (20) illustrates the use of causal relations, in this case *Explanation*. We expect that, everything else being equal, discourse units related by a causal relation will have a higher proportion of DIFFERENT markers than discourse units related by a non-causal relation.

- (20) [π_a *Urutau ma je nd-o-vy’a-i*] [π_b *i-juru guaxu vaipa*] *vy* .
 potoo BDY HSY NEG-A3-happy-NEG B3-mouth big very SM
 ‘The potoo was unhappy because he had such a large mouth.’ (Dooley, 2011)

Table 1 lists the rhetorical relations used for annotation, grouped by levels of additivity and polarity.⁶

additivity			polarity		
additive	causal	additive/causal	positive	negative	positive/negative
Attribution	Consequence	Alternation	Attribution	Contrast	Alternation
Background	Explanation	Contrast	Background		
Commentary	Goal	Elaboration	Commentary		
Continuation	Result		Consequence		
Narration			Continuation		
Precondition			Elaboration		
Source			Explanation		
			Goal		
			Parallel		
			Precondition		
			Result		
			Source		

Table 1: Classification of Rhetorical Relations by additivity and polarity

⁶Note that the rhetorical relation *goal* is absent from Reese et al. (2007) and was adopted from Muller et al. (2012).

3.3 Spatio-temporal continuity

In some SR systems, DIFFERENT markers can indicate spatial or temporal discontinuity (see [Stirling, 1993](#); [Roberts, 2017](#)). [Roberts \(1988\)](#) notably showed that in Amele, DIFFERENT marking can be used to indicate a change of place or time, among other parameters. Note that in that case, the factor that conditions the use of SM versus DF marking is the spatio-temporal identity of the situations described in the discourse units related by SR. Specifically, DIFFERENT marking can be used to indicate that two situations unfold at different times or in different places. We capture this form of spatio-temporal continuity through two variables, time and place.

The variable time has two levels: sequence and containment. The former is used when the situations described by the marked and reference clauses unfold at different times, that is to say, they follow one another in time (regardless of the lapse between them). The latter was used when the two situations are simultaneous, or when one situation includes the other in time.⁷ Note that several rhetorical relations in SDRT have temporal entailments. The most frequent ones in our corpus are *narration*, which entails sequence, and *background*, which entails containment.

The two levels of the variable place are different and containment. Similarly to temporal sequence, spatial different is used when the situations described by the marked and reference clauses occur at locations that do not stand in a containment relation. Spatial containment is used when the location of one situation includes that of the other one, or when the two situations occur at the same location.

If Mbyá Guaraní SR were similar to Amele SR in its sensitivity to spatio-temporal continuity, we would expect that, everything else being equal, discourse units that describe situations that unfold at different times (time=sequence) or in different places (place=different) should have higher proportions of DIFFERENT markers.

With respect to annotation, the spatial and temporal locations of the situations described by the marked and reference clause were inferred based on the use of spatial and temporal modifiers, the type of events or states described by the marked and reference clause, and contextual cues provided by narratives. Note that because Mbyá Guaraní is a tenseless language, tense inflection cannot serve as a resource in the annotation of spatio-temporal continuity.

3.4 Mirativity

[Dooley \(1992\)](#) argues that one of the factors that govern non-canonical uses of SR markers in Mbyá is whether the SR construction describes a sequence of events that unfolds in a predictable fashion. To a certain extent, this factor is captured by the variable polarity, since *contrast* is the only negative rhetorical relation, and counter-expectational *contrast* conveys that the outcome of an event was not as it was expected. In addition, the Mbyá language also has mirative particles at its disposal, whose use in SR constructions gives cues about the speaker's perception of predictability. Mirativity is a linguistic category whose function is "to mark sentences which report information which is new or surprising to the speaker" ([DeLancey, 1997](#)). When a mirative particle occurs in the second member of a pair of DUs related by an SR marker, it may convey that the event described by this DU is surprising in the context provided by the first DU. In addition, surprise of a protagonist

⁷If an interval B extends past an interval A in time, interval A and B are said to stand in a sequence relation, even if the two intervals have a non-empty intersection. In other words, sequence stands for partial rather than complete precedence in a period structure.

may be lexically encoded by verbs such as *nhemondyi* (‘be startled’). We expect that, everything else being equal, the presence of mirative particles in the second member of a pair of discourse units related SR will be associated with a higher proportion of DIFFERENT marking.

Every occurrence of SR marker in the corpus was annotated for the presence of a mirative particle or a verb of surprise in the second member of the pair of DUs it relates.⁸ The following three mirative particles were considered in our annotation:

(21) Mirative particles (definitions from Dooley, 2016):⁹

- a. *ra’e*: ‘Indicates a discovery, that is, indicates that a fact is verified only in the reported instant.’
- b. *ri ty*: ‘Indicates surprise about something in the context.’
- c. *tu*: ‘Indicates intensity and even abruptness.’

We define the variable *mirativity*, with two levels: TRUE (a mirative particle or verb of surprise is attested in the second DU) and FALSE. The following example illustrates. The second occurrence of the DIFFERENT marker *rã* relates the DUs labelled π_a and π_b . Since two mirative markers (*ri ty* and *ra’e*) occur in the second DU, the SR construction that relates π_a and π_b was annotated TRUE for the variable *mirativity*. Incidentally, we note that SR is used non-canonically in this construction:

(22) *O-ma’ẽ rã je* , *h-edy-pa rei merami rã je* ,
 A3-look DF HSY B3-glowing-COMP EXCL apparently DF HSY

[π_a *o-i-kuaa pota*] *rã je*
 A3-3-know try DF_{NC} HSY

[π_b *oo o-vera-pa va’e py ri ty ra’e o-ĩ* , *kunha va’e* .]
 house A3-shine-COMPL NMLZ LOC MIR MIR A3-be woman NMLZ

‘She looked, and it seemed that everything was glowing; she tried to understand: she was in a house where everything was shining!’ (Dooley, 2011)

3.5 Clause type

The last class of predictors of SR marker choice we considered consists only of one variable, *clause_type*, which encodes whether an SR construction is full or reduced. There are differences between the grammar of these two construction types that may affect the rate of non-canonical uses of SR markers in each. Indeed, SR markers in full constructions relate clauses associated with a well defined set of arguments identifiable by grammatical function. By contrast, SR markers in reduced constructions relate a clause to a pronoun that is anaphoric to a preceding discourse unit. The resolution of the pronoun’s antecedent may fail to retrieve the grammatical function of the arguments that introduced the discourse referents mentioned in the antecedent. This in turn may contribute to a higher rate of non-canonical uses, since canonical uses of SR markers track the referential continuity of pivots defined by grammatical function (i.e., subjects).

⁸That is to say, using the terminology introduced in section 3.1, SR constructions were annotated for the presence of a mirative particle in the current DU.

⁹The definitions from Dooley (2016) were translated from Brazilian Portuguese into English. Note that Dooley (2016) does not discuss the classification of these particles as mirative. Any errors in this respect are ours.

4 Corpus exploration

In this section, we explore the distribution of SR marker types in their canonical and non-canonical uses. Table 2 summarizes our data set. All our predictor variables except `additivity` and `time` are significantly associated with `marker_type` at the $p < 0.05$ significance level. We note in particular that the ratio of `DIFFERENT` to `SAME` markers is significantly higher in the presence of mirative markers than in their absence, and with rhetorical relations of negative polarity as compared to positive polarity.

