

Differential Subject Expressions in Lelepa: A Quantitative Investigation

Yuchen Li

Bachelor of Arts (Honours)

The Australian National University

October 2023

This thesis is submitted in partial fulfilment of the requirements for the
degree of Bachelor of Arts with Honours in Language Studies in the
College of Arts and Social Sciences

Declaration

I hereby declare that, except where it is otherwise acknowledged in the text, this thesis represents my own original work. All versions of the submitted thesis (regardless of submission type) are identical. Chapters 1, 2, 3, and Section 4.2 partially draw on materials submitted for assessment in LING4013 *Essentials for Thesis Writing in Linguistics* and LING4106 *Advanced Readings in Linguistics*.

Yuchen Li

October 2023

Table of contents

Declaration	ii
Table of contents	iii
List of tables	vii
List of figures	ix
Abbreviations.....	xi
Acknowledgements	xii
Abstract.....	xiii
Chapter 1 Introduction	1
1.1 Overview	1
1.2 Language background	1
1.2.1 Linguistic classification	1
1.2.2 The speakers.....	3
1.2.3 Previous work	5
1.3 Aim of the study	5
1.3.1 Motivation	5
1.3.2 Research questions.....	6
1.4 Structure of the thesis.....	7
Chapter 2 Background	8
2.1 Overview	8
2.2 Referential choice.....	8
2.2.1 Overview	8
2.2.2 Between lexical and pronominal forms	9
2.2.3 Amongst pronominal references	14
2.3 Subject expressions in Austronesian languages	17
2.3.1 Overview	17
2.3.2 Free personal pronouns.....	17

2.3.3	Subject indexes	19
2.4	Subject expressions in Lelepa	21
2.5	Summary	25
Chapter 3	Methodology	27
3.1	Overview	27
3.2	Data selection.....	27
3.3	Data re-annotation.....	29
3.4	Token extraction and coding.....	32
3.5	Sample selection	33
3.5.1	Overview	33
3.5.2	Free subject pronouns.....	33
3.5.3	Subject proclitics.....	36
3.6	Statistical analyses	37
3.6.1	Overview	37
3.6.2	Study 1	38
3.6.3	Study 2	43
3.6.4	Study 3	43
3.6.5	Study 4	45
Chapter 4	Results	48
4.1	Overview	48
4.2	Study 1: GLMM analysis of free pronouns	48
4.2.1	Overview	48
4.2.2	Switch in referent.....	50
4.2.3	Humanness	54
4.2.4	Referent type.....	56
4.2.5	Direct speech	60
4.3	Study 2: Recursive partitioning analysis of free pronouns.....	66
4.3.1	Overview	66

4.3.2	Switch in referent.....	70
4.3.3	Person and number	70
4.4	Study 3: Mblogit analysis of free pronouns	78
4.4.1	Overview	78
4.4.2	Switch in referent.....	79
4.4.3	First occurrence in speech.....	82
4.4.4	Direct speech	85
4.4.5	Humanness	88
4.5	Study 4: Subject proclitic deletion based on multiple models	91
4.5.1	Overview	91
4.5.2	Preceding phoneme	94
4.5.3	Syllable structure	97
4.5.4	NP type	99
4.5.5	Mood	105
4.5.6	Person and number	109
4.6	Summary	114
Chapter 5	Discussion and conclusion	116
5.1	Overview	116
5.2	Free subject pronouns.....	116
5.2.1	Distribution	116
5.2.2	Functions.....	119
5.3	Subject proclitics.....	122
5.3.1	Distribution	122
5.3.2	Status	126
5.4	Referential choice.....	128
5.4.1	Referential choice and information structure.....	128
5.4.2	Referential choice and redundancy.....	130
5.5	Conclusion.....	135

