THE ROLE OF SEMANTIC REFERENTIAL DEFICIENCY IN CABO-VERDEAN CREOLE SUBJECT EXPRESSION: VARIABLE CHOICE OF THIRD PERSON REDUCED REFERRING DEVICE

O PAPEL DA DEFICIÊNCIA SEMÂNTICA REFERENCIAL NA EXPRESSÃO DO SUJEITO NO CRIOULO CABO-VERDIANO: ESCOLHA VARIÁVEL DOS RECURSOS REFERENCIAS REDUZIDOS DA TERCEIRA PESSOA

Adrián Rodríguez-Riccelli[1]
The State University of New York at Buffalo

Abstract: This study examines the influence of 'semantic referential deficiency'—consisting of nonhuman, nonspecific, and indefinite reference—in variable third-person subject expression in a corpus of naturalistic Cabo-Verdean Creole discourse collected from respondents from the islands of Santiago and Maio. The methodology follows the Probabilistic Linguistics program (CLAES, 2017) in combining variationist sociolinguistics with cognitively-oriented discourse analysis, whereas the notion of semantic referential deficiency is adopted from Generative Grammar and from research on Brazilian Portuguese argument expression. The coded corpus data were submitted to a suite of descriptive and inferential statistical analyses in R (R CORE TEAM, 2021). Results show a promoting effect from nonhuman and collective referents on the selection of zero or null subjects, alongside other predictors related to referential coherence and accessibility.

Keywords: Cabo-Verdean Creole; Null subject; Animacy; Subject clitic; Discourse anaphora.

[1] arricel@buffalo.edu

Resumo: *Nesta pesquisa investiga-se a influência da 'deficiência semântica referencial' —a qual consiste da referência não humana, não específica e indefinida— na expressão variável dos sujeitos da terceira pessoa nas variedades do crioulo cabo-verdiano faladas nas ilhas de Santiago e do Maio. A metodologia que se segue é a Linguística Probabilística (CLAES, 2017), que mistura o variacionismo-sociolinguístico com a análise discursiva orientada à cognição, embora na pesquisa atual o foco interdisciplinar é ampliado ainda mais, já que se toma emprestado da Gramática Gerativa, e das pesquisas na expressão de argumentos no português brasileiro, o conceito da deficiência referencial semântica. O corpus foi sometido a uma série de análises estatísticas descritivas e inferenciais no R (R CORE TEAM, 2021). Os resultados mostram um efeito promotor dos referentes não humanos e coletivos na seleção de sujeitos nulos, em conjunto com outros preditores que se relacionam com a acessibilidade referencial.*

Palavras-chave: *Crioulo cabo-verdiano; Sujeito nulo; Animacidade; Clítico de sujeito; Anáfora discursiva.*

INTRODUCTION

Following formalizations of the semantics of *strong* and *deficient* pronominals in the Generative Grammar tradition (RIZZI, 1986; CARDINALETTI & STARKE, 1999; *inter alia*.), the foundational work of linguists such as Maria Eugênia L. Duarte, Mary A. Kato, and Sonia M. L. Cyrino, among others, compiled a preponderance of empirical evidence to demonstrate the powerful conditioning effect of *semantic referential deficiency* on the morphophonological form selected to encode verbal arguments in Brazilian Portuguese (BP) discourse-anaphoric reference. Cardinaletti's and Starke's (1999) use of the term 'semantic referential deficiency' referred to the semantic specifications associated with different classes of pronoun and the type of referents that they can index: *strong* pronouns index fully semantically specified referents with the properties [+human], [+specific], [+definite]; on the other hand, *deficient* pronouns (divided into *weak* and *clitic* in trinary classifications), can also index [-human], [-specific], [-definite] referents.

In the present study, we explore the role of semantic referential deficiency in conditioning the discourse-anaphoric referring devices selected in subject expression in the Portuguese-lexified language Cabo-Verdean Creole (CVC). Drawing on a corpus of naturalistic speech collected from speakers from the islands of Santiago and Maio, we conducted a series of quantitative analyses using variationist sociolinguistic methods informed by cognitively-oriented discourse analysis, as well the aforementioned notion from Generative Grammar. We show that, just as in BP, semantic referential deficiency plays a major role in constraining the organization of the pronominal paradigm and patterns of subject expression in CVC.

In CVC subject reference, arguments are represented in most active voice constructions with a *singleton atonic person marker* historically classified as a *subject clitic* (see Tables 1 and 2 below for two classifications of the nominative person marker/subject pronoun paradigm). The third-person atonic nominative markers in the Santiago (popularly called *Badiu* < Pt. '*vadio*') and Maio (popularly called *Djarmai* < Pt. '*Ilha do Maio*') varieties are singular *=e(l)=* and plural *=es=*, where the equals signs indicate that they can attach to a phonological host in either direction (though usually to elements in the Verb Phrase during rightward attachment), and where the singular form variably occurs with a liquid coda (they do not encode grammatical gender). In example (1), the subject referent is introduced as a Noun/Determiner Phrase in one clause and is then resumed by *e* in a subsequent independent clause.

1) *Ten* *[un* *santu]*$_i$ *[…]* ***e***$_i$*=txoma, uhh, a* *Nossa*
PRES DET.INDEF saint 3.SG=call DM DET POSS.1PL.F
Sinhora di Fátima.
lady of F.
'There is a saint […] she is called, umm, Our Lady of Fatima'. (P20, *Marluce*,[2] F, dialect zone: *Santiago sul*, age: 22)

[2] Respondents have been assigned pseudonyms.

This can be considered the default form of nominative referential expression in CVC. For example, a quantitative analysis of nominative person marking in the Santiago and Maio varieties of CVC found rates of 81.9% subject clitic expression across 3,651 observations of all person-number iterations in the paradigm (RODRÍGUEZ-RICCELLI, 2021) (see Figure 4, Section 4, below).

Reduced referring devices are deictic (i.e., Speech Act Participant) or discourse-anaphoric (i.e., external reference) morphemes (including phonologically vacuous null/zero 'morphemes') selected to *index* or *resume* a referent associated with a verbal argument (KIBRIK, 2012). When reduced referring devices are taken to be a matter of speaker's choice (as is assumed if applying the notion of *the Principle of Accountability* [LABOV, 1969] to subject expression), then, in addition to the singleton clitic, there are two reduced referring devices that speakers of CVC might choose from based on the discursive and morphosyntactic context: (2) a *double subject pronoun construction* or *conomination* (see, HASPELMATH, 2013) in which a tonic independent pronoun (in this case, *el*) 'doubles' or indexes a second time the same subject referent as the clitic (in this case, *e*) in the same local clause; or, (3) a null or zero-subject.[3]

2)

E$_i$=ka	*podi sa ta fazi kantu*	*trabadju,*	*kiii*	***el***$_i$	***e***$_i$*=sa ta*
3.SG=NEG	can TMA do so much work		COMP	3.SG	3.SG=TMA

ganha txeu, ki	*kelotu sa ta*	*ganha poku!*
earn a lot COMP	other TMA	earn little

'He can't do so much of the work, that he earns so much, while that other guy earns so little!' (P28, *Nunú,* M, dialect zone: *Santiago centro*, age: 28)

[3] First observed in Baptista (2002: p. 259-260), Rodríguez-Riccelli (2019, 2021) refers to these type of null subjects as "*true zero-subject*s" since there is no person marking whatsoever in the local clause. Contrast this with the use of null/zero subjects in 'prototypical *pro*-drop' languages, which involves person marking solely on verbal suffixes, while the independent tonic pronoun is excluded or rendered phonologically null.

3) *E$_i$=kai na lagoua, i ali, dja Ø$_i$ labanta*
3.SG$_i$=fall in pond CONJ here TMA 3.SG$_i$ rise
'He fell in the pond, and here, now [he] rises.' (P16, *Ebanir,* M, dialect zone: *Santiago centro* [childhood]; *Santiago sul* [adluthood], age: 19).

In variationist sociolinguistic inferential statistical analysis, the numerical and visual output usually provides some indication of the *probability* (e.g., the *odds ratios*, *log odds*, or *predicted probabilities*) with which a given *dependent, outcome,* or *response variable* will occur given the conditions associated with the set of *predictor* or *independent variables* that were hypothesized by the researcher to condition some outcome. In the case of CVC subject expression outlined thus far, such an analysis would provide some information about the probability that a double subject pronoun construction (2) or a null subject (3) would be chosen by the speaker over a singleton clitic (1), given a set of predictor variables representing a range of morphosyntactic, discourse, and individual speaker-specific sociolinguistic conditions (and given certain distributional characteristics of the dataset).

This was the method applied in recent variationist studies on subject expression in CVC (RODRÍGUEZ-RICCELLI 2019, 2021); the findings for the discursive and morphosyntactic distribution of double subject constructions and null subjects (compared against the singleton subject clitic baseline) can be summarized as follows (see also, RODRÍGUEZ-RICCELLI, [2021: 159, Table 7]): i.) Double subject constructions were identified as switch reference and contrastive devices, resembling the 'doubling' of inflectional suffixes by *independent tonic pronouns* in languages traditionally considered 'prototypically *pro*-drop'; ii.) Null subjects were promoted in the third-person by a *persistence of* *morphophonological form* effect (i.e., *perseveration*),[4] and by prosodic and syntactic linking across adjacent clauses containing antecedent and anaphor (e.g., Chafe 1987, 1993, 1994; Du Bois *et al.* 1993; Torres Cacoullos & Travis 2019). Further, with respect to the issue of semantic referential deficiency, in Rodríguez-Riccelli (2019, 2021) nonhuman and/or indefinite/nonspecific and collective referents strongly promoted resumption with a discourse-anaphoric zero/null subject, as in (4).[5]

4)	*Morna$_i$ abes*	*Ø$_i$*	*ta*	*leba txeu instrumentu,*
	morna$_i$ sometimes	3.SG$_i$	TMA	take a.lot instrument

Ø$_i$ *ta*	*leba violinhu*
Ø$_i$ TMA	take guitar

‘Morna, sometimes [it] has a lot of instruments, [it] has guitar.’ (P3, *Danilo,* M, dialect zone: *Santiago norte*, age: 30)

Unlike null subjects, Baptista (2002: p. 240-241) asserted that double subject pronoun constructions should never resume an antecedent bearing nonhuman reference, based on data from her corpus of Sotavento CVC (5a,b).[6]

[4] For a recent example of the persistence/perseveration effect in subject expression (in Spanish bilinguals), see Prada Pérez (2020).