	DF, N = 394 ¹	SM, N = 368	p-value ²
<code>referential_continuity</code>			<0.001
none	165 (100%)	0 (0%)	
OO	34 (100%)	0 (0%)	
OS.SO	189 (95%)	10 (5%)	
SS	6 (2%)	358 (98%)	
<code>additivity</code>			0.060
additive	308 (54%)	265 (46%)	
causal	86 (46%)	103 (54%)	
<code>polarity</code>			<0.001
negative	26 (84%)	5 (16%)	
positive	368 (50%)	363 (50%)	
<code>mirativity</code>			<0.001
FALSE	339 (49%)	352 (51%)	
TRUE	55 (77%)	16 (23%)	
<code>place</code>			0.003
different	36 (37%)	61 (63%)	
containment	358 (54%)	307 (46%)	
<code>time</code>			0.5
containment	146 (50%)	147 (50%)	
sequence	248 (53%)	221 (47%)	
<code>clause_type</code>			<0.001
full	157 (40%)	235 (60%)	
reduced	237 (64%)	133 (36%)	

¹: n (%)
²: Pearson's Chi-squared test with Yates continuity correction

Table 2: Data set summary

A first generalization about the distribution of canonical and non-canonical SR markers in our corpus can be extracted from the distribution of `referential_continuity` in table 2. Note that canonical uses of SR markers correspond either to `SAME` markers with an `SS` value of this variable (in which case the subjects of the marker and reference clauses corefer), or to `DIFFERENT` markers with a value other than `SS` (in which case the subjects do not corefer). Focusing on non-canonical uses of SR markers then, we note that all 10 non-canonical uses of `SAME` markers in the corpus are attested

with an OS.S0 value of `referential_continuity`. Non-canonical SAME marking is unattested with values 00 and none. This distribution is unlikely to be observed under the null hypothesis that non-canonical SAME markers are equally likely to occur in each of the three non-SS levels of `referential_continuity` (Exact Multinomial Test: $p = 5e-5$).

The previous observation suggests that SR marker choice is sensitive to finer degrees of referential continuity than the distinction between coreferential subjects versus non-coreferential subjects. More precisely, they suggest that SAME marking requires a high degree of referential continuity not only in its canonical uses but also in its non-canonical uses, the latter of which are only attested when the subject of the marked or reference clause corefers with one non-subject in the other clause. Consequently, this observation suggests that SR marker choice is always sensitive to referential continuity, both in its canonical uses and in its non-canonical uses. This is problematic for a view of SR that would posit a categorical distinction between canonical and non-canonical SR according to which the former would serve a function of reference tracking, while the latter would track non-referential dimensions of thematic continuity. By contrast, it is consistent with the view that canonical and non-canonical uses of SR markers are generated by a single process that is uniformly sensitive to referential continuity.

We now turn to a more detailed examination of the distribution of non-canonical uses of SR marking in the corpus. Table 3 presents the combination of predictor values for which non-canonical SAME markers are attested in the corpus, together with counts of non-canonical SAME markers and canonical DIFFERENT markers for each combination.¹⁰ As we already observed, non-canonical SAME markers are unattested at levels of `referential_continuity` lower than OS.S0. In addition, we observe that they are only attested with positive rhetorical relations (`polarity = positive`) and in the absence of mirative markers (`mirativity = FALSE`). This is consistent with the hypothesis that non-canonical SAME markers indicate thematic continuity in non-referential dimensions, or at least the absence of discontinuity that would come from rhetorical relations of contrast, or surprise expressed by mirative markers.

<code>referential_continuity</code>	<code>place</code>	<code>time</code>	<code>additivity</code>	<code>polarity</code>	<code>mirativity</code>	<code>clause type</code>	SM noncan.	DF can.
OS.S0	different	containment	causal	positive	FALSE	reduced	1	0
OS.S0	different	sequence	additive	positive	FALSE	reduced	1	10
OS.S0	containment	containment	additive	positive	FALSE	reduced	1	14
OS.S0	containment	containment	causal	positive	FALSE	reduced	3	3
OS.S0	containment	sequence	additive	positive	FALSE	full	1	18
OS.S0	containment	sequence	additive	positive	FALSE	reduced	2	72
OS.S0	containment	sequence	causal	positive	FALSE	reduced	1	8

Table 3: Distribution of non-canonical SAME markers

Table 4 presents the combination of predictor values for which non-canonical DIFFERENT markers are attested in the corpus, together with counts of non-canonical DIFFERENT markers and canonical SAME markers for each combination. We observe that non-canonical different markers are

¹⁰By definition, canonical SAME markers and non-canonical DIFFERENT markers are unattested at levels of `referential_continuity` other than SS.

only attested with negative rhetorical relations (polarity = negative) or in the presence of mirative markers (mirativity = TRUE). This conforms to the hypothesis that non-canonical DIFFERENT markers indicate thematic discontinuity in non-referential dimensions.

referential continuity	place	time	additivity	polarity	mirativity	clause type	DF noncan.	SM can.
SS	containment	containment	additive	negative	FALSE	reduced	1	0
SS	containment	containment	additive	positive	TRUE	full	1	3
SS	containment	containment	causal	negative	FALSE	full	2	3
SS	containment	sequence	additive	positive	TRUE	full	1	5
SS	containment	sequence	additive	positive	TRUE	reduced	1	3

Table 4: Distribution of non-canonical DIFFERENT markers

Finally, in tables 3 and 4, we observe that canonical SR markers are attested in all but two contexts in which non-canonical markers are attested in the corpus. Here, “context” refers to a combination of predictor values. To the extent that our predictor variables capture the set of factors that govern SR marker choice, this supports a probabilistic approach to SR marker choice, according to which both SAME and DIFFERENT markers may have a non-zero probability of use in any given context.¹¹

¹¹An anonymous reviewer points out that reduced SR constructions seem to favour non-canonical SAME rather than DIFFERENT marking and asks why this may be the case. We note that `clause_type` is significantly associated with `marker_type` overall (canonical and non-canonical uses included), as can be seen in table 2. We also note that the proportion of DIFFERENT markers overall is *greater* in reduced clauses. While this may seem contradictory at first, remember that both canonical uses of DIFFERENT markers and non-canonical uses of SAME markers indicate non-coreferential subjects. Remember also that 98% of all SR markers in the corpus are used canonically, so that the greater proportion of DIFFERENT markers in reduced SR constructions entails a greater proportion of non-coreferential subjects in these constructions. There is indeed a significant association between subject coreference (coreferent vs non-coreferent subjects) and `clause_type` (Fisher’s exact test: $p < 0.001$). In light of this, we hypothesize that the association between `clause_type` and `marker_type` observed both in table 2 and in tables 3 and 4 is indirect, subject coreference acting as a mediating variable (note that subject coreference is the same as `referential_continuity` with all non-SS levels are collapsed into one for simplicity). That is to say, we hypothesize that `clause_type` influences subject coreference, which in turns influences `marker_type`. We use the Baron and Kenny method to establish this (Baron and Kenny, 1986). First, we regress `marker_type` on `clause_type`; `clause_type` significantly influences `marker_type` ($p < 0.001$). Secondly, we regress subject coreference on `clause_type`; `clause_type` significantly influences subject coreference ($p < 0.001$). Third, we regress `marker_type` on `clause_type` controlling for subject coreference; subject coreference significantly influences `marker_type` ($p < 0.001$) but `clause_type` does not have a significant effect on `marker_type` when subject coreference is controlled. Finally, we regress `clause_type` on `marker_type` controlling for subject coreference. `marker_type` does not have a significant effect on `clause_type` ($p = 0.14$) when subject coreference is controlled. These results support our mediation hypothesis and suggest that the effect is complete. Bootstrap estimation with the `Mediation` package in R confirms that the indirect effect of subject coreference is significant ($p < 0.001$). The real question, then, is why there is a greater proportion of non-coreferential subject in reduced SR constructions than in full SR constructions. We speculate that this may be an effect of discourse coherence constraints that put more pressure on keeping the sentence topic constant inside than across sentences.