References	138
-------------------------	------------

List of tables

Table 2-1: Typology of argument indexes based on the obligatoriness of coreferential NPs (conominals)	20
Table 2-2: Realis and irrealis subject proclitics in Lelepa and Nafsan	22
Table 2-3: Free personal pronouns in Lelepa.....	25
Table 3-1: Metadata for the recordings chosen from Lacrampe (2017)	28
Table 3-2: GRAID symbols used in this thesis	31
Table 4-1: Results of the GLMM analysis of free subject pronoun presence with principled undersampling with significant predictors in bold	49
Table 4-2: Results of the GLMM analysis of free subject pronoun presence with random undersampling with significant predictors in bold	50
Table 4-3: Free subject pronoun presence based on switch in referent	50
Table 4-4: Free subject pronoun presence according to humanness	54
Table 4-5: Free subject pronoun presence based on referent type	56
Table 4-6: Free subject pronoun presence based on whether in direct speech	60
Table 4-7: Free subject pronoun presence in terminal nodes.....	69
Table 4-8: Free subject pronoun presence under the first split.....	70
Table 4-9: Free subject pronoun presence in nodes 3-6 based on morphological ambiguity and referent switch	73
Table 4-10: Types of NP referencing 3sg subject referents.....	77
Table 4-11: Results of the mblogit analysis with reference levels and significant predictors in bold.....	78
Table 4-12: Types of NP based on switch in referent.....	79
Table 4-13: Types of NP based on first occurrence in speech	82
Table 4-14: Types of NP based on direct speech.....	85
Table 4-15: Types of NP in direct speech based on person	87
Table 4-16: Types of NP according to humanness.....	89
Table 4-17: Presence of subject proclitics based on preceding phoneme (1)	94
Table 4-18: Presence of subject proclitics based on preceding phoneme (2)	95
Table 4-19: Presence of subject proclitics based on syllable structure	97
Table 4-20: Presence of subject proclitics based on NP type.....	99
Table 4-21: Presence of subject proclitics based on mood	105

Table 4-22: Presence of subject proclitics based on mood and proclitic type	107
Table 4-23: Presence of subject proclitics based on number	110
Table 4-24: Presence of subject proclitics based on whether the referent is 2sg	113
Table 4-25: Results of the hypothesis-testing GLMM analysis with significant predictors in bold	114
Table 5-1: Preferred subject expressions in Lelepa based on accessibility .	130
Table 5-2: Prototypical semantic redundancy hierarchy of subject expressions in Lelepa.....	131
Table 5-3: Person and number values of 1du.excl, 1pl.excl, 3du, and 3pl free pronouns and subject proclitics.....	132
Table 5-4: Redundancy resolution for cross-references between a subject NP and a proclitic in Lelepa	134

List of figures

Figure 1-1: Family tree of Lelepa and some related languages, based on Lynch et al. (2002, pp. 4, 108–112) and Lynch & Crowley (2001, pp. 107–113)	3
Figure 1-2: A map of Lelepa and Efate in Vanuatu.....	4
Figure 3-1: A screenshot of re-annotation using GRAID in ELAN with the additional annotation tier named <i>GRAID@TSN</i>	30
Figure 4-1: Free subject pronoun presence based on switch in referent	51
Figure 4-2: Free subject pronoun presence according to humanness	55
Figure 4-3: Free subject pronoun presence according to referent type	57
Figure 4-4: Free subject pronoun presence based on whether in direct speech	61
Figure 4-5: Probability of free subject pronoun presence based on referent type and direct speech.....	65
Figure 4-6: Conditional classification tree of free subject pronoun presence .	68
Figure 4-7: Free subject pronoun presence in terminal nodes.....	69
Figure 4-8: Free subject pronoun presence under the first split.....	71
Figure 4-9: Free subject pronoun presence in nodes 3-6 based on morphological ambiguity and referent switch	73
Figure 4-10: Types of NP referencing 3sg subject referents	77
Figure 4-11: Types of NP based on switch in referent.....	80
Figure 4-12: Types of NP based on first occurrence in speech	83
Figure 4-13: Types of NP based on direct speech	86
Figure 4-14: Types of NP in direct speech based on person.....	88
Figure 4-15: Types of NP according to humanness	89
Figure 4-16: Importance scores of predictors of subject proclitic presence ...	92
Figure 4-17: Conditional classification tree of subject proclitic presence	93
Figure 4-18: Presence of subject proclitics based on preceding phoneme	95
Figure 4-19: Presence of subject proclitics based on syllable structure.....	97
Figure 4-20: Presence of subject proclitics based on NP type	100
Figure 4-21: Presence of subject proclitics based on mood	106
Figure 4-22: Presence of subject proclitics based on mood and proclitic type	107
Figure 4-23: Presence of subject proclitics based on number	110

Figure 4-24: Presence of subject proclitics based on whether the referent is 2sg	113
---	-----