[5] In this linguistic example, the first lexical nominal (which introduces the referent), *Morna* (a Cabo-Verdean genre of music), occurs at the left-edge of the clause where it is separated from the ‘subject clitic slot’ by the adverb *abes* ‘sometimes’, and the clitic slot is empty (though it need not be, i.e., the clitic is variable in that context). When these constructions were coded as antecedents, they were assigned a unique category of *DP + intervening material + Ø* (see Sections 3 and 4 below). The [-human] referent *Morna* is then resumed by the target subject token of interest which occurs in the following *intonational unit* (see, Section 3 below) and independent clause: a null subject.

[6] These examples involve the doubling of an object person marker by a tonic independent pronoun, but the same holds for the subject clitic (5c,d; introspectively contrived examples):

c.)								
	Omi$_i$ kumesa	*ta*	*balansia*	*n-um*		*rama*	*di*	*arvi,*
	man start	TMA	sway	PREP-DET.INDEF		branch	PREP	tree
	[*Kelotu*	*mos*]$_j$	*ben*	*ku*	*redi di saguridad*		*ma,*	
	DEM.3SG	guy	come	PREP	security.net		CONJ	
	ael$_i$*,*	***e***$_i$*=kai*		*di*	*arvi*			
	3.SG		3SG=fall		PREP	tree		

(5) a.) ***Ael***$_i$, *N*$_j$=*gosta* *d*-***el***$_i$ [+human] (object reference)

3SG 1SG=like PREP-3SG

b.) ****Ael***$_i$, *N*$_j$=*gosta* *d*-***el***$_i$ [-human] (object reference)

3SG 1SG=like PREP-3SG

The role of semantic referential deficiency in CVC subject expression is further analyzed in the present study. Previous variationist studies of CVC subject expression modeled the entire person marking paradigm of reduced referring devices. Given the Principle of Accountability, this means that inanimate referents were excluded from those previous analyses, thus preventing a full consideration of semantic referential deficiency (in the inferential statistics) beyond the property of specificity. Thus, the present study models third-person subject reference apart from all others in order to further explore the role played by nonhuman and indefinite/nonspecific reference in the effect exerted by semantically referentially deficient antecedents on speaker's choice of nominative referring expression. For reasons also related to the Principle of Accountability, and as is explained further in Section 2 below, in the inferential statistical analysis advanced in the present study, double subject pronoun constructions are not considered, and only variation between a null subject and subject clitic is considered.

'A man started swaying on a tree trunk. This guy came with a security net but, he fell from the tree.'

d.) d.) *Mangui*$_i$ *kumesa ta balansia n=um rama di arvi,*

mango start TMA sway PREP=DET.INDEF branch PREP tree

ma mbora bentu para di sopra

DM although wind stop PREP blow

(****ael***$_i$), (***e***$_i$)=*kai di arvi*

3.SG 3SG=fall PREP tree

'A mango starts swaying on a tree trunk. Even though the wind stopped blowing, it fell from the tree.'

In Section 1, we describe the view of argument expression in *Probabilistic Linguistics* (CLAES, 2O17). We then turn to a basic description of third-person subject pronouns and other nominative referring devices in CVC in Section 2, including a discussion of previous research on the role of semantic referential deficiency in this language and in BP. Comparing subject expression in CVC and BP is not a primary goal of the present study (that topic is reserved for future analysis), but previous research on BP argument expression directly informed the way that semantic referential deficiency was conceptualized and represented in the quantitative analyses. As such, a brief summary of semantic referential deficiency in BP is warranted. In Section 3, we describe the fieldwork and data collection from which the corpus was compiled, and the statistical procedures and analyses applied to it. We present the results in Section 4 and draw some generalizations in the discussion in Section 5. To conclude, we consider whether semantic referential deficiency is a relevant property for argument expression in Lusophone varieties, in so-called topic-prominent languages, or simply in languages in which the same pronouns or person markers are used to resume fully specified and semantically deficient referents alike.

1 VARIABLE CHOICE OF REFERRING EXPRESSION IN 'PROBABALISTIC LINGUISTICS'

The present study proceeds within the investigative program proposed by Jeroen Claes (2017), Probabilistic Linguistics. This approach fuses the precise quantitative tools of variationist sociolinguistics with theoretical and explanatory frameworks from related subfields of linguistic inquiry such as usage-based approaches, Functionalism, and Cognitive Linguistics. The present study widens the inter-sub-disciplinary scope yet further by also engaging research in the Generative Grammar tradition, with the qualification that the linguistic properties described be understood as *probabilistic tendencies*, rather than discrete features.[7]

Under this approach, "referential coherence"—or the organizational properties constraining naturalistic discourse reference—is conceptualized as "mental processing instructions" (GIVÓN, 1992: p. 5-56). Discourse-anaphoric markers of grammatical person thus serve as reference points in "discourse-model management procedures used by speakers and hearers to adjust or maintain the accessibility (activation or saliency) level of referents in the evolving mental model of the discourse." (SIEWIERSKA, 2004: p. 174). Therefore, the morphophonological form of the referring device selected to encode a given referent—in those instances in which there are two or more referring devices to choose from—is determined by referential *continuity*, *cohesion*, or *coherence*, or by the extent to which a given referent is *activated* or *accessible* (GIVÓN, 1976, 1983, 1992, 2017; ARIEL, 1990, 2001; CHAFE, 1987, 1994; CORNISH, 1999; HUANG, 2000; *inter alia.*), all of which may be subsumed under processes of working memory (KIBRIK, 1997, 2011).

A confluence of factors are assumed to inform the degree of referential accessibility associated with a given referent in the discourse, among these: the morphophonological form selected for that referent at last mention, the morphosyntactic and discourse configuration between a referring device or discourse anaphor and its referent's last mention, their respective morphosyntactic roles in their local clause, the grammatical person-number of the referent, and the animacy and definiteness/specificity of the referent, among many other conditions.

[7] A similar approach, in which considerations of probabilistic linguistic variation inform (micro-)parametric models of typological variation, has enjoyed a strong tradition in research on BP morphosyntax and discourse (TARALLO, 2015[1987]; TARALLO & KATO, 2007[1989]; *inter alia.*). See Section 5 for further discussion.

Reduced referring devices are considered the most referentially continuous. This includes atonic person markers like clitic pronouns and zero/null anaphora, as can be seen in Givón's (2017: p. 6) depiction of "referent coding devices and referential continuity" (Figure 1).

Figure 1. "Referent coding devices and referential continuity"

highest continuity

a.) zero anaphora
b.) unstressed anaphoric pronouns
c.) stressed independent pronouns
d.) definite NPs
e.) indefinite NPs
f.) modified NPs

lowest continuity

These notions have found empirical support when represented as predictor or independent variables in statistical models of referential choice in production studies based on naturalistic discourse, in studies of linguistic corpora, or in reference processing studies based on experimental methods. For instance, the degree of prosodic and morphosyntactic *interclausal linking* between *intonational* or *discourse units* has been observed to promote null subject expression, and this appears to obtain independently of language-specific constraints on pronominal expression (TORRES CACOULLOS & TRAVIS, 2019).

In a comparative analysis of the variable-rule systems underlying variation between overt and null subject expression in Spanish and English discourse, Torres Cacoullos and Travis (2019) showed that maximal interclausal linking—by means of prosodic continuity characterized by rising or steady intonation at the right-edge boundary of the intonational unit, or by means of syntactic linking through the use of conjunctions and discourse markers at either of the intonational unit's edges—promoted null subject expression in both languages once the *variable context* or *envelope of variation* had been delimited according to language-specific constraints on subject expression observable in their corpora. Following that study, interclausal linking was also found to promote null subject expression in quantitative models of nominative referential choice in CVC (RODRÍGUEZ-RICCELLI, 2019, 2021). In Section 3 below, more details are provided on how intonational units were defined and how interclausal linking was identified, and four examples are given demonstrating each condition: [+/- prosodic linking] and [+/- syntactic linking].

Another important probabilistic factor is also related to referential accessibility and appears to be active cross-linguistically when language specific conditions are held constant: *persistence of morphophonological form* (also called *perseveration*). Persistence refers to the tendency for the morphophonological form of a referring device to be repeatedly selected upon resumption of the same referent, usually when antecedent and anaphor are proximate to one another in the discourse. For example, when an activated referent is resumed by a null subject, subsequent and proximate resumptions of that referent are more likely to redeploy a null subject. This often takes the form of *anaphoric chains*, as in (5), below, drawn from the present corpus. Note, also, that in (5), each of the successive clauses or intonational units are also prosodically linked, in that there is rising or steady intonation at the end or right-edge of each intonational unit; this is represented orthographically with a comma and a line break (as in, CHAFE, 1993, 1994; DU BOIS *et al.*, 1993). Such prosodic linking across units of discourse is thought to enhance referential accessibility, thereby increasing the probability of null subject expression when there is coreference across the adjacent intonational units, as well as when other predictive factors intervene, such semantic referential deficiency (in the case of [5], an indefinite/nonspecific referent introduced by the lexical nominal *kriansa* 'children' [plural number is contextually interpreted]).

6)	*I*	***kriansa***$_i$		*ta*	*brinka*	*ku-el*$_j$*,*		
	CONJ	children		TMA	play	PREP-3sg.OBL		
	Ø$_i$	*ta*	*fazi*	*m-Ø*$_j$*-é*			*di*	*baka,*
	3PL.SBJ	TMA	make	COMP-3SG-COP			PREP	cow
	Ø$_i$	*ta*	*brinka*	*ku-el*$_j$*,*				
	3PL.SBJ	TMA	play	PREP-3SG.OBL				
	Ø$_i$	*ta*	*toka*	*Ø*$_j$	*dianti,*			
	3PL.SBJ	TMA	touch	3SG.OBJ	in.front			
	Ø$_i$	*ta*	*pega*	*munti*				
	3PL.SBJ	TMA	catch	many				
	Ø$_i$	*ta*	*inxi*	*garrafa,*				
	3PL.SBJ	TMA	fill	bottle				
	Ø$_i$	*ta*	*bira*		*ta*	*brinka*	*ku-el*$_j$*.*	
	3PL.SBJ	TMA	set.about		TMA	play	PREP-3SG.OBL	

'And the kids play with it, [they] make like [it] is a cow, [they] play with it, [they] chase [it], [they] catch a lot [of them], [they] fill the bottle, [they] set about playing with it.'