5 Modelling switch-reference marker choice

We now proceed to a multi-factorial analysis of SR marker choice. Our hypothesis is that canonical and non-canonical uses of SR markers are generated by the same process, which maps a context (characterized by a combination of predictor values) to a probability of SAME marker use (or equivalency of DIFFERENT marker use). In this perspective, the status of SR markers as canonical or non-canonical is neither a parameter of the context nor a class of outcomes in the process that generates SR markers. It is merely an epiphenomenon of SR marker choice. Consequently, neither the predictors of our model nor its outcome variable will encode the distinction between canonical and non-canonical uses. The outcome variable of our model, `marker_type`, has two levels, SAME (SM) and DIFFERENT (DF), which stand for the SR markers *vy* and *ramo~rã* respectively. The predictors are the variables discussed in section 3: `referential_continuity`, `polarity`, `additivity`, `mirativity`, `place`, `time` and `clause_type`.

The reader may wonder how to evaluate the success of such a model in predicting (non-)canonical uses of SR markers. The performance of our model as a classifier is of limited interest. Indeed, we saw in the previous section that in virtually all contexts where non-canonical uses of SR markers are attested, canonical uses are also attested with a higher frequency. Therefore, a probabilistic classifier that always gives a higher probability to the canonical outcome in these contexts will have better classification performance. Consequently, although we will present standard evaluation metrics to assess the success of our model in predicting SAME versus DIFFERENT marking, its success in predicting non-canonical uses of SR markers will be assessed by simulating uses of SR markers from the fitted model and comparing them to occurrences observed in our corpus. More precisely, we will use the predictions of our model on holdout data to simulate vectors of SAME and DIFFERENT markers, and we will ask (i) whether the simulated proportions of non-canonical SR markers match the proportions that are observed in our corpus and (ii) whether the contexts in which non-canonical SR markers are produced in simulations match the contexts in which they are used in the corpus.

The supplementary materials present a model of `marker_type` as a function of pivot coreference only. The interested reader is invited to compare the results presented in the present section to those obtained from this simpler baseline model.

5.1 Penalized regression model of SR marker choice

We train a logistic regression model of SR marker choice with the variables discussed in section 3 as predictors, as specified in formula (23):

$$(23) \quad \text{marker_type} \sim \text{referential_continuity} + \text{polarity} + \text{additivity} + \text{mirativity} + \text{place} + \text{time} + \text{clause_type}$$

When verifying that the assumptions of logistic regression are met for this model, we observe that quasi-complete separation leads to infinite maximum likelihood estimates for several parameters. There is also high multicollinearity between `referential_continuity`, `polarity` and `mirativity`. In order to address these issues, we fit our model with Least Absolute Shrinkage and Selection Operator (lasso) regularization. Lasso regression shrinks coefficients towards zero by introducing a penalty term scaled by a parameter λ (Hastie et al., 2015). In classical lasso regression, a possibly different penalty score is calculated for each predictor. In an extension known as grouped

lasso regression, the set of predictors is partitioned in non-overlapping groups, and a penalty score is calculated for each group of predictors. In the case of lasso regression with categorical predictors, this ensures that levels of a categorical predictor coded as different variables are subject to the same penalty (Yuan and Lin, 2006).

We fit a grouped lasso model of marker choice with the formula specified in (23) using the R package `gglasso` (Yi and Hui, 2015). Categorical variables are treated using one-hot coding, where each level is coded as a separate variable.¹² Variables obtained from the same predictors by one-hot coding are grouped together.

The penalty parameter λ is estimated by 10-fold cross-validation on the model training set. Because non-canonical uses of SR markers are rare and we need them to be represented both in the training and test data, we cannot train and evaluate the model using a single train-test split. We resort instead to nested cross-validation implemented using the `nestedcv` package in R (Lewis et al., 2023). We use Leave-One-Out Cross-Validation (LOOCV) for the outer loop of nested CV and 10-fold cross-validation for the inner loop. That is to say, for each observation in our data set, we train a model on the 761 remaining observations (this is the outer-loop). The best λ parameter for this model is assessed by 10-fold cross-validation on these 761 observations (this is the inner loop). Crucially, when evaluating our model performance in predicting SR marking, we use the model predictions on the holdout data of the outer loop of cross-validation. Note that these predictions are obtained from fitting the model to 762 different training sets, one per fold of LOOCV, although these training sets differ from one another by at most one observation. We report coefficient estimates for a model fitted on the whole data set.

Table 5 presents the coefficient estimates of our model fitted to the complete data set at an optimal value of $\lambda = 3.98e - 4$. Note that we do not provide p -values or 95% confidence intervals.¹³ This is because penalization introduces substantial bias in the estimates, and it is challenging to adjust confidence intervals accordingly (Taylor and Tibshirani, 2015). While there are several proposals for statistical inference with penalized regression, this is an open area of research that goes beyond the scope of the present manuscript (for a recent overview, see Kammer et al., 2022). Note also that the ordered factor `referential_continuity` was coded using polynomial contrasts with linear (L), quadratic (Q) and cubic (C) effects.¹⁴

While we cannot provide confidence intervals for the estimates, we can approximate their sampling distribution using bootstrap resampling (Hastie et al., 2015: §6.2). We draw 1000 samples with replacement of 762 observations from the corpus (as many as there are observations in the corpus). For each sample, we select an optimal value of λ by 10-fold cross-validation, and we fit the model on the whole data set at this value of λ . Figure 2 displays the distribution of model estimates across the bootstrap samples, as well as the proportion of times each estimate is zero in the bootstrap distribution. We observe that, if we disregard outliers, `mirativity`, `polarity` and

¹²When fitting a Lasso regularized regression model, one-hot coding is preferable to dummy coding, since the latter would encode the reference level of categorical predictors in the intercept, but the lasso does not regularize the intercept.

¹³The reader may wonder whether the interpretation of coefficients is reliable given the presence of high multicollinearity. Multicollinearity inflates standard errors of coefficient estimates in traditional regression models, which results in inaccurate tests of significance. Note that this issue is put aside with Lasso regression, since there is no attempt to provide standard errors given the strong bias that is introduced by regularization. The emphasis of Lasso regression is on prediction rather than on inference. Furthermore, note that the coefficients of our model remain fairly stable across bootstrap samples, as can be seen in figure 2, which suggests that multicollinearity among the predictors does not result in instability in variable selection in our data set.

¹⁴These were treated as different groups for regularization.