Abbreviations

1	first-person	NOM	nominative
2	second-person	OBL	oblique
3	third-person	ORD	ordinal
ACC	accusative	pl	plural
ADD	addressee	PRF	perfect
ASS	associative	PURP	purposive
BEN	benefactive	R	realis
CERT	certainty	RED	reduplication
COME	directional (hither)	REL	relativiser
COMP	complementiser	REP	repetitive
COP	copula	SBST	substantiviser
DEM	demonstrative	SEQ	sequential
du	dual	sg	singular
EXCL	exclusive	STAT	stative
GO	directional (thither)	TR	transitiviser
HESIT	hesitation		
INCL	inclusive		
INDEF	indefinite		
INTERJ	interjection		
IPFV	imperfective		
IRR	irrealis		
LIM	limitative		
LOC	locative		
NEG	negative		
NMLZ	nominaliser		

Acknowledgements

First of all, my deepest thanks to my supervisor Rosey Billington for her tireless support and invaluable guidance along the way. I also wish to thank her for encouraging me to go on with morphosyntax when I was having a hard time deciding on my thesis topic, and for generously sharing lots of her knowledge, experiences, and materials with me. It would be impossible for me to come this far without Rosey's expertise.

I am also grateful to the Lelepa speakers whose recordings have facilitated this current study, and to Sébastien Lacrampe who collected and managed the Lelepa speech data and wrote the reference grammar. Although I did not have a chance to closely work with them, they really have inspired this thesis.

I also wish to express my thanks to a long list of academics at ANU who have taught or helped me at different stages since I commenced my journey in linguistics: Alexandra Marley, Alice Richardson, Amanda Laugesen, Beth Evans, Candida Spence, Carmel O'Shannessy, Catherine Travis, Danielle Barth, Evan Kidd, Harold Koch, I Wayan Arka, Jane Simpson, Jennifer Hendriks, John Giacon, Ksenia Gnevsheva, Lauren Reed, Mae Carroll, Michael Carne, Nick Evans, Sam Passmore, Shunichi Ishihara, Simon Gonzalez, Susy Macqueen, Toshiyuki Nakamura, Yuko Kinoshita, and Zhengdao Ye.

Finally, I would like to extend my gratitude to my family for their love and support all my life, to my peer linguistics students for the chats and fun we had together, to my roommates for the weekly hot pots, barbecues, and mahjongs, to Zed for those sleepless Overwatch and Minecraft nights, and my utmost thanks to Xin for the caring and accompaniment throughout the year.

Abstract

In Lelepa, an Oceanic language spoken in central Vanuatu, syntactic subjects are described as being expressed obligatorily through a subject proclitic at the beginning of the verb complex and optionally with a preverbal subject NP, headed by either a lexical noun or a free pronoun (Lacrampe, 2014). However, evidence has suggested that subject proclitics are sometimes deleted. Hence, Lelepa features four different subject expressions: a sole proclitic, a lexical NP, a free pronoun, or zero subject, whose contexts are nonetheless not well understood. This thesis draws on a corpus of Lelepa natural speech and investigates the variability in subject expressions quantitatively with logistic regression and recursive partitioning, accompanied by descriptive statistical data. Results have supported existing description that subject proclitics are the most prevalent subject expression, although they are also found deleted in clause-chains or in the presence of a coreferential lexical NP. In addition, Lelepa speakers' realisation of subject references in the corpus is usually the optimal choice that is neither ambiguous nor redundant, hence lexical NPs and free pronouns are commonly present where ambiguous but disfavoured elsewhere. However, redundant free subject pronouns may still be used for pragmatic functions such as floor-shifting or attention-getting, especially for human, direct-speech, or deictic speech-act participant referents. This thesis complements existing description of Lelepa morphosyntax, and extends our typological understandings of referential choice in Oceanic languages. The data pre-processing and statistical modelling in this thesis also demonstrate practical approaches to performing quantitative analyses of small and/or unbalanced datasets.

Chapter 1 Introduction

1.1 Overview

As home to 138 distinct languages, Vanuatu is known for its linguistic density and diversity (François et al., 2015). This thesis presents a comprehensive investigation into differential subject expressions in Lelepa, an underdescribed Oceanic language in central Vanuatu. Referential choice for subjects has shown notable variability both across languages, in that some languages prefer overt pronouns while some prefer bound forms or zero, and within languages, where both options are present but may be used to different extents. Lelepa is a language in which both options are present (Lacrampe, 2014), but there has not yet been a detailed investigation of factors influencing their usage. Drawing on available corpus data, this study aims to investigate Lelepa speakers' choice of differential subject expressions, and compare how the patterns found in Lelepa fit into cross-linguistic typologies of referential choice. This chapter is structured as follows: § 1.2 introduces the classification, speaker community, and existing studies of Lelepa, § 1.3 addresses the motivation and research questions of this thesis, and § 1.4 outlines the structure of the thesis.