Like interclausal linking, persistence appears to condition variable referential choice cross-linguistically, as it has been found to be predictive in variationist studies of subject expression in Santomean Portuguese (BOUCHARD, 2018), across varieties of Spanish (OTHEGUY & ZENTELLA, 2012; CARVALHO, OROZCO, & LAPIDUS SHIN, eds. 2015; PRADA PÉREZ, 2020; *inter alia.*), in comparisons of variable-rule systems across Romance varieties (DUARTE & SOARES DA SILVA, 2016), and across varieties of English (TORRES CACOULLOS & TRAVIS, 2014, 2019; WAGNER, 2016), while experimental and corpus studies provide comparable evidence from cross-linguistic studies of *anaphora resolution* (HOLLER & SUCKOW, eds., 2016, and references therein). A persistence effect was also found for null subjects across adjacent, prosodically linked clauses containing coreferential antecedent and anaphor in CVC (RODRÍGUEZ-RICCELLI, 2019, 2021).

In this section, I highlighted the compelling and ever mounting evidence for the cognitively-oriented discourse analysis conceptualization of referential choice espoused here as part of the Probabilistic Linguistics program. However, in the next section, we explore certain semantic properties of discourse referents that, counter to the predictions of the cognitively-oriented discourse analysis accounts, point to a promoting effect from referentially deficient discourse antecedents on null subject expression.

2 PREVIOUS RESRESEARCH ON SEMANTIC REFERENTIAL DEFICIENCY IN CVC AND BP SUBJECT EXPRESSION

Cardinaletti's and Starke's (1999) *Typology of Referential Deficiency* was a highly influential taxonomy that postulated a universal trinary classification by which strong pronominals were distinguished from deficient ones on the basis of a range of morphophonological, syntactic, semantic, and referential criteria. These properties were further used to distinguish two categories within the deficient class, weak and clitic pronominals. This taxonomy was invoked by Bapista (2002) and Pratas (2004) to classify subject pronouns in the Sotavento varieties of CVC (Table 1).

Table 1. CVC subject pronouns according to the Typology of Referential Deficiency *(based on BAPTISTA [2002: p. 245-250] and PRATAS [2004: p. 50-58]).*

Person/Number	Strong	Weak	Clitic
1SG	*ami*	*mi*	*=N= (=M=)*
2SG	*abo*	*bo*	*=bu=* (*=u*)
2SG (POLITE)	*anho* (M)/*anha* (F)	*nho* (M)/*nha* (F)	*=nhu=* (M)/*=nha=* (F)
3SG	*ael*	*el*	*=el= (=e=)*
1PL	*anos*	*nos*	*=nu=*
2PL	*anhos*	*nhos*	*=nhos=*
3PL	*aes*	*es*	*=es=*

Since the functional, distributional, and formal properties of each of these pronominal categories has already been explored extensively in Baptista (2002: p. 23-74, 213-267) and Pratas (2004: p. 39-61), we refer the reader to those studies. To summarize here, both "strong" and "weak" pronouns are [+tonic] and phrasal (XP, as opposed to syntactic heads), although weak pronouns are thought to be missing an additional complementizer phrase projection that is present for the strong forms (hence "deficient", CARDINALETTI & SRTARKE, 1999). Weak pronouns were claimed to differ from strong forms (and to share with "clitics") some deficient properties, such the inability to be coordinated or modified, the need to receive specification from a discourse antecedent, and the properties of semantic referential deficiency discussed in the introductory section of the present paper. Clitics share all these properties of deficiency with weak pronominals, but additionally, they are never stressed (and thus depend upon a phonological host), and they are not phrasal but rather syntactic heads (X^0). The properties of each class are summarized in Table 2.

Table 2. The properties of pronominal categories in the Typology of Referential Deficiency *(based on CARDINALETTI & STARKE [1999]).*

Strong	Weak	Clitic
- +animate - +specific - +definite - +tonic - +modification - +coordination - XP - Does not receive specification from antecedent (has "functional case features")	- +/-animate - +/- specific - +/-definite - +tonic - -modification - -coordination - XP (but 'deficient', i.e., lacking a functional projection= CP) - Receives specification from antecedent	- +/-animate - +/- specific - +/-definite - -tonic - -modification - -coordination - X° - Receives specification from antecedent

Other studies of pronominal and argument expression in CVC that work with a framework other than one from Generative Grammar adopt a binary classification of CVC subject pronouns distinguished solely on the basis of stress ([+/-tonic]) (VEIGA, 1996: p. 176-77, 332-35, 2002: p. 83; QUINT, 2000: p. 161-62, 2015: p. 44-45; LANG, 2012). With respect to the disyllabic forms containing initial *a-*, they have either been said to appear only in clause-initial position (QUINT, 2015: p. 44-45), or to be used only in emphatic and contrastive contexts (LANG, 2012). The paradigm under the binary classification is shown in in Table 3.

Table 3. The binary classification of CVC subject pronouns.

Person/Number	Tonic	Atonic
1SG	*(a)mi*	*=N= (=M=)*
2SG	*(a)bo*	*=bu= (=u)*
2SG (POLITE)	*(a)nho* (M)/*(a)nha* (F)	*=nhu=* (M)/*=nha=* (F)
3SG	*(a)el*	*=el= (=e=)*
1PL	*(a)nos*	*=nu=*
2PL	*(a)nhos*	*=nhos=*
3PL	*(a)es*	*=es=*

Importantly, singleton tonic subject pronouns (whether di- or monosyllabic) rarely occur as the sole morpheme indexing an active voice subject argument in the Sotavento CVC simple and embedded clause. Instead, they almost always appear in double subject constructions/conominations where they 'double' the subject clitic, as in examples (2) and (5c) above, and (6) and (7) below. Only seven of 8,548 isolated intonational units/clauses in the present corpus contained a singleton tonic subject pronoun with a verb other than *é* 'to be' (individual-level copula).[8]

7)

Pamodi	*ran*	*sta*	*di*	*un*	*ladu*			
because	frog	COP	PREP	DET	side			
i	***es***$_i$	***es***$_i$*=sta*			*di*	*kel*	*otu*	*ladu.*
CONJ	3SG	3SG=COP		PREP	that	other	side	

'Because the frog is on one side and they, they are on that other side' (P16, *Ebanir,* M, dialect zone: *Santiago centro* [childhood]; *Santiago sul* [adluthood], age: 19).

8)

*Dj-**el**,*	*kantu **e**=tropesa na*			*kel*	*pau,*	*dj=el*		*kai*
TMA-3SG	when 3SG=trip PREP			that	stick	TMA=3SG		fall
dentu di	*agu*							
PREP	water							

(P23, *Jesias*, M, dialect zone: *Maio*; age: 28)

[8] The major exception to the infrequency of singleton tonic subject pronouns generalization can be found with the verb *é* 'to be' (> Pt. *ser*). It is the only verb in CVC that categorically excludes subject clitics, thus allowing for variation only between a [+tonic] independent pronoun and a null subject. More recently in the history of the language, corresponding forms inflected for aspect, *foi* and *era* (invariably based on the Portuguese third-person, i.e., they do not inflect for person), were 'borrowed' into CVC from the Cabo-Verdean Portuguese superstrate (QUINT, 2012), and these appear to allow subject clitics at least at very low rates (RODRÍGUEZ-RICCELLI, 2021: p. 19fn11). However, based on the lexical kinship among *e, era,* and *foi*, all subject observations for these verbs were kept separate from the corpus used in the present study, which contains only subject observations from the remaining verbs in the language, all of which allow speaker's choice, given certain morphosyntactic and discursive conditions, from among the double subject construction, singleton clitic, or null subject options.

Thus, in the coding scheme for the present study, all double subject pronoun constructions were coded alike regardless of whether the conominating tonic independent pronoun was di- or monosyllabic. This was done for two reasons: (i) Given the inter-sub-disciplinary nature of the study, those distinguishing properties across the pronominal paradigm that are agreed upon across theoretical perspectives are favored over those which have not achieved consensus (i.e., everyone agrees on a stress based distinction but not on one based on assumptions of hierarchical clause structure which are specific to Generative Grammar); (ii) The sample size requirements implied by having two, rather than one category of double subject pronoun construction (one involving a disyllabic and another involving a monosyllabic [+tonic] form) represented in the response/outcome/dependent variable would have been too burdensome to conduct inferential statistical analysis on the present corpus (because this would weaken statistical power and require a much larger sample size).

Another important concern is with the position of the tonic independent pronoun relative to the subject clitic: the conominating tonic independent pronoun can be left dislocated, where it is separated from the clitic by some intervening material, as in (7), or they could be directly adjacent one another, as in (6) (in many cases the clitic can attach directly to the tonic pronoun itself, especially in the first-person where *N* attaches to *[A]mi* yielding *[A]mi=N*). But crucially, the variationist analysis brought to bear on the present corpus provides the methodological tools to consider the discourse function of reduced referring devices, whereas other methods may be needed to explore the hierarchical structure underlying the left-dislocated conominating independent pronoun as opposed to the one that is adjacent the 'subject slot'. For this reason, double subject pronoun constructions were categorized as alike regardless of the position of the tonic pronoun relative to the subject clitic.

Returning to the issue of semantic referential deficiency, recall that CVC singleton third-person clitics can resume referentially deficient antecedents, as was seen in (4) above, and as can be seen in (9), below, in which third-person singular *e* resumes the indefinite and nonhuman discourse antecedent *um moldura* 'a mold'.

9) *Ta faze-du* [*um, um* *moldura*]$_i$, *dipôs* ***e***$_i$=*ta*

TMA make-PASS DET, DET mold then 3sg.SBJ=TMA

fika *kunpridu* Ø$_i$ *ta* *seka,*

become long 3sg.SBJ TMA dry

'One makes a mold, then it becomes elongated [and] [it] dries.'

Another important property of deficient reduced referring devices in CVC can also be observed in (4) and (9), and in (6), above: in those examples, an indefinite/nonspecific (in the case of [6]) or an inanimate (in the cases of [4] and [9]) lexical nominal referent is resumed immediately by a null subject (in the case of [6]), or first a clitic in the independent clause immediately following the clause containing the lexical nominal, and then a null subject in the next independent clause after that (in the case of [9]). This is consistent with the longstanding classification of *pro*—the phonologically silent morpheme taken to underlie null subjects in Generative Grammar—and clitics—as deficient pronouns (CARDINALETTI & STARKE, 1999, p. 68, 89-91; *inter alia.*).

Recall that, on the other hand, in the case of double subject pronoun constructions like (2), (5), (7), and (8), above, Baptista (2002: 240-42) had asserted they should never resume nonhuman antecedents. In a quantitative analysis of double subjects involving both *DP + clitic* and *tonic pronoun + clitic* amalgamations across Lusophone varieties, Silva (2013), Silva, Carvalho, and Ziober (2016, 2017), and Silva and Ziober (2017), found quantitative evidence that they tend to encode human, definite, and specific referents, but rarely encode semantically referentially deficient ones. The variationist analyses of CVC subject expression in Rodríguez-Riccelli (2019) provided yet more evidence in support of this notion: the jitter plots in Figures 2 and 3 display every subject observation of the types in (1)-(3) (including first- and third-person referents in the corpus) as a plotted point. In Figure 2 it can be observed that double subject pronoun constructions never resumed an antecedent that bore nonhuman reference, and did so only twice when the antecedent bore collective reference. In Figure 3, it can be seen that a double subject pronoun construction never resumed an indefinite and nonspecific referent, and only resumed an indefinite and specific referent once.