	Estimate
additivity [additive]	-0.33
additivity [causal]	0.33
clause_type [full]	-0.15
clause_type [reduced]	0.15
mirativity [FALSE]	0.70
mirativity [TRUE]	-0.70
place [different]	0.16
place [containment]	-0.16
polarity [negative]	-0.71
polarity [positive]	0.71
referential_continuity.L	5.64
referential_continuity.Q	2.10
referential_continuity.C	0.33
time [containment]	0.05
time [sequence]	-0.05

Table 5: Model coefficients at $\lambda = 3.98e-4$

referential_continuity are never set to zero in the resampling.

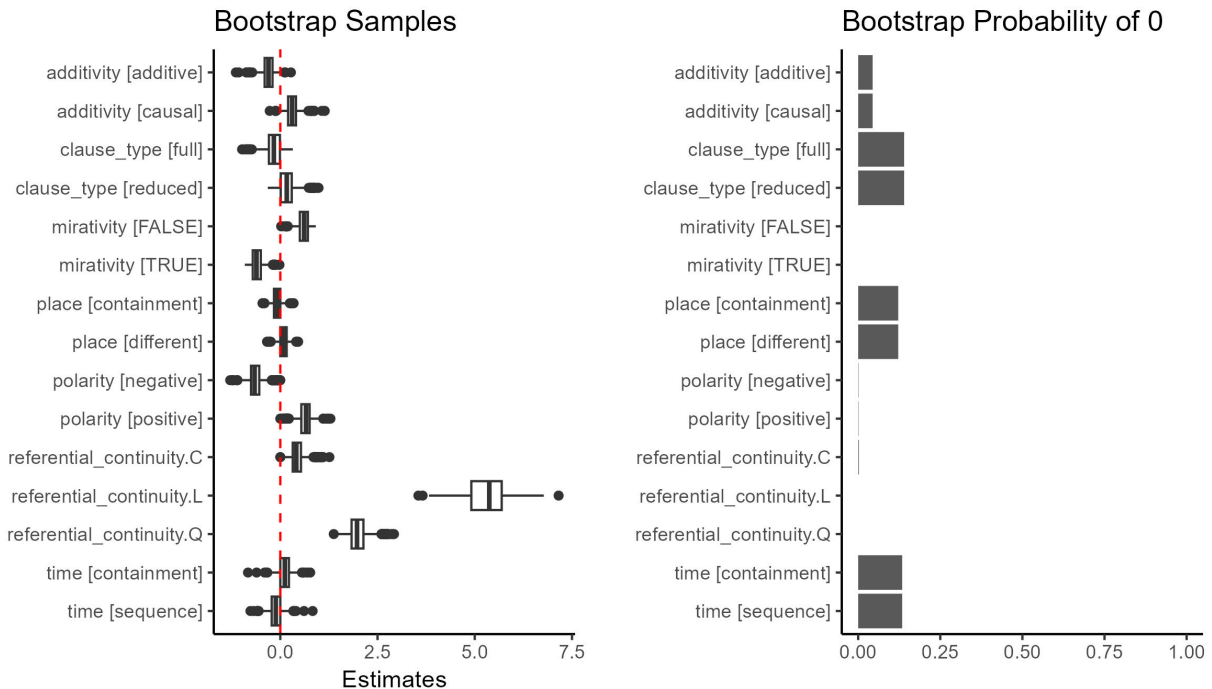


Figure 2: Bootstrap approximation of sampling distribution

It is worth noting that grouped lasso regression performs variable selection in addition to regularization. Indeed, optimization of the objective function for grouped lasso regression may result

in setting coefficient estimates to 0 for some predictors. These predictors are effectively excluded from the model. We can observe in table 5 that no coefficient has been shrunk to 0, therefore all variables have been retained.

We assess the model performance on holdout data of the LOOCV process. The model has excellent discrimination, with a C-index of 0.99.¹⁵ This is not surprising since 98% of all SR markers in the corpus are canonical and can therefore be perfectly predicted by `referential_continuity`. For comparison, a logistic regression model of `marker_type` with pivot coreference as its unique predictor¹⁶ has a C-index of 0.96, and an intercept-only model has a C-index of 0.5.

5.2 Analysis of model predictions

We wish to know (i) whether the frequencies of non-canonical SAME and DIFFERENT markers in data generated by the model match the frequencies observed in the corpus and (ii) whether the contexts in which the model tends to generate non-canonical SAME and DIFFERENT markers match the contexts in which they are observed in the corpus.

5.2.1 Predicted probabilities of SAME marking across contexts

Before we address these questions directly, let us compare the predicted probabilities of SAME marker choice for different classes of SR markers observed in our corpus. The predicted probabilities are obtained from holdout data in the LOOCV process. As a result, a single probability of SAME marking is predicted for each observation in the corpus. We can therefore group these predicted probabilities according to the type of SR markers that are attested in the corpus for their matching observations, and whether these markers are used canonically or not in the corpus. This results in four classes of predictions, each of which corresponds to a class of SR markers attested in the corpus: canonical SAME markers (SM_C), non-canonical DIFFERENT markers (DF_NC), non-canonical SAME marker (SM_NC) and canonical DIFFERENT markers (DF_C). Figure 3 displays the probabilities of SAME marker choice predicted by the model for each class of observations.

Note that observations for which a canonical SAME marker was attested in the corpus are similar to observations for which a non-canonical DIFFERENT marker was attested, insofar as the value of `referential_continuity` is SS for both types of observations. These observations are grouped in the two leftmost box plots in figure 3. Likewise, observations for which a non-canonical SAME marker was attested in the corpus are similar to observations for which a canonical DIFFERENT marker was attested, insofar as the value of `referential_continuity` is not SS for either type of observations. These observations are grouped in the two rightmost box plots.

Figure 3 shows that the probabilities of SAME marking predicted by the model are significantly higher with observations for which a canonical SAME marker is attested in the corpus, than with observations for which a non-canonical DIFFERENT marker is attested (two-sample permutation test for the mean: $p = 4e-04$). In other words, the model correctly predicts that the use of a DIFFERENT marker is significantly more probable with observations for which non-canonical DIFFERENT

¹⁵For binary outcomes, the C-index is a measure of the probability that a randomly selected observation that has a positive value of the outcome variable has a higher predicted probability of positive outcome than a randomly selected observation that has a negative value of the outcome variable. The C-index is equivalent to the area under the Receiver Operating Characteristic curve in the case of binary outcomes (Harrell, 2015: §10.8)

¹⁶Where pivot coreference has two levels: coreferential subjects versus non-coreferential subjects.

markers are actually attested than with observations for which canonical SAME markers are attested.

Likewise, the probabilities of SAME marking predicted by the model are significantly lower with observations for which a canonical DIFFERENT marker is attested in the corpus, than with observations for which a non-canonical SAME marker is attested (two-sample permutation test for the mean: $p = 7e-04$). In other words, the model correctly predicts that the use of a SAME marker is significantly more probable with observations for which non-canonical SAME markers are actually attested than with observations for which canonical DIFFERENT markers are attested.