1.2 Language background

1.2.1 Linguistic classification

Lelepa (ISO 639-3: lpa, Glottocode: lele1267) is a member of the South Efate

dialect network under the Nuclear Southern Oceanic linkage¹ within the Austronesian language family (Lynch et al., 2002, pp. 4, 108–112). The closest sister languages of Lelepa include Nafsan, Nguna, and Eton, from the same dialect network² (Lynch & Crowley, 2001, pp. 107–113). Compared with its sisters such as Nguna, which is described as the most conservative of the Efate dialects (R. Clark, 1985), and Nafsan, in which a number of sound changes have been completed (Billington et al., 2020; Thieberger, 2006), Lelepa exhibits signs of language change currently in progress (Lacrampe, 2014, p. 15). A family tree of Lelepa and related languages is illustrated in Figure 1-1 below.

¹ The term linkage, following Ross (1988, p. 8), refers to a group of languages formed by dialect differentiation, thus having no shared proto-language.

² Nafsan is also known as South Efate. Nguna is also known as Nakanamanga and North Efate.

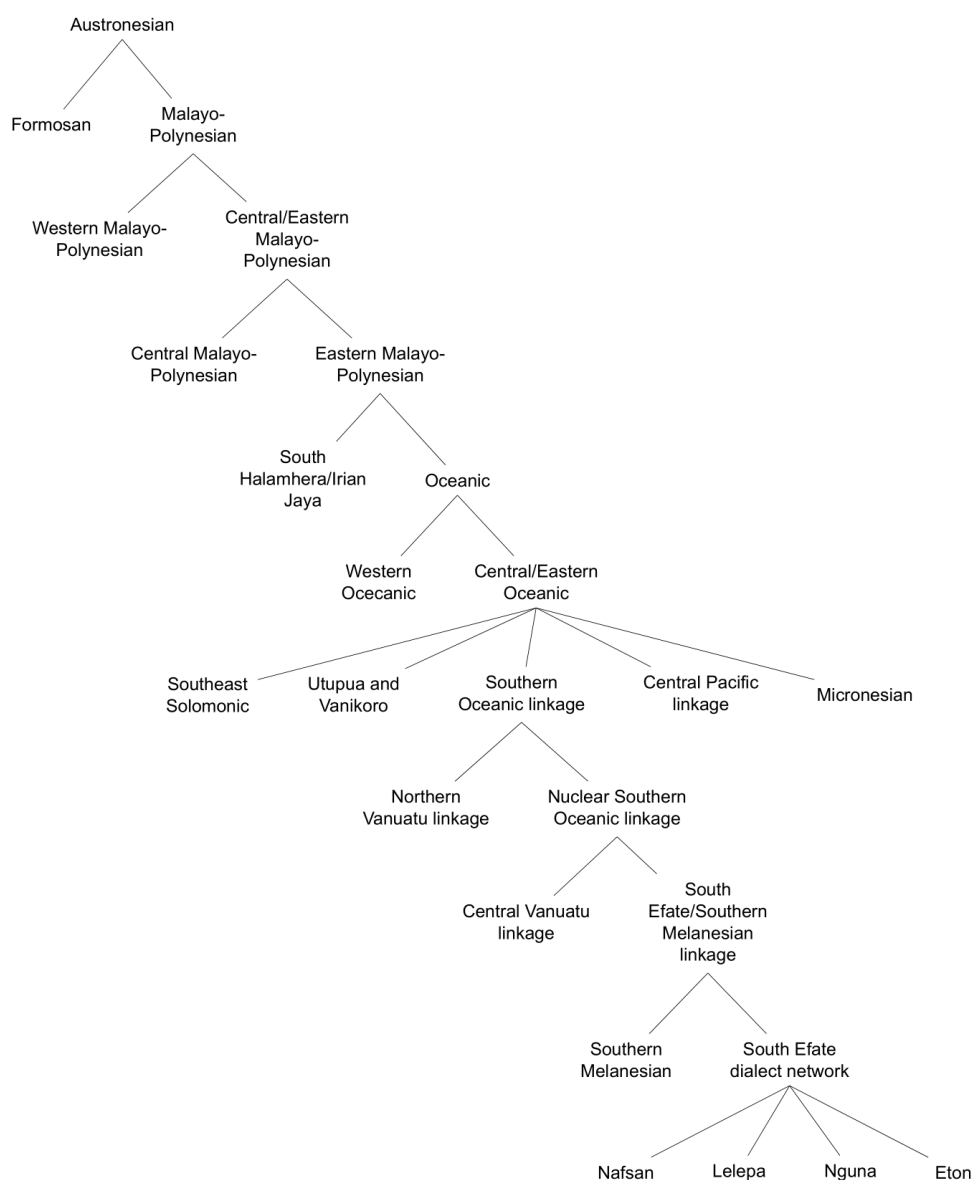


Figure 1-1: Family tree of Lelepa and some related languages, based on Lynch et al. (2002, pp. 4, 108–112) and Lynch & Crowley (2001, pp. 107–113)

1.2.2 The speakers

Lelepa is spoken by about 500 people on two islands: Lelepa and Efate (Lacrampe, 2014, p. 1). As of 2009, 387 Lelepa speakers reside in the Nataþao village at the south-eastern edge on the island of Lelepa, and a smaller Lelepa-speaking community of around 100 people live in the village of Mangaliliu on