*Figure 2. Jitter plot of subject pronoun expression (*spe*) by animacy of the referent (double subject pronoun construction =* dbl*; [in]animate =* [IN]ANIM*) (RODRÍGUEZ-RICCELLI, 2019: p. 311)*

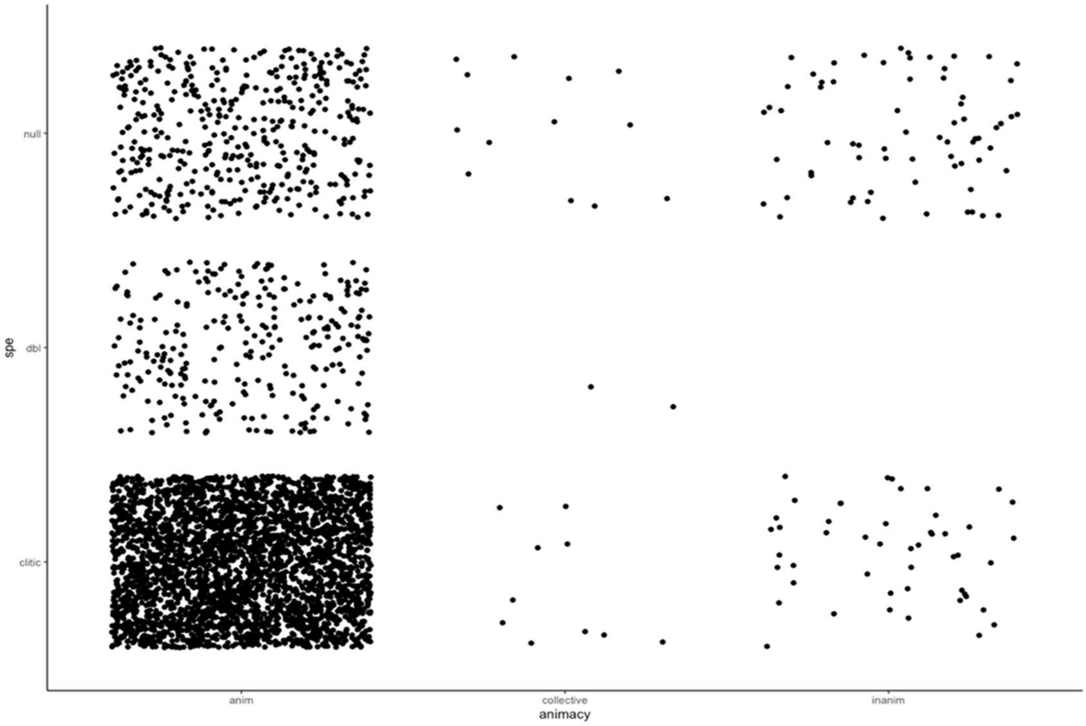
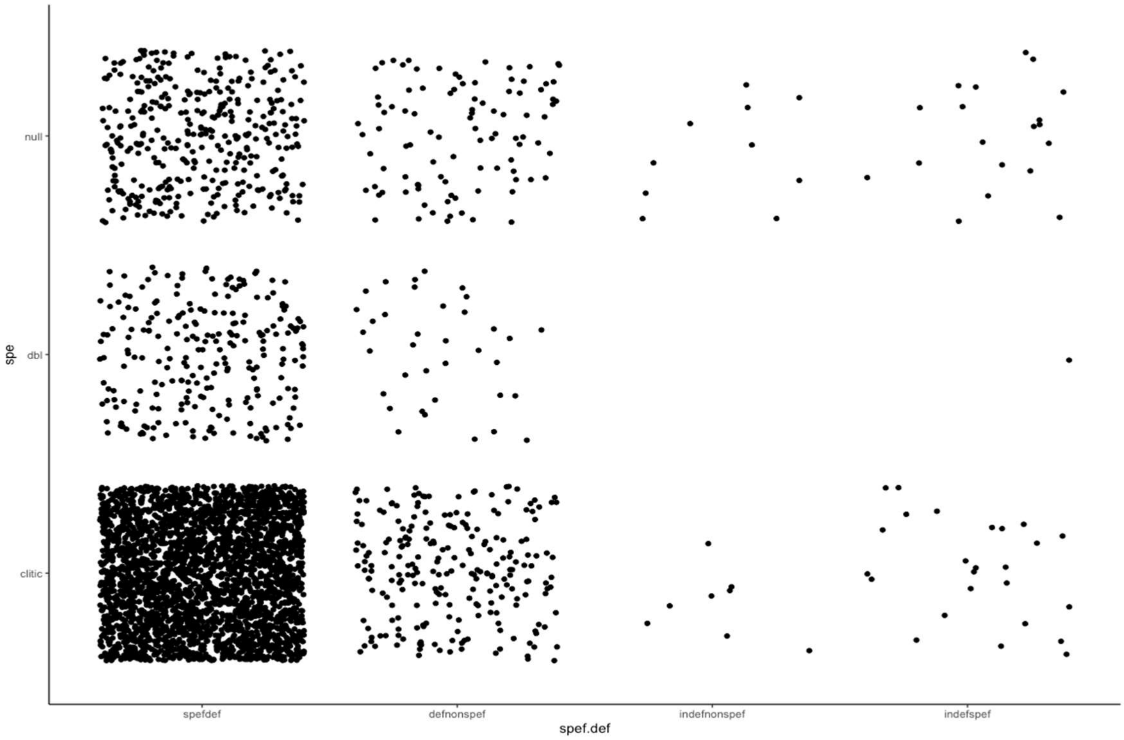


*Figure 3. Jitter plot of subject pronoun expression (*spe*) by specificity-definiteness of the referent (double subject pronoun construction =* dbl*; [non]specific =* [non]spef*; [in]definite =* [in]def*) (RODRÍGUEZ-RICCELLI, 2019: p. 319)*



Further, in the multinomial regression analyses conducted in Rodríguez-Riccelli (2019, 2021), nonspecific reference was found to have a promoting effect on the selection of null subjects. These findings mirror quantitative evidence from BP, which also suggests a role for semantic referential deficiency in probabilistic models of argument expression. For instance, Othero *et al*. (2018) applied quantitative methods to measure rates of null/overt object expression in a corpus of tweets and other text approximating naturalistic BP speech. Animacy (though not specificity) was found to be an important conditioning factor in BP object expression, with nonhuman antecedents promoting resumption by null objects.

Another recent example can be found in Duarte and Soares da Silva (2017). In that study, the authors updated some of the quantitative methods that had been applied to the corpus data explored in Duarte (1995); they found a strong promoting effect from nonhuman and/or nonspecific referents on null subject expression in the variable-rule system for BP (thereby differentiating BP from all the other Romance varieties that they analyzed). They recast Montalbetti's (1984) famous "avoid pronoun principle" in terms of strong and deficient pronominals for BP, this time as the "avoid referentially deficient pronoun" principle, which could be understood to apply probabilistically. After finding similar results for CVC subject expression, Rodríguez-Riccelli (2019, 2021) postulated that the "avoid referentially deficient pronoun" principle was active in that language too.

In the next section, I turn to a description of the methods, from data collection, to transcription, to coding, and the statistical analysis.

3 METHODS

All recordings were collected between 2015 - 2017 from inhabitants of the islands of Santiago and Maio who were born and raised in the Republic of Cabo Verde. The participants were recruited through word-of-mouth snowball sampling led by local community member informants. The goal was to collect speech samples from speakers representing as broad a range of ages and socioeconomic backgrounds as was possible.

In the summer of 2015, recordings were collected in Palmarejo and adjacent neighborhoods of the capital city, Praia, including on the campus of the Universidade de Cabo Verde. Respondents were also recruited in the municipality of Vila do Maio, the only substantial urbanization on the eponymous island. Of the recordings obtained in the first summer, 11 interviews were retained from Praia-based speakers and five from Maio-based speakers.

The next summer, I resided in the Achada Igreja urbanization of the *vila* 'town' of Picos, in the *concelho* 'municipality' of São Salvador do Mundo in the center of Santiago. Participants were also recruited from the adjacent *concelhos* of São Lourenço dos Órgãos and Santa Catarina. For the recordings collected during this second summer, a colleague and informant from Achada Igreja facilitated participant recruitment and assisted in asking questions of the participants during the sociolinguistic interview, and in issuing instructions prior to the picture description narrative task. We retained ten recordings for analysis from this phase of the fieldwork.

Recordings in the third summer were collected from speakers from the town of Tarrafal in the eponymous *concelho* in *Santiago norte* 'Northern Santiago'. Like in the previous summer, a local informant who was born and raised in Tarrafal helped to identify and recruit participants, to ask questions during the interview, and to explain instructions. Eight of the recordings that we collected in 2017 were retained for the present analysis.

Elicitation methods involved a guided sociolinguistic interview and a picture description narrative task. The interviews were intended to be conversational in nature and to elicit the vernacular. Each interview relied on the same battery of questions, but if the interviewer(s) found a topic to be of interest to a respondent, they would pursue additional questions on that topic or allow the respondent to continue as they pleased. Demographic information was elicited as part of the interview. 25 of 33 respondents completed a picture description narrative task in which they were asked to narrate the sequence of events depicted in the unannotated illustrations in *A Boy, a Dog, and a Frog* (MAYER, 1967).

The speakers' ages ranged from 18 to 57, with an average age of 29.3. There were 12 male respondents with an average age of 36.2, and 21 female respondents with an average age of 25.4 (see, RODRÍGUEZ-RICCELLI [2019, 2021] for a full report of the demographics of the respondents). The data were recorded with a Zoom H2N Handy recorder. The first 15 minutes of the interview portion of each recording were first transcribed by the present author; for recordings that also contained a Frog Story narrative (n = 25/33), it was transcribed in its entirety. The transcriptions were then sent to three assistants for review, editing, and correction. The transcription assistants were verbally informed of the topic under investigation and were issued an instructional style sheet which established the orthographic norms and how to represent common morphophonological elements (such as clitics, Tense-Mood-Aspect markers, and zero subjects, etc.).