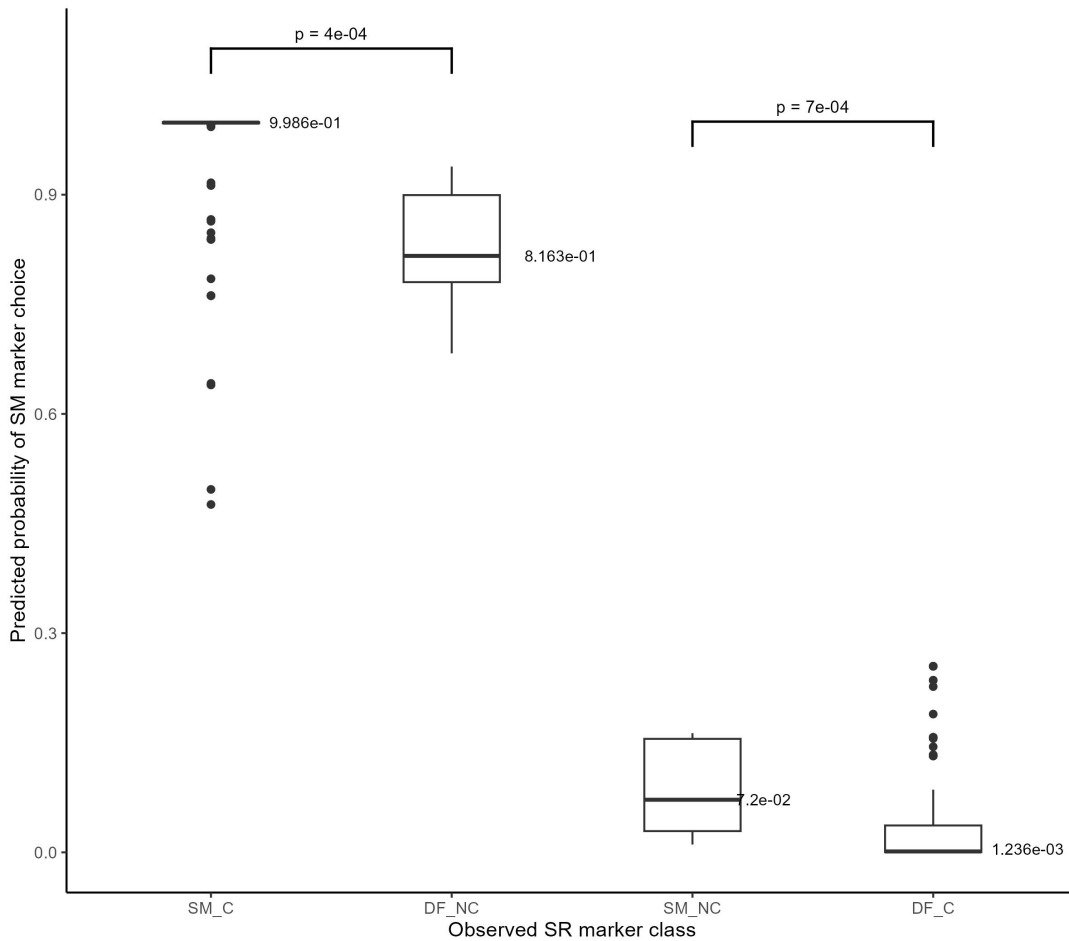


Figure 3: Predicted probabilities of SAME marker choice for different SR marker classes

5.2.2 Frequency of non-canonical SR markers

In order to investigate the model’s predictions more in depth, we use our model to simulate new observations, and we explore the frequency of non-canonical SR markers in these simulations, as well as the contexts in which non-canonical SR markers are generated. To do so, we gather the probabilities of SAME marking predicted by the model¹⁷ into a vector of 762 probabilities, one for each observation in our corpus. For each element p in this vector, we generate an SR marker randomly, with a probability p of generating a SAME marker, and a probability $1 - p$ of generating a DIFFERENT marker. This results in a vector of 762 SR markers. We repeat the process 1,000 times, for a total of 1,000 vectors of 762 SR markers. Note that each SR marker generated in this way is associated with a combination of predictor values. We can therefore determine whether an SR marker generated as part of this simulation is an instance of canonical or non-canonical SR marking: non-canonical markers are SAME markers that were generated with a `referential_continuity` value other than `SS`, or DIFFERENT markers that were generated with an `SS` value of `referential_continuity`.

Figure 4 displays the frequencies of non-canonical SAME and DIFFERENT markers in the simulated data sets, with the mean represented by a red dot in each plot. The frequency of non-canonical markers attested in the corpus is superimposed to each plot as a dashed red line. For both DIFFERENT and SAME markers, we observe that the median frequency of non-canonical markers across simulated data sets is identical to the frequency observed in the corpus. This corresponds to 6 non-canonical DIFFERENT markers and 10 non-canonical SAME markers. In addition, the frequency of non-canonical DIFFERENT markers observed in the corpus is only -0.15 standard deviations away from the mean of simulated frequencies (one-sample permutation test for the mean: $p < 0.0001$). Likewise, the frequency of non-canonical SAME markers observed in the corpus is only -0.15 standard deviations away from the mean of simulated frequencies ($p < 0.0001$). We conclude that the frequencies of non-canonical SR markers in data sets generated by the model are similar on average to the frequencies observed in the corpus.

¹⁷Remember that these are predictions made on holdout data in the leave-one out cross-validation process, therefore the observation for which each prediction is made was not included in the corresponding training set.