the western coast of the Efate island (Lacrampe, 2014, p. 4). Mangaliliu was recently settled in 1983 by migrants from Lelepa, and residents in both villages remain in daily contact (Lacrampe, 2009, p. 9, 2014, p. 4). There is reportedly no dialectal variation between Lelepa spoken in the two settlements (Lacrampe, 2014, p. 4). A map of Lelepa and Efate and their position in Vanuatu is shown in Figure 1-2 below:

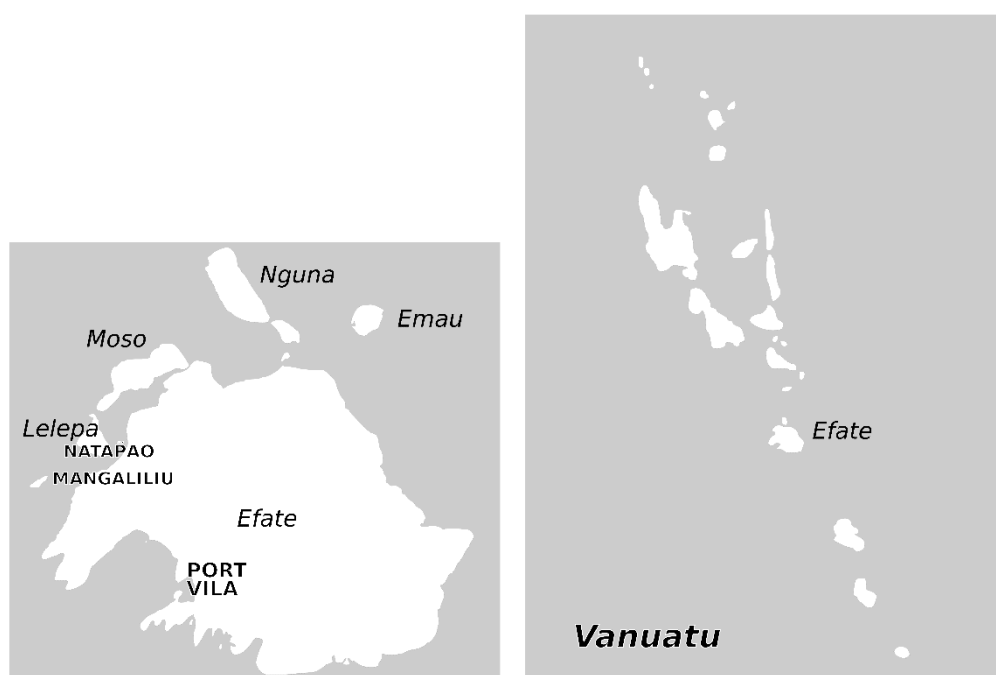


Figure 1-2: A map of Lelepa and Efate in Vanuatu³

Lelepa speakers are also highly multilingual. Lelepa is the first language to most children in the community (Lacrampe, 2014, p. 4). Bislama, an English-lexified creole that is the national language of Vanuatu, is also used as the household language in exogamous families. Lelepa is only taught at kindergarten level, while Lelepa children are educated in either English or French when they enter

³ This map is by courtesy of my Supervisor Dr Rosey Billington.

primary school (Lacrampe, 2014, pp. 4, 6). Lelepa speakers, especially those over forty, are also usually familiar with closely related Nafsan, Nguna, or both, through Christianisation practice or literacy teaching, since Lelepa has not been used as a church language, and there is also no strong writing tradition in the Lelepa community (Lacrampe, 2014, p. 5).