Once the transcription assistants returned the edited and corrected transcriptions, the author reviewed and finalized them to reflect a *prosodic transcription format*, following a simplified method based the one used in Torres Cacoullos and Travis (2019). This method requires the transcriber to represent the prosodic contours of the discourse using punctuation, line spacing, and by grouping intonationally interconnected utterances into larger 'chunks' (~paragraphs). The basic segment is an intonational unit (IU), which represents a segment of speech "uttered under a single, coherent intonation contour" (DU BOIS *et al.*, 1993: p. 58-60; CHAFE, 1994: p. 58-60). Each IU occupies a line of text in the document. The end boundary of an IU is defined by a "terminal pitch contour": a terminal pitch contour is characterized by low or sharply falling intonation, it is represented orthographically with a period <.>, and it is associated with non-continuing intonation or prosodically unlinked IUs; continuing intonation is defined by a rising or steady intonational contour at the end of an IU, it is represented orthographically with a comma <,>, and it is associated with prosodic linking between IUs (DU BOIS *et al*. 1993, p. 53; CHAFE 1994, p. 60-61).

IUs can also be syntactically linked by coordinating conjunctions, subordinators and relativizers, or discourse markers.[9] Syntactic linkers are

[9] CVC syntactic linkers included coordinating conjunctions *i* 'and', *ô(u)* 'or', and *ma(s)* 'but', and complementizers, relativizers, and discourse markers like *ki* 'that', *ma* 'that', *(ã)ntôn* 'so, then', *(lógu) dipôs (ki)* 'then, after', *(a)gó* 'now/well/then', *tipu (ki)* 'like, sort of like', *ó(ra(s)) ki* / *kuandu* / *kantu ki* 'when', *(na) undi (ki)* 'where', *pa* 'to, for to', *nkuantu* 'during', *gósi (ki)* 'now (that)', *te/ti (ki)* 'until', *kazu (ki)* 'in case', *sima (ki)* 'like, like if', *(ẽ)mbora (ki)* 'even though, although, despite', *(di) modi ki* 'like if, as if, in a way that', and *pamo(di)* 'because'. For the purposes of syntactic linking, I made no distinction between complementizers and relativizers, but this distinction should be implemented in all coding procedures in future analyses.

usually placed at IU boundaries when the content of the IU “introduces separate new information” (CHAFE 1994, p. 188). Most IUs contain a single independent clause. However, when there was subordination or embedding in which the matrix and embedded clauses both fall under a single overarching intonational contour (i.e., there is no audible pause between the main and embedded clauses), these IUs were coded as simultaneously syntactically and prosodically linked (where relevant: in the presence of coreference across the IUs). Given space limitations, only one example of a referring device (double subject, null subject, subject clitic) per interclausal linking condition ([+/-prosodic linking], [+/-syntactic linking]) is shown here (10)-(13) (RODRÍGUEZ-RICCELLI, 2021: p. 24-25). To see examples of all three reduced referring devices in each condition, please consult Rodríguez-Riccelli (2019: p. 274-277).

10) Prosodic linking (IU-final rising or steady intonation)

E$_{i}$*=ta* *pila* *kana*,

3SG=TMA crush cane

Ø$_{i}$ *ta* *fazi* *si* *groginhu*

3SG TMA make 3SG.POSS cane.spirit

‘He presses the sugar cane, [he] makes his aguardente.’ (P24, *Rogério*, M, age: 55, dialect zone: *Santiago Centro*)

11) Syntactic linking (IU-initial or -final conjunction, complementizer, relativizer, or discourse marker)

Rapazinhu$_{i}$ *buâ* *n-el*.

boy jump PREP-3SG

I *li* ***e***$_{i}$*=sa* *ba ku* *redi* *p=e*$_{i}$

CONJ here 3SG=TMA go PREP net COMP=3SG

panha=l

catch=3SG

'The boy jumped on it. And here he is going with the net to catch it.'
(P45, Danira, F, age: 21, dialect zone: *Santiago centro*)

12) Both prosodic and syntactic linking

[*Kel sapu*]$_i$ *ta fuji mas um bes,*
[that frog]$_i$ TMA escape more DET time
i **Ø**$_i$ *ba fika sukundidu riba d-um pédra*
CONJ 3SG$_i$ go stay hidden on.top.of-DET rock
'That frog escapes one more time, and [he] goes to hide on top of a rock.' (P6, *Kátia*, F, age: 28, dialect zone: *Santiago sul*)

13) No linking

Ulisis$_i$*, k-eli*$_i$ *é prizidenti di kambra li di Praia***.**
U. REL.FOC-3SG COP municipal executive here PREP Praia
El$_i$*=**e***$_i$ *kumesa fazi um bon trabadju, kalseta rua,*
3SG=3SG begin do DET good job cobble street
fazi prasas, fitines park, así.
make square fitness park like that
'Ulises, who is the municipal executive, here from Praia. He started to do a good job, cobbling the streets, making plazas, fitness parks, things like that.' (P11, *Jandira*, F, age: 27, dialect zone: *Santiago sul*)

Space restriction here prevent a detailed elaboration of the *variable context* (i.e., *the envelope of variation* or the range of variants available for the speaker to choose from, in this case, the morphophonemes that CVC allows in active voice nominative reference) for CVC subject expression, but the reader is directed to the extensive expositions in Rodríguez-Riccelli (2019, 2021). But, differently from those previous studies, and in order to explore the effects of semantic referential deficiency in CVC subject expression in the present study, third-person referents were separated from first-person ones (since only third-person referents can be nonhuman and indefinite/nonspecific). As such, for any given observation of a nominative referential device to be admitted for analysis, the discourse referent to which it referred must have been introduced by lexical nominal DP/NP antecedent earlier in the discourse. Further, since, as we saw from Figures 1 and 2 that double subject pronoun constructions did not resume nonhuman referents, only contexts where a plausible variable choice between the subject clitic (1) and the null subject (3) expressions were measured in the quantitative analysis.

Before fitting the regression model for the primary analysis, a model building and selection procedure was carried out to select the most important of all the variables that had been coded for in the corpus (for a complete reporting of all of the variables that were coded for, see RODRÍGUEZ-RICCELLI, 2019). First, the descriptive statistical relationships among the variables were examined in R (R CORE TEAM, 2021) using cross tabulations and by inspecting visual distributions of the data using the 'ggplot2' package (WICKHAM, 2016). Then, a variable importance plot of a random forests analysis (STROBL *et al*., 2007, 2008) that included all the coded predictors was inspected using 'party' (HOTHORN, HORNIK, & ZEILEIS, 2006). This was compared with the output of forward-and-backward stepwise Likelihood Ratio tests, which select an optimal model from the full set of predictors by using ANOVA to compare changes to the nested models' Akaike Information Criterion score (AKAIKE, 1998[1973]) when predictors are iteratively added or removed from the nested models. Next, the outputs of the two procedures were compared and the predictors that ranked highest in the variable importance plot, and that were also retained following the stepwise procedure, came to form the parameter structure for a base model.

The predictor variables that were ultimately retained to fit the mixed-effects binomial regression with response outcomes *subject clitic* (1), and *null subject* (3), in order of descending variable importance, were: *morphophonological form of the antecedent*, *interclausal linking*, and *animacy of the referent*, with *participant* as a random effect. *Morphophonological form of the antecedent* is self explanatorily named; its outcomes were *double subject* (*dbl*, both those with a *DP* + *clitic* or *tonic pronoun* + *clitic* amalgamations), *DP* (when directly adjacent to the verb and without a clitic), *DP* + *intervening material* + *Ø* (with a left dislocated DP and the clitic slot empty), *incorporated 'ta'* serial verb-like *construction* (*incpta*), *possessive pronoun* (*posspro*), singleton *tonic pronoun,* and *zero* (null) anaphor; the application value (against which the prior levels were compared) was a singleton *clitic. Interclausal linking* referred to the use of prosodic and syntactic linkers (described earlier in this section) and applied when there was coreference across adjacent IUs. The outcomes were *both* syntactic and prosodic linking, only *prosodic linking*, or only *syntactic linking*; the application value was *no linking*. *Animacy of the referent* was self explanatorily named and bore the levels *collective*, and *inanimate* (*inanim*), and had the application value *animate* (see RODRÍGUEZ-RICCELLI, 2019, for linguistic examples exemplifying the various levels of the predictor variable).

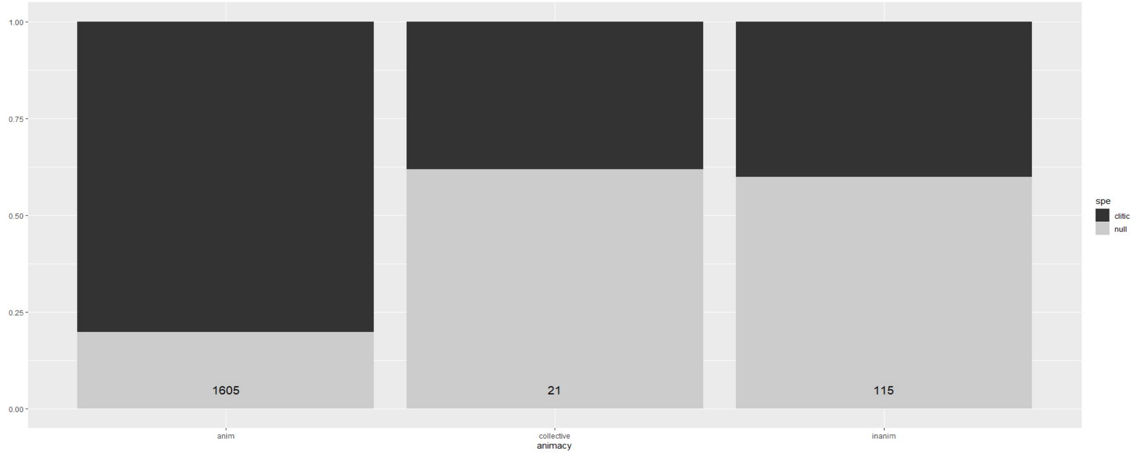
To more closely observe interactions among the variables, and to adopt an alternative non-parametric perspective, a conditional inference tree that included the three aforementioned predictor variables was grown using 'party'. Post-hoc, the regression model was tested for singularity using the 'issingular' function and for multicollinearity using the 'vif' function in the 'usdm' package (NAIMI *et al*. 2014),[10] resulting in non-singular models with low collinearity.

[10] We followed the standardized "rule of thumb" recommendations for VIF (BELSLEY, 1991, p. 56; KASSAMBRA, 2018). (GVIF$^{(1/2*df)}$) < 5 indicated "weak dependency" (low collinearity). If one of the parameters were to have exceeded this value, it would have been removed from the model. See Rodríguez-Riccelli (2019) for a full reporting of the post-hoc statistics.