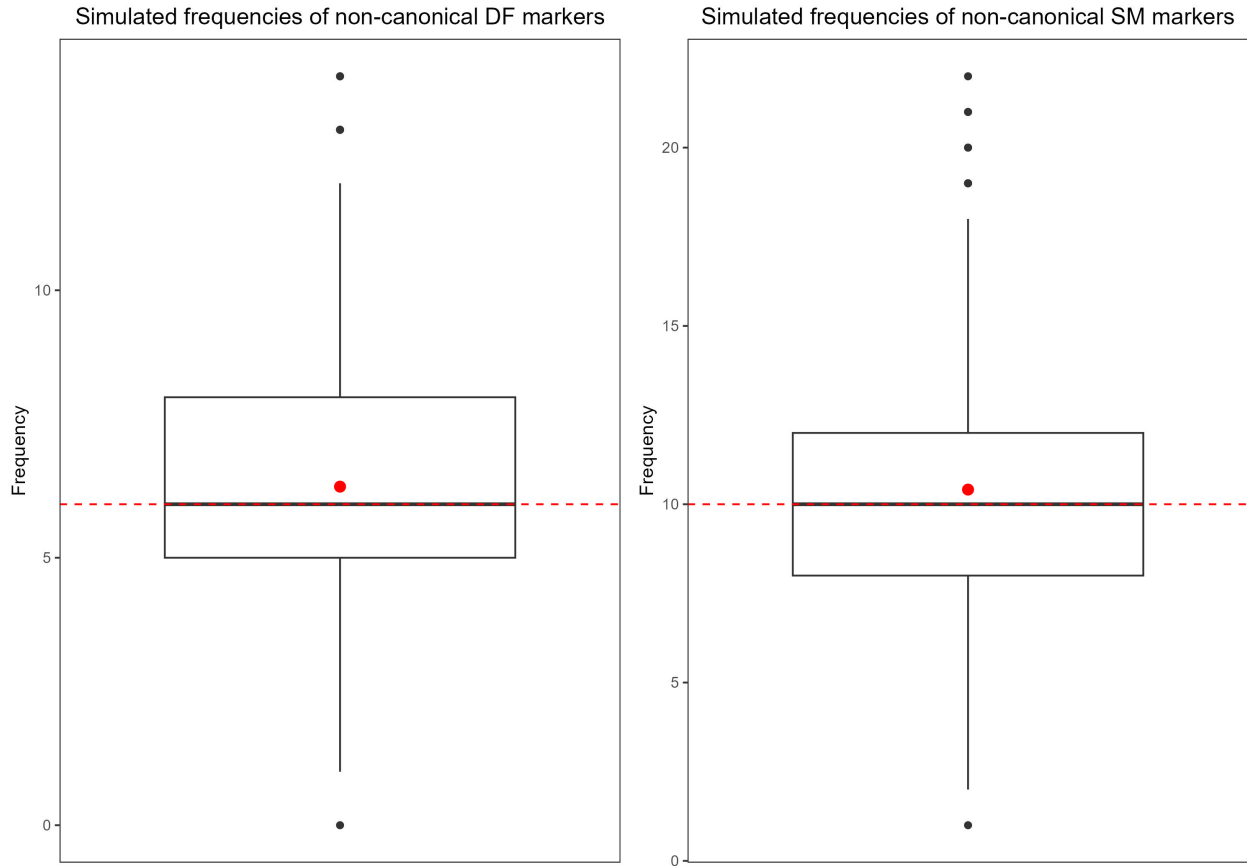


Figure 4: Frequencies of non-canonical SR markers in 1,000 simulated data sets

5.2.3 Contexts of occurrence of non-canonical SR markers

Next, we ask whether the contexts in which the model generates non-canonical SR markers are similar to the contexts in which SR markers are used non-canonically in the corpus. To do so, we merge the 1,000 data sets generated using the procedure described in section 5.2.2 into a single data set with 762,000 simulated observations of SR markers. We then fit a classification tree model of SR marker choice to this data set, as specified in formula (24). Note that our goal in fitting this model is to explore the structure of the simulated data set, rather than to make predictions about new observations.

$$(24) \text{ marker_type} \sim \text{referential_continuity} + \text{polarity} + \text{additivity} + \text{mirativity} + \text{place} + \text{time} + \text{clause_type}.$$

The classification tree is fitted using the `ctree` function of the `partykit` package in R (Hothorn and Zeileis, 2015). The model is fitted by choosing a predictor to split the data set in two subsets. The same process is repeated on each subset, until a stopping criterion is met. At each step of the process, the predictor that is used to perform the split is chosen by performing a series of permutation tests of independence between each predictor and the outcome variable. The predictor with the most

significant association to the outcome variable is selected.¹⁸ The process stops once no predictor is significantly associated with the outcome variable at a level of significance α . Other stopping criteria can be specified, such as maximal tree depth or minimum number of observations in subsets at each split. We set the level for predictor selection at $\alpha = 0.05$, and the maximal tree depth at 3.¹⁹

The fitted tree is displayed in figure 5. At the root of the tree, `referential_continuity` splits the whole data set into two groups: observations with an SS value of this predictor are gathered in the subset to the right²⁰ and all other observations are gathered in the subset to the left. The left subset therefore contains canonical DIFFERENT markers and non-canonical SAME markers, while the right subset contains canonical SAME markers and non-canonical DIFFERENT markers. Focusing on the left subset first, we observe that non-canonical SAME markers are only attested with an OS.SO value of `referential_continuity`, and are more frequent with causal rhetorical relations. Moving on to the right subset, we observe that non-canonical DIFFERENT markers are only attested with NEGATIVE rhetorical relations or in the presence of mirative markers.

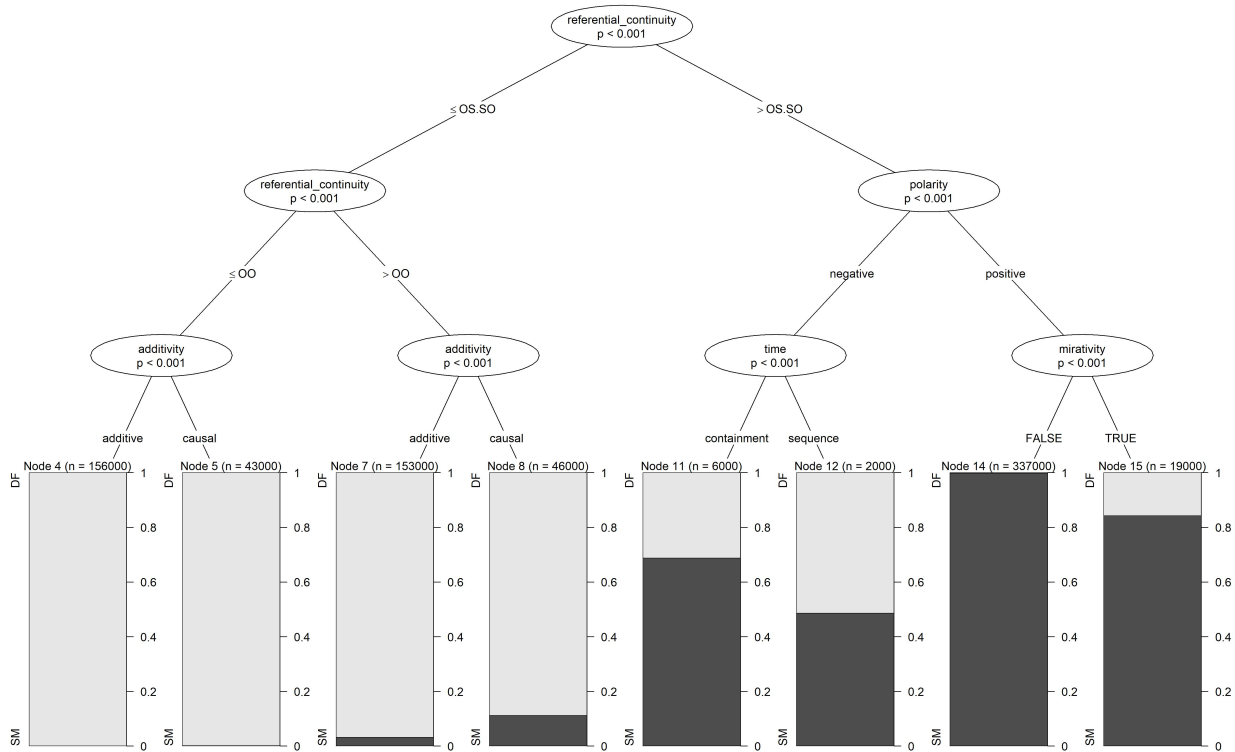


Figure 5: Classification tree of `marker_choice` in simulated data

By comparing this classification tree with tables 3 and 4 of section 4, we observe that the contexts

¹⁸If the predictor has more than two levels, a similar procedure is applied to all possible binary splits of the outcome variable along the predictor, and the split with the most significant association to the outcome variable is chosen.

¹⁹Trees with more depths would have terminal nodes that are subsets of the terminal nodes displayed in figure 5. While such trees might perform differently for classifying new observations, their added complexity is not needed for the exploratory purpose of this section.