4 RESULTS

Following the modifications to previous iterations of the envelope of variation for subject expression in CVC described above, 1,741 observations were analyzed. Of these, 1,331 (76.9%) were singleton subject clitics and 400 were null subjects (23.1%) (n=90 double subject pronoun constructions were removed). Figure 4 shows the proportion of *subject clitics* and *null subjects* for each level of *animacy of the referent*, where the y-axis represents 100% of the observations in each category of the predictor. The number at the base of each bar is the total number of observations in that category and the shading represents the proportion associated with each referring device. It can be seen that 61.2% (n = 13) of all *collective* referents and 60% (n = 69) of all *inanimate* referents were resumed by *null subjects*.

*Figure 4. Referential choice in third person subject pronoun expression (*spe*) by* animacy of the referent*.*

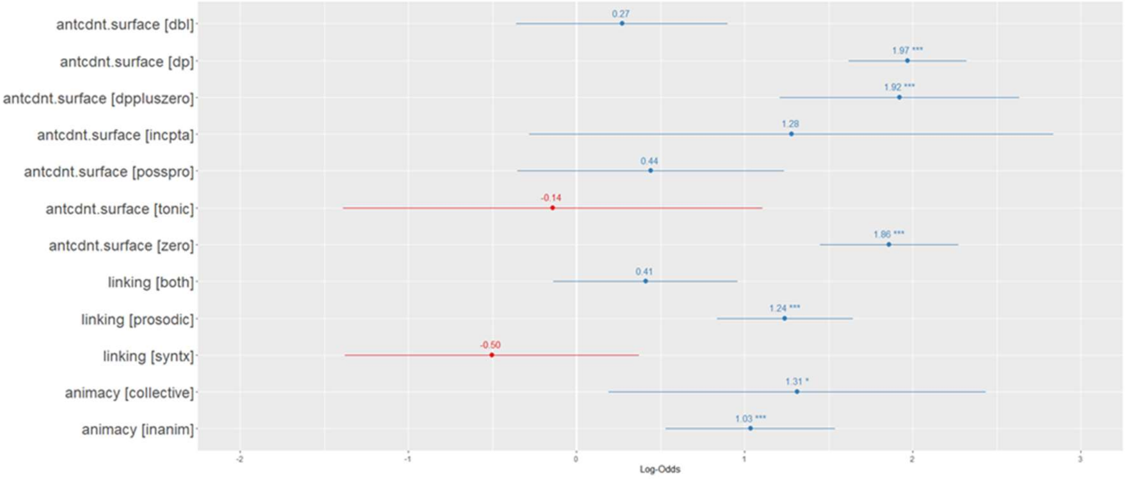


To follow, in Figure 5,[11] the output for the fixed-effects from the binomial mixed-effects regression is reported in a visual plot using ‘sjPlot’ (LÜDECKE, 2019). In this plot, the application value *subject clitic* for the

[11] AIC = 1413.9; Deviance = 1361.9; Residual d.f. = 1715; Random effects: Variance = 0.147; Std. Dev. = 0.383.

response/dependent/outcome variable is held as a baseline against which the log-odds of realizing a *null subject,* given each level of the predictor variables (except for the application value for each predictor, which is also held constant as a point of comparison), are plotted correspondingly along the y-axis. The numerical value of the log-odds is listed above each plotted point; log-odds greater than zero are plotted in blue and to the right half of the x-axis; these indicate a favoring effect from that level of the predictor variable on the realization of a *null subject*; log-odds less than zero and plotted in red to the left of the x-axis indicate the converse, a disfavoring effect from that level of the predictor on the realization of a *null subject.* Statistical significance is indicated with asterisks (* = p < 0.05, ** = p < 0.01, *** = p < 0.001).

Figure 5. Binomial mixed-effects regression for third-person null subject *expression in Santiago and Maio CVC.*



The results in Figure 5 show a strong promoting effect from lexical nominal *DP* (log-odds = 1.97, p < 0.001) and *DP + intervening material + Ø* (log-odds = 1.92, p < 0.001) antecedents on *null subject* expression. Upon consulting the results for the conditional inference tree analysis in Figure 6 below, it should become apparent that two effects obtained: one can be attributed to the high degree of discourse coherence associated with referents that were recently activated by lexical nominals; the other can be attributed to the promoting effect on null subject expression from semantically referentially deficient lexical nominals, namely those bearing inanimate and collective reference.

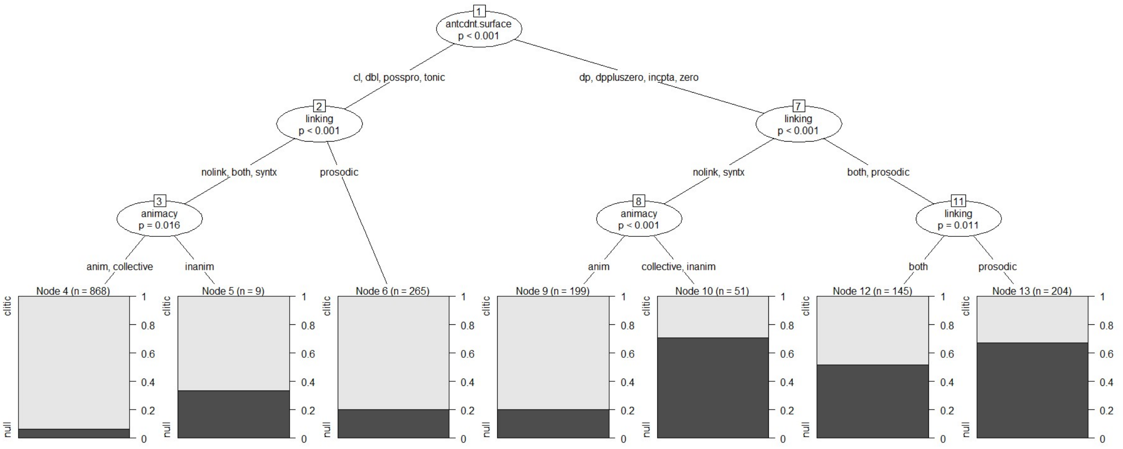
There was also a strong promoting effect from *zero* (null) anaphor antecedents on the realization of subsequent coreferential *null subjects* (log-odds = 1.86, p < 0.001). This reinforces previous observations of such a persistence effect. *Prosodic interclausal linking* also exerted a strong promoting effect on *null subject* expression (log-odds = 1.24, p < 0.001), usually occurring in anaphoric chains like (6) above. It seems plausible that *interclausal linking* and *persistence of morphophonological form* work in tandem to enhance referential accessibility.

The aforementioned effect for *animacy* can be seen in the favoring effect from *collective* (log-odds = 1.31, p < 0.05) and *inanimate* (log-odds = 1.03, p < 0.001) referents on *null subject* expression. That this effect is somewhat independent from the linking and persistence effects is supported by the results of the conditional inference tree analysis in Figure 6 below.

The conditional inference tree in Figure 6 shows the relative importance of the predictor variables in determining *subject clitic* v. *null subject* outcomes. The branches emanating from the numbered bubbles with the name of the predictor variables inside them represent statistically significant splitting criteria. The relevant levels of the predictor variable are listed along each branch, and it can be seen that antecedents with the morphophonological form *dp, dp + intervening material* + Ø, or *zero* (null anaphor antecedent), patterned apart from all others. In that condition, outcomes were significantly different when there was only *prosodic* or *both* prosodic and syntactic *linking* (as opposed to when there was *no linking* or only *syntactic linking*). The numbers above each bubble are the node numbers and the total number of observations associated with each terminal node is shown above the bars along the bottom of the plot. The p-value for each bubble is calculated from a significance test on independence between the covariates and the predictor outcomes (HOTHRON & EVERITT, 2006); branches split when *p* is less than 0.05. The light grey shading at the top of each bar represents singleton *subject clitic* responses, while the dark grey shading at the bottom of each bar represents *null subject* realizations.

We draw the reader's attention to the right-branching half of the tree coming out of Node 1. Observe that when the antecedent was a *dp, dp + intervening material + Ø,* or *zero*, and when the target anaphor and antecedent were in *unliked* or only *syntactically linked* IUs (left-branching out of Node 7), then the effect for *animacy* obtained (Node 8; p < 0.001), with *collective* and *inanimate* referents in this context accounting for the largest overall proportion of *null subjects*. The right branch emerging from Node 7 reveals the second and third largest contributors of *null subjects* to the sample (in Nodes 12 and 13). Here is where the effects of interclausal linking and persistence obtained (often in anaphoric chains like [6] above), thus increasing the likelihood of a null subject being selected, even when (in some cases) it was not promoted by the "avoid referentially deficient pronoun" effect (i.e., even when the referent was [+human] and/or [+definite/specific]).

Figure 9. Conditional inference tree for subject pronoun expression *by* morphophonological form of the antecedent, interclausal linking, *and* animacy.



5 DISCUSSION

In the present study, we elaborated a series of quantitative analyses of CVC subject expression in order to examine the role of semantic referential deficiency as a variable constraint on speaker's choice of nominative reduced referring device. This analysis culled a newly defined envelope of variation from a corpus of naturalistic discourse collected from speakers from Santiago and Maio, modeling third-person/external reference subject expression on its own (apart from deictic/Speech Act Participant referring devices), although notions of semantic referential deficiency were rooted in previous studies on CVC and BP. Ultimately, only the predictor variable encoding *animacy of the referent* was retained following the elimination of the less predictive/important variables related to semantic referential deficiency that had been coded for in the corpus (i.e., specificity/definiteness).

The results from a series of statistical analyses revealed two main effects: one involved both *persistence of morphophonological form* and *interclausal linking*, and especially *prosodic linking* between adjacent IUs containing coreferential antecedent and anaphor. Both these forces can be seen as tied to issues of referential activation, which in turn is related to working memory (KIBRIK, 1997). The other significant effect involves *semantic referential deficiency*, by which inanimate and indefinite/nonspecific referents disfavor the expression of double subject pronoun constructions, and by which nonhuman and collective referents exert a promoting effect on null subject expression, as in (13)-(14).