²⁰SS is the only level l of `referential_continuity` such that $l > OS.S0$.

in which the model generates non-canonical SAME and DIFFERENT markers correspond to those in which these markers are attested non-canonically in the corpus. In particular, non-canonical uses of SAME markers are only attested with a high degree of referential continuity, and non-canonical DIFFERENT markers are only attested with negative rhetorical relations or in the presence of mirative markers.

In sum, it appears that our model of *marker_type*, although blind to the difference between canonical and non-canonical uses of SR markers, generates non-canonical uses with adequate frequency in adequate contexts.

6 Discussion

The results of the previous section suggest that the distinction between canonical and non-canonical uses of SR markers can be analyzed as a side effect of a probabilistic and multifactorial process of SR marker choice that is itself blind to canonicity. More precisely, it appears that SR marker choice is sensitive to a multiplicity of factors besides referential continuity, including notably mirativity and the polarity of rhetorical relations, which manifest non-referential aspects of thematic continuity. While referential continuity appears to have the highest weight in determining SR marker choice, non-referential aspects of thematic continuity also affect this process, albeit with a lesser weight. The competition between these factors results in a non-negligible probability of non-canonical uses of SR markers in certain contexts. One virtue of such a probabilistic model of SR marker choice is that it allows us to formulate predictions not only about the contexts in which non-canonical uses are likely to be attested, but also about the expected frequencies of non-canonical uses of SR markers in different contexts.

Our analysis fits within a broader tradition of regression- and corpus-based research on syntactic alternations well illustrated by [Bresnan et al.'s \(2007\)](#) work on the English dative alternation (see also [Gries 2017](#) for a review of research in this tradition). The probabilistic and multifactorial models produced in such studies can be readily interpreted within usage-based theories of grammar, which posit that linguistic representations emerge from repeated exposure to particular instances of constructions, and retain specific features of these instances ([Bybee, 2006](#)). In the present study, the SR markers *vy*, *ramo* and *rã* may be seen as constructions whose representations include features that instantiate the independent variables of our model, such as referential continuity, mirativity and polarity. Crucially, the strength of association between such features and the phonetic form of the markers varies with the frequency of occurrence of the markers in the different contexts in which the speaker was exposed to their use ([Bybee, 2006](#): §13). It is these variable associations of form and meaning that determine the probabilities of canonical and non-canonical uses of SR markers in production.

Our analysis contrasts with categorical accounts of SR marker choice, according to which SAME and DIFFERENT markers are associated with necessary and sufficient conditions of use. One example of such an account is [Stirling's \(1993\)](#) analysis of SR marker choice in Eastern Pomo, Lenakel and Amele. Here, we illustrate with Stirling's analysis of Lenakel. [Stirling \(1993: 152\)](#) observes that Lenakel speakers use Same Subject (SS) markers unless the marked and reference clauses have different tenses or the reference of the main protagonist changes across clauses, in which case a Different Subject (DS) marker is used:

(25) For Lenakel:

if tense changes, use DS;

otherwise, if reference changes, use DS;

otherwise, if tense and reference stay the same, use SS. (Stirling, 1993: 152)

Stirling argues that SR markers indicate agreement between parameters of the events described by the marked and reference clauses. In the case of Lenakel, a difference between future and non-future tense is argued to indicate that the two events differ in actuality. Consequently, if the marked and reference clauses describe events a_i and a_j , SS marking indicates that $\text{Protagonist}(a_i) = \text{Protagonist}(a_j) \ \& \ \text{Actuality}(a_i) = \text{Actuality}(a_j)$, while DS marking indicates that $\text{Protagonist}(a_i) \neq \text{Protagonist}(a_j) \ \vee \ \text{Actuality}(a_i) \neq \text{Actuality}(a_j)$.

Our argument against categorical analyses of SR marker choice in Mbyá was that canonical uses of SR markers are attested in virtually all contexts where non-canonical uses are attested. Therefore, there is no combination of predictor values that act as necessary and sufficient conditions for SAME or DIFFERENT marker choice. This supports a probabilistic approach to SR marker choice, which attempts to model the relative frequencies of canonical and non-canonical uses of SR markers in different contexts.

An analysis of SR related to ours was defended in Pustet's (2013) study of clause linkers in Lakota.²¹ Pustet argues that the function of Lakota clause linkers is to indicate higher or lower degrees of discourse cohesion (which corresponds to what we called thematic continuity in the present manuscript). Furthermore, she argues that discourse cohesion in Lakota should be decomposed into four parameters: "Switch Reference" (i.e., referential continuity across pivots), "probability" (i.e., whether the second clause in a linkage construction expresses a deviation from the expected course of events), temporal cohesion and contrast. A limitation of Pustet's study is that while it offers a quantitative analysis of the association between clause linker choice and these different parameters, it does not present a multivariate model of clause linker choice but relies on a collection of univariate models instead. In the spirit of Pustet's (2013) study, we advocated for a multifactorial analysis of SR marker choice that acknowledges both the referential and the non-referential dimensions of thematic continuity. A further contribution of the present manuscript is the implementation of such an analysis as a probabilistic model of SR marker choice, and the discussion of strategies for evaluating the extent to which such a model captures both canonical and non-canonical uses of SR markers.

Glossing abbreviations for Mbyá examples

ALSG: first person singular active; ABL: ablative; ANA: anaphoric pronoun; ASP: aspect; ATTN: attention; BLSG: first person singular inactive inflection; BDY: information structure boundary; CF: counterfactual; CONV: converb; COMIT: comitative causative; COMP: completive marker; CONT: continuative; DAT: dative; DIM: diminutive; DIR: directional; DF: DIFFERENT Switch-Reference marker;

²¹Note that Pustet (2013) does not use the term "Switch Reference" to refer to a class of expressions, but rather to the reference tracking function of clause linkers. What she calls "Switch Reference" corresponds roughly to our `referential_continuity` predictor.

EXCL: exclusive; FRUST: frustrative; FUT: future; HAB: habitual; HSY: hearsay evidential; INT: intensifier; LOC: locative; MAN: manner; MIR: mirative; NEG: negation; NMLZ: nominalizer; NPOSSD non-possessed; OBJ: object marker; Q: question particle; PL: plural; R: linking morpheme; SM: SAME Switch-Reference marker; SRC: source; SEQ: sequential.

References

- Asher, N. and Lascarides, A. (2003). *Logics of Conversation*. Cambridge University Press.
- Baron, R. M. and Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, 51:1173–1182.
- Behrens, B. and Fabricius-Hansen, C. (2010). The relation accompanying circumstance across languages: conflict between linguistic expression and discourse subordination? In Shu, D. and Turner, K., editors, *Contrasting meaning in languages of the East and West*, pages 531–552. Peter Lang.
- Behrens, B., Fabricius-Hansen, C., and Solfjeld, K. (2012). Competing structures: the discourse perspective. In Cathrine, Solfjeld, K., and Haug, D., editors, *Big events, small clauses*, pages 179–225. De Gruyter.
- Bresnan, J., Cueni, A., Nikitina, T., and Baayen, H. (2007). Predicting the dative alternation. In Bouma, G., Krämer, I. M., and Zwarts, J., editors, *Cognitive foundations of interpretation*, pages 69–94. Royal Netherlands Academy of Arts and Sciences.
- Bybee, J. (2006). From usage to grammar: the mind’s response to repetition. *Language*, (82):711–733.
- DeLancey, S. (1997). Mirativity: The grammatical marking of unexpected information. *Linguistic Typology*, 1(1):33–52.
- Dooley, R. A. (1989). Switch reference in Mbyá Guaraní: a fair-weather phenomenon. *Work Papers of the Summer Institute of Linguistics, U. of North Dakota Session*, 33:93–119.
- Dooley, R. A. (1992). When switch reference moves to discourse: developmental markers in Mbyá Guaraní. In Hwang, S. J. J. and Merrified, W. R., editors, *Language in context: essays for Robert E. Longacre*, pages 97–108. Summer Institute of Linguistics and The University of Texas at Arlington Publications in Linguistics.
- Dooley, R. A. (1999). A noncategorical approach to coherence relations: switch reference constructions in Mbyá Guaraní. In Loos, E., editor, *Logical relations in discourse*, pages 219–242. Summer Institute of Linguistics.
- Dooley, R. A. (2011). Mbyá Guaraní collection of Robert Dooley. www.ailla.utexas.org. Media: text. Access: 100% restricted. PID ailla:119734.