13)	*Ten*	[*otus*	*izemplus*	*di*	*país*	*ki*	*ka*	*é*
	PRES	other	examples	PREP	country	REL	NEG	COP
	riku]$_i$	*ki*	*konsigi implementa kela,*					
	rich	COMP	able.to implement DEM					

i	*purizemplu*	Ø$_i$	*ta*	*txoma*	*mas*	*turistas*
CONJ	for.example	3.PL$_i$	TMA	call	more	tourists

pa	*ses*$_i$	*lugar,*
PREP	3.PL.POSS	place

‘There are other examples of countries that are not rich that were able to implement that, and for example [they] attract more tourists to their places.’ (P4, *Renata*, F, dialect zone: *Santiago sul,* age: 27)

14) Q:

Bu	*ten*	*[um pratu, así*	*terra-terra,*	*ki*	*mas*
2.SG	have	DET plate like.that	earth-earth	COMP	most

bu	*gosta*]$_i$*?*
2.SG	like

A:

mmm	*não purakazu*	*N=ta*	*kumé*	*di*	*tudu*
DM	NEG so.happens	1.SG=TMA	eat	PREP	all

um poku,	*ma si*	*for*	Ø$_i$	*ta*
DET little	DM COMP	be.SBJNCT	3.SG	TMA

ser-ba	*fejoada*
be-TMA	feijoada

‘Q: Do you have a favorite plate, like a national dish, that you like the most? A: hmm it so happens that I eat a little of everything, but if I had to choose [it] would be feijodada.’ (P4, *Renata*, F, dialect zone: *Santiago sul,* age: 27)

This latter effect (from semantically referentially deficient antecedents) is somewhat counterintuitive for cognitively-oriented discourse analysis approaches, which usually take inanimacy, nonspecificity, and indefiniteness, as a hindrance to referential coherence, and therefore as a force that disfavors referential choice of zero/null arguments. Yet, as we have discussed above, similar effects for semantic referential deficiency to those found here for CVC can also be found in BP argument expression. Thus, it appears that where the same person marker or pronominal form can be deployed to resume both fully human/animate and nonhuman/inanimate discourse referents alike, semantic effects can intervene to disrupt or act independently from the forces associated with referential accessibility or coherence.

Before turning to the conclusion section in which we propose modifications to the methods used here, and new directions for future research, we find it constructive to acknowledge the value that has come from defying the habit in the field of linguistics to assume that research in the rationalist tradition is somehow incompatible with empiricist approaches. Such inter-sub-disciplinary blending has long been practiced in research on BP morphosyntax and discourse (TARALLO, 2015[1987]; TARALLO & KATO, 2007[1989]; DUARTE, 1995, 2000; DUARTE & SOARES DA SILVA, 2016; OTHERO *et al.*, 2018). We advocate for a *Kantian Compromise* (e.g., JOHNSON, 1974; PLOTKIN, 2008; *inter alia.*), first, by providing an additional example of how notions from rationalist paradigms can serve as hypotheses to be tested empirically for their probabilistic conditioning effects. Secondly, we urge investigators to explore and engage with research beyond their disciplinary-paradigmatic and methodological 'comfort zones', with an eye towards advancing a more holistic and integrated model of natural language use, processing, and cognitive representation.

6 CONCLUSIONS AND FUTURE DIRECTIONS

A number of improvements could be made to the coding procedure and analysis adopted in the present study. Firstly, animacy, definiteness, and specificity might be subsumed in a single predictor variable encoding something like the *animacy hierarchies* of Foley and Van Valin (1984) or those reviewed in

Vihman and Nelson (2019). Similarly, notions of interclausal linking might be subsumed with another predictor variable (which was not retained in the present analysis, but which has proved predictive in other iterations of studies drawn from the present corpus), *antecedent accessibility pattern*, or the clausal configuration between antecedent and anaphor. Other adjustments might include fusing the separate DP categories into a single level for lexical nominals within the predictor *morphophonological form of the antecedent*, and perhaps eliminating distinctive predictor levels for noncanonical argument slots like those in the *incorporated ta* serial verb-like construction.

This study leaves many future directions for analysis. Crucially, a full model of referential choice of discourse anaphora would include both object and subject reference (since the three-way variation between conomination/singleton clitic/null subject is also active for object reference in CVC, see [5a-d] above), and perhaps would redefine these in terms of *thematic relations*, such as for the agent, patientive, and benefactive roles.

With respect to the effects of semantic referential deficiency on variable referential choice, the present study contributes to the ever-mounting evidence for its active status in discourse and morphosyntactic organization across Lusophone varieties, though confirmation of this generalization awaits additional empirical testing across more varieties and types of discourse. Interestingly, similar effects have also been found for argument expression in so-called *topic-prominent* languages like Mandarin (LI & THOMPSON 1976; CHEN 1986; PU 1997; HUANG 2000). Future research should identify if the relationship between semantic referential deficiency and variable choice of discourse-anaphora is due to a typological orientation towards topic-prominent discourse organization (as has long been claimed for BP; see, for example: PONTES, 1987; NEGRÃO & VIOTTI, 2000; CAVALCANTE DA CUNHA, 2010; NUNES, 2016; OLIVEIRA DA SILVA & ALVES FONSECA, 2018; *inter alia.*), or if the relationship can simply be attributed to the availability of a variable choice between an overt and a zero/null anaphor, both of which can resume semantically referentially deficient antecedents. Disentangling such questions might be achieved by establishing comparisons among variable-rule systems for referential choice cross-linguistically, following the *Variationist Typology* method proposed in Torres Cacoullos and Travis (2019).

REFERENCES

AKAIKE, Hirotogu. Information Theory and an extension of the maximum likelihood principle. In: PARZEN, Emanuel; TANABE, Kunio; KITAGAWA, Genshiro. *Selected Papers of Hirotugu Akaike*, New York: Springer, 1998. p. 199-213.

ARIEL, Mira. *Accessing noun-phrase antecedents*. Abington, Oxon, UK: Routledge, 1990.

ARIEL, Mira. Accessibility theory: An overview. In SANDERS, Ted J.M.; SCHILPEROORD, Joost; SPOOREN, Wilbert. *Text representation:* Linguistic and psycholinguistic aspects. Amsterdam/Philadelphia: John Benjamins, 2001. p. 29–87.

BAPTISTA, Marlyse. *The Syntax of Cape Verdean Creole:* The Sotavento Varieties. Amsterdam/Philadelphia: John Benjamins. 2002.

BELSLEY, David A. *Conditioning diagnostics: Collinearity and weak data in regression*. New York: Wiley. 1991.

BOUCHARD, Marie-Eve. Subject pronoun expression in Santomean Portuguese. *Journal of Portuguese Linguistics,* Caderno 17, 1, p. 5, 2018.

CARDINALETTI, Anna; STARKE, Michal. The typology of structural deficiency: A case study of the three classes of pronouns. In: RIEMSDIJK, Henk van.*Eurotyp*. Berlin/New York: De Gruyter Mouton, 1999 p. 145 - 234.

CAVALCANTE DA CUNHA, Antônio Sérgio. Estrutura tópico-comentário, a tradição gramatical e o ensino de redação. *Soletras*, Caderno 20, São Gonçalo, p. 53-63, jul./dez. 2010.

CHAFE, Wallace L. Cognitive constraints on information flow. In: TOMLIN, Russel S. *Coherence and Grounding in Discourse:* Outcome of a Symposium, Eugene, Oregon, June 1984, Amsterdam/Philadelphia: John Benjamins, 1987. p. 21-51.

CHAFE, Wallace L. Prosodic and functional units of language. In: EDWARDS, Jane A.; LAMPERT, Martin D. *Talking data: Transcription and coding in discourse research*, Hillsdale, NJ: Lawrence Erlbaum Associates, 1993.

CHAFE, Wallace L. *Discourse, consciousness, and time:* The flow and displacement of conscious experience in speaking and writing. Chicago: University of Chicago Press, 1994.

CHEN, Ping. *Referent introducing and tracking in Chinese narratives*. 1986. Tese. University of California, Los Angeles.

CLAES, Jeroen. Probabilistic Grammar: The view from cognitive sociolinguistics. *Glossa:* a journal of general linguistics, Caderno 2, 1, 2017, p. 62.

CORNISH, Francis. *Anaphora, discourse, and understanding:* Evidence from English and French. Oxford, UK: Clarendon Press. 1999.

CYRINO, Sonia M. L.; DUARTE, Eugênia M. L.; KATO, Mary A. Visible subject and invisible clitics in Brazilian Portuguese. In: KATO, Mary Aizawa; NEGRÃO, Esmeralda Vailati. *Brazilian Portuguese and the Null Subject Parameter*. Frankfrut a. M./Madrid: Vervuert Verlagsgesellschaft, 2000, p. 55-74.

CYRINO, Sonia M. L. Construções com *se* e a promoção de argumento no português brasileiro: uma investigação diacrônica. *Revista de ABRALIN,* Caderno 6, 2, Segripe, 2007.

CYRINO, Sonia M.L.; MATOS, Gabriela. Anáfora do Complemento Nulo: anáfora profunda ou de superfície? Evidência do Português Brasileiro e Europeu. *Letras De Hoje*. Caderno 41, 1, Porto Alegre, 2006.

DUARTE, Maria Eugênia L. *A Perda do Princípio "Evite pronome" no Português Brasileiro*. 1995. Tese. Universidade Estadual de Campinas.

DUARTE, Maria Eugênia L. 2000. The loss of the 'Avoid Pronoun' principle in Brazilian Portuguese. In Mary Aizawa Kato & Esmeralda Vailati Negrão (eds.), *Brazilian Portuguese and the Null Subject Parameter* (Ediciones de Iberoamericana 4). Frankfurt am Main, Madrid: Iberoamericana Vervuert.

DUARTE, Maria Eugênia L.; SOARES DA SILVA, Humberto. Microparametric variation in Spanish and Portuguese. In: KATO, Mary A.; ORDÓÑEZ Francisco. *The Morphosyntax of Portuguese and Spanish in Latin America*. Oxford, UK: Oxford University Press, 2016. p. 1-26.

DUARTE, Maria Eugênia L. Analyzing a parametric change in Brazilian Portuguese: a sociolinguistic investigation, In: BARBOSA, Pilar P.; PAVIA, Maria da Conceição de; RODRIGUES, Celeste Rodrigues. *Studies on Variation in Portuguese.* Amsterdam/ Philadelphia: John Benjamins, p. 234–253, 2017.

DU BOIS, John W.; SCHUETZE-COBURN, Stephan; CUMMING, Susanna; PAOLINO, Danae. Outline of discourse transcription. In: EDWARDS. Jane A. ; LAMPERTS, Martin D. *Talking data:* Transcription and coding in discourse research, Hillsdale, NJ: Lawrence Erlbaum. 1993, p. 45-89.

FOLEY, William A.; VAN VALIN, Robert D. Jr. *Functional Syntax and Universal Grammar*. Cambridge, UK: Cambridge University Press. 1984.

GIVÓN, Talmy. Topic, pronoun and grammatical agreement. In: LI, Charles N. *Subject and Topic*. New York: Academic Press, 1976. p. 151-88.