- Dooley, R. A. (2015). *Léxico Guarani, dialeto Mbyá: Introdução*. Associação Internacional de Linguística - SIL Brasil.
- Dooley, R. A. (2016). *Léxico Guarani, dialeto Mbyá*. Associação Internacional de Linguística - SIL Brasil.
- Givón, T. (2001). *Syntax: an Introduction*, volume 2. John Benjamins Publishing Company.
- Gries, S. T. (2017). Syntactic alternation research: Taking stock and some suggestions for the future. *Belgian Journal of Linguistics*, 31(1):8–29.
- Grosz, B., Joshi, A. K., and Weinstein, S. (1995). Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):203–225.
- Grosz, B. and Sidner, C. L. (1986). Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3):175–204.
- Guérin, V. (2019). *Bridging constructions*, volume 24. Language Science Press.
- Guillaume, A. (2011). Subordinate clauses, switch-reference, and tail-head linkage in Cavineña narratives. In van Gijn, R., Haude, K., and Muysken, P., editors, *Subordination in Native South American Languages*, pages 109–140. John Benjamins.
- Haiman, J. and Munro, P. (1983). Introduction. In Haiman, J. and Munro, P., editors, *Switch-reference and universal grammar*, pages ix–xv. John Benjamins.
- Harrell, F. E. (2015). *Regression Modelling Strategies*. Springer.
- Hastie, T., Tibshirani, R., and Wainwright, M. (2015). *Statistical Learning with Sparsity*, volume 143. CRC Press.
- Hobbs, J. R. (1985). On the coherence and structure of discourse. Technical Report 85-37, Center for the Study of Language and Information, Stanford, CA. CSLI Publications.
- Hothorn, T. and Zeileis, A. (2015). partykit: A modular toolkit for recursive partytioning in r. *Journal of Machine Learning Research*, 16(118):3905–3909.
- Kammer, M., Dunkler, D., Michiels, S., and Heinze, G. (2022). Evaluating methods for lasso selective inference in biomedical research: a comparative simulation study. *BMC Medical Research Methodology*, 22(206):1–13.
- Kamp, H. and Reyle, U. (1993). *From Discourse to Logic*. Kluwer Academic Publishers.
- Ladeira, M. I. (2018). Guarani mbya. In Ricardo, F. P., editor, *Povos Indígenas no Brasil*. Instituto Socioambiental.
- Lewis, M. J., Spiliopoulou, A., Goldmann, K., Pitzalis, C., McKeigue, P., and Barnes, M. R. (2023). nestedcv: an R package for fast implementation of nested cross-validation with embedded feature selection designed for transcriptomics and high-dimensional data. *Bioinformatics Advances*, 3(1):vbad048.

- Mann, W. C. and Thompson, S. A. (1988). Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243–281.
- McKenzie, A. (2014). On the emergence of discourse functions from semantics and pragmatics. Invited talk, 8th Conference on the Semantics of Under-represented Languages of the Americas (SULA 8).
- Mithun, M. (1993). Switch-reference: Clause combining in central pomo. *International Journal of American Linguistics*, 59(2):119–136.
- Mithun, M. (2021). Switch-reference: Clause combining in central pomo. In Epps, P., Law, D., and Pat-El, N., editors, *Historical Linguistics and Endangered Languages*, pages 182–211. Routledge.
- Muller, P., Vergez-Couret, M., Prévot, L., Asher, N., Farah, B., Bras, M., Le Draoulec, A., and Vieu, L. (2012). Manuel d’annotation en relations de discours du projet ANNODIS. Technical Report 21, CLLE-ERSS, Toulouse.
- Pustet, R. (2013). Switch-reference or coordination? A quantitative approach to clause linkage in Lakota. *International Journal of American Linguistics*, 79:153–188.
- Reese, B., Hunter, J., Asher, N., Denis, P., and Baldrige, J. (2007). *Reference Manual for the Analysis and Annotation of Rhetorical Structure*. inria-00514933.
- Roberts, J. (1988). Amele switch-reference and the theory of grammar. *Linguistics Inquiry*, 19:45–63.
- Roberts, J. (2017). A typology of switch reference. In Aikhenvald, A. Y., editor, *The Cambridge Handbook of Linguistic Typology*, pages 538–573. Cambridge University Press.
- Sanders, T. J., Demberg, V., Hoek, J., Scholman, M. C., Asr, F. T., Zufferey, S., and Evers-Vermeul, J. (2021). Unifying dimensions in coherence relations: How various annotation frameworks are related. *Corpus Linguistics and Linguistic Theory*, 17(1):1–71.
- Stirling, L. (1993). *Switch-reference and discourse representation*. Cambridge University Press.
- Taylor, J. and Tibshirani, R. (2015). Statistical learning and selective inference. *Proceedings of the National Academy of Sciences*, 112(25):7629–7634.
- Thomas, G., Antono, G., Bradford, L., Kiss, A., and Winkelman, D. (2021). Switch-reference and its role in referential choice in mbyá guaraní narratives. *Corpus Linguistics and Linguistic Theory*, 17(3):563–597.
- van Gijn, R. (2012). Switch-attention (aka switch-reference) in South-American temporal clauses: Facilitating oral transmission. *Linguistic Discovery*, 1(10):112–127.
- van Gijn, R. (2016a). Switch-reference: An overview. In van Gijn, R. and Hammond, J., editors, *Switch-reference 2.0*, volume 114 of *Typological Studies in Language*, pages 1–54. John Benjamins.

- van Gijn, R. (2016b). Switch-reference in western South-America. In van Gijn, R. and Hammond, J., editors, *Switch-reference 2.0*, volume 114 of *Typological Studies in Language*, pages 53–206. John Benjamins.
- Veríssimo, T. A. (2002a). *Opa mba'e re nhanhembo'e aguã 1*. Nhombo'ea Guarani: Mundo Indígena.
- Veríssimo, T. A. (2002b). *Opa mba'e re nhanhembo'e aguã 2*. Nhombo'ea Guarani: Mundo Indígena.
- Watkins, L. J. (1993). The discourse functions of kiowa switch-reference. *International Journal of American Linguistics*, 59(2):137–164.
- Yi, Y. and Hui, Z. (2015). A fast unified algorithm for solving group-lasso penalized learning problems. *Statistics and Computing*, 25:1129–1141.
- Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67.