GIVÓN, Talmy. Topic continuity in discourse: The functional domain of switch reference. In: HAIMAN, John Haiman; MUNRO, Pamela Munro. *Switch reference and Universal Grammar:* Proceedings of a symposium on switch reference and universal grammar, Winnipeg, May 1981. Amsterdam/Philadelphia: John Benjamins. 1983.

GIVÓN, Talmy. The grammar of referential coherence as mental processing instructions. In: DITTMAR, Norbert Dittmar. *Topic:* special issue of Linguistics, Caderno 30, 1992. p. 5–56.

GIVÓN, Talmy. 2017. *The story of zero*. Amsterdam/Philadelphia: John Benjamins.

HOLLER, Anke; SUCKOW, Katja. *Empirical Perspectives on Anaphora Resolution*. Berlin/Boston: Walter de Gruyter. 2016.

HOTHORN, Torsten; EVERITT, Brian S. *A handbook of statistical analyses using R*. Boca Raton, FL: CRC Press. 2006.

HOTHORN, Torsten; HORNIK, Kurt; ZEILEIS, Achim. Unbiased Recursive Partitioning: A Conditional Inference Framework. *Journal of Computational and Graphical Statistics*, Caderno 15, 3, London, p. 651–674, 2006.

HUANG, Yan. 2000. *Anaphora: A cross-linguistic approach*. Oxford, UK: Oxford University Press.

JOHNSON, OLIVER A. Kant and The Great Compromise. *Akten des 4. Internationalen Kant-Kongresses*: Mainz, April 1974. Berlin/Munich/Boston: Walter de Gruyter, 1974. p. 109-114.

KASSAMBARA, Alboukadel. *Machine learning essentials:* Practical guide in R. Marseilles: STHDA, 2018.

KATO, Mary A. Strong and weak pronominals in the null subject parameter. *Probus.* Caderno 11, 1, p. 1-38, Bad Feilinbach, 1999.

KIBRIK, Andrej, A. Reference and working memory: cognitive inferences from discourse observations. In: HOEK, Karen Van; KIBRIK, Andrej, A.; NOORDMAN, Leo. *Discourse Studies in Cognitive Linguistics*: Selected papers from the 5th International Cognitive Linguistics Conference, Amsterdam, July 1997. Amsterdam/Philadelphia: John Benjamins Publishing Company, 1999. p. 29-52.

KIBRIK, Andrej A. *Reference in Discourse*. Oxford, UK: Oxford University Press, 2011.

LI, Charles N.; THOMPSON, Sandra A. Subject and topic: A new typology of language. In: LI, Charles N. *Subject and topic*. New York: Academic Press. 1976. p. 457-89.

MAYER, Mercer. *A boy, a dog, and a frog*. New York: Dial Press, 1967.

MONTALBETTI, Mario M. *After binding:* on the interpretation of pronouns. 1984. Tese. Massachusetts Institute of Technology.

NAIMI, Babak; HAMM, Nicholas A. S.; GROEN, Thomas A.; SKIDMORE, Andrew K. Skidmore; TOXOPEUS, Albertus G. Where is positional uncertainty a problem for species distribution modelling? *Ecography*, 37, 2, Lund, Sweeden, p. 191–203, 2014.

NEGRÃO, Esmeralda V.; VIOTTI, Evani. Brazilian Portuguese as discourse oriented language. In: KATO, Mary A.; NEGRÃO, Esmeralda V., *Brazilian Portuguese and the Null Subject Parameter.* Madrid: Iberoamericana Vervuert, 2000.

NUNES, Jairo. Subject and topic hyper-raising in Brazilian Portuguese. In: KATO, Mary A.; ORDÓÑEZ, Francisco, *The Morphosyntax of Portuguese and Spanish in Latin America.* Oxford: Oxford University Press, 2016.

OLIVEIRA DA SILVA, Andressa Christine; ALVES FONSECA, Aline. Prosody and Processing: Comprehension and Production of Topic-Comment and Subject-Predicate Structures in Brazilian Portuguese. *Revuista de Estudos de Linguagem*. Caderno 26, 4, p. 1601-1646, Universidade Federal de Juiz de Fora, Juiz de Fora, Brasil, 2018.

PLOTKIN, Henry. The Central Problem of Cognitive Science: The Rationalist-Empiricist Divide. *The Journal of Mind and Behavior*. Caderno 29, 1/2, p. 1-16, The University of Maine, Orno, 2008.

PONTES, Eunice. *O tópico no português do Brasil*, Campinas: Pontes Editores, 1987.

PRADA PÉREZ, Ana de. *The Interaction of Functional Predictors and the Mechanical Predictor Perseveration in a Variationist Analysis of Caribbean Spanish Heritage Speaker Subject Pronoun Expression.* Caderno 5, 4, p. 36, Basel, Switzerland, 2020.

PRATAS, Fernanda. *O Sistema Pronominal Do Caboverdiano (Variante de Santiago):* Questões de Gramática. Lisboa: Edições Colibri. 2004.

PU, Ming-Ming. Zero anaphora and grammatical relations in Mandarin. In: GIVÓN, Talmy. *Grammatical Relations:* a functionalist perspective. Amsterdam/Philadelphia: John Benjamins. 1997, p. 281-321.

OTHEGUY, Ricardo; ZENTELLA, Ana Celia. *Spanish in New York:* Language Contact, Dialectal Leveling, and Structural Continuity. Oxford, UK: Oxford University Press. 2012.

OTHERO, G. de Ávila; CYRINO, S. M. L.; MADIRD, L. T. Alves; ROSITO, R. B. V.; SCHABBACH, G. R. Objeto nulo e pronome pleno na retomada anafórica em PB: uma análise em corpora escritos com características de fala. *Revista Da Anpoll*. Caderno 1, 45, Universidade Federal de Santa Catarina, 2018. p. 68–89.

QUINT, Nicolas. *Grammaire de la langue cap-verdienne: étude descriptive et compréhensive du créole afro-portugais des Iles du Cap-Vert*. Paris: Harmattan. 2000.

QUINT, Nicolas. Les influences du portugais contemporain sur le système verbal du capverdien santiagais. In: CHAMOREAU, Claudine; GOURY Laurence; *Changement linguistique et langues en contact: approches plurielles du domaine prédicatif*. Paris: CNRS. 2012. p. 155-178.

QUINT, Nicolas. *Let's Speak Capeverdean: Language and Culture*. London: Battlebridge Publications. 2015.

R CORE TEAM. *The R Project for Statistical Computing*. 2021.

RIZZI, Luigi. On the status of subject clitics in Romance. In: JAEGGLI, Osvaldo A.; SILVA CORVALÁN, Carmen; *Studies in Romance linguistics*. Dordrecht, Netherlands: Foris, 1986. p. 391–419.

RODRÍGUEZ-RICCELLI, Adrián. *The Subject Domain in Cabo-Verdean Creole: Combining variationist sociolinguistics and formal approaches.* 2019. Tese. The University of Texas at Austin.

RODRÍGUEZ-RICCELLI, Adrián. Variable Subject Expression in Cabo-Verdean Creole: Some language-internal factors. In: LÉGLISE, Isabelle; MIGGE, Bettina; QUINT, Nicolas. *Journal of Pidgin & Creole Languages:* Creoles and variation: new trends and perspectives. Caderno 36, 1, p. 109-174, Amsterdam/Philadelphia: John Benjamins. 2021.

SIEWIERSKA, Anna. *Person*. Cambirdge, UK: Cambridge University Press. 2004.

SILVA, Claudia R.T. Comportamento e natureza dos sujeitos duplicados no crioulo caboverdiano e no português falado em comunidades quilombolas. In: MOURA, M. D. D.; SILBADO, M. A. *Para a história do português brasileiro:* Sintaxe comparativa entre o português brasileiro e línguas crioulas de base lexical portuguesa. Maceió: EDUFAL. 2013, p. 167-206.

SILVA, Claudia R.T., CARVALHO, Danniel; ZIOBER, Fernanda M. Traços de pessoa e duplos sujeitos no português. In: *XXXI Econtro nacional da ANPOLLGT*–Teoria da Gramática. Campinas, São Paolo. 2016.

SILVA, Claudia R.T.; CARVALHO, Danniel; ZIOBER, Fernanda M. Licenciamento de duplos sujeitos em variedades do português: pessoa, definitude e estrutura de traços. *Letras Escreve,* Caderno 7, 2, Universidade Feredal do Amapá, p. 91, 2017.

SILVA, Claudia R.T.; ZIOBER, Fernanda M. Sobre os sujeitos pré-verbais duplicados: uma análise contrastiva entre o português vernacular brasileiro, o caboverdiano eo santomé. *Estudos linguísticos e literários,* Caderno 57, Salvador, p. 164-85, 2017.

STROBL, Carolin; BOULESTEIX, Anne-Laure; KNEIB, Thomas; AUGUSTIN, Thomas Augustin; ZEILEIS, Achim. Conditional variable importance for random forests. *BMC Bioinformatics.* Caderno 9, 1, New York, p. 307, 2008.

STROBL, Carolin, BOULESTEIX, Anne-Laure; ZEILEIS, Achim; HOTHORN, Torsten. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics.* Caderno 8, 1, New York, p. 25, 2007.

TARALLO, Fernando. Por uma sociolingüística românica “paramétrica”: fonologia e sintaxe. *Cadernos de Linguística e Teoria da Literatura,* Caderno 7, 13, Universidade Federal de Minas Gerais, p. 53–85, 2015.

TARALLO, Fernando; KATO, Mary. Harmonia trans-sistêmica: variação intra- e inter linguística. *Revista Diadorim,* Caderno 2. Universidade Federal do Rio de Janeiro, 2007.

TORRES CACOULLOS, Rena; TRAVIS, Catherine E. Variationist typology: Shared probabilistic constraints across (non-)null subject languages. *Linguistics,* Caderno 57, 3, Bad Feilinbach, p. 653-92, 2019.

VIHMAN, Virve-Anneli; NELSON, Diane. Effects of Animacy in Grammar and Cognition: Introduction to Special Issue. *Open Linguistics,* Caderno 5, Warsaw, p. 260–267, 2019.

WAGNER, Susanne. Never saw one – first-person null subjects in spoken English. *English Language and Linguistics*, Caderno 22, 1, Cambridge, p. 1-34. 2016.

WICKHAM, Hadley. *Ggplot2:* Elegant graphics for data analysis. New York: Springer Verlag. 2016.

Nota do editor:
Artigo submetido para avaliação em: 21 de setembro de 2021.
Aprovado em sistema duplo cego em: 29 de dezembro de 2021.