1.2.3 Previous work

Existing studies on the Lelepa language are limited. The main sources of reference are a descriptive grammar (Lacrampe, 2014), and associated work by the same author, including a master's thesis on the possessive constructions (Lacrampe, 2009), and a conference paper on vowel reduction and epenthesis (Lacrampe, 2012). Historical comparative studies also contribute a list of cognate lexicons between Lelepa and related languages (R. Clark, 1985; Tryon, 1976). Prior to these, there is an old sketch grammar which provides valuable diachronic references (Miller, 1945, cited in Lacrampe, 2014). The most recent study on Lelepa is an investigation of syllable structures (Sun, 2023).

1.3 Aim of the study

1.3.1 Motivation

Lelepa has a wide range of subject expressions. According to Lacrampe (2014), subjects in Lelepa can be expressed through an obligatory proclitic and an optional lexical or pronominal noun phrase (NP). However, quantitative studies across languages have demonstrated the presence of an overt NP is not just optional but conditioned by pragmatic factors. It remains to be seen whether this is also the case for Lelepa. Lacrampe (2014, pp. 241–212, 351) also

suggests there are instances of proclitic omission. However, the factors influencing this are not yet clear.

The Lelepa reference grammar (Lacrampe, 2014) is based on fieldwork undertaken by Lacrampe between 2005 and 2012, and a total of 20.5 hours' recordings across genres have been collected (Lacrampe, 2009, pp. 11–13, 2014, pp. 11–12). This existing corpus is available as an archival collection (Lacrampe, 2017), and allows for a comprehensive investigation to build on the grammatical description. Facilitated by the current corpus, a quantitative examination will help understand referential choice in Lelepa in more detail. Furthermore, as suggested by Simard & Wegener (2017), descriptive work in a more specific area of the language can also serve as a basis for future theoretical and typological studies.

1.3.2 Research questions

The general aim of this study is to find out the distribution and functions of differential subject expressions in Lelepa, and how they relate to the cross-linguistic typology of references. Consequently, this thesis aims to answer the following three research questions:

1. What factors relate to the presence of a free subject pronoun in Lelepa, and what are the functions of free subject pronouns when they occur?
2. What factors relate to the deletion of subject proclitics in Lelepa, and do deletion patterns suggest whether Lelepa subject proclitics are bound pronouns or agreement markers?

3. What conditions the final choice of differential subject expressions in Lelepa, between lexical NPs, free pronouns, proclitics, and zero forms?

1.4 Structure of the thesis

Following the introduction presented here, Chapter 2 comprises a literature review on referential choice across languages as well as a typology of subject expressions in Austronesian languages, with a focus on Lelepa. Chapter 3 details the data, procedure, and statistical models implemented in this thesis, followed by Chapter 4 presenting the results of the statistical analyses. Chapter 5 situates the findings within existing studies and offers concluding comments for the thesis.

Chapter 2 Background

2.1 Overview

This literature review chapter aims to contextualise the thesis and set the stage for subsequent chapters by establishing a theoretical framework for referential choice and presenting a typology of differential subject expressions. Section 2.2 explores critical studies on referential choice across languages. Section 2.3 constitutes a typological overview of two non-lexical subject expressions in Austronesian languages: free personal pronouns and argument indexes. This is followed by a concise summary of subject expressions in Lelepa according to the reference grammar (Lacrampe, 2014) with notes on related languages in 2.4. Section 2.5 discusses the gaps in existing studies and motivations for this current study.

2.2 Referential choice

2.2.1 Overview

Languages have different strategies to reference entities in discourse, known as *referential choice* (Kibrik et al., 2016). The basic referential choice is between reduced pronominal and full lexical forms, while the differential realisation between proper names and descriptions within lexical references is also acknowledged in Kibrik et al. (2016, p. 2). Since the realisation between multiple lexical references is beyond the scope of this thesis, this section focuses on the choice between lexical, overt pronominal, and zero referential forms across languages. Section 2.2.2 outlines the choice between lexical and pronominal

forms, and Section 2.2.3 turns to specifically the overt and zero realisations of pronominal references.

2.2.2 Between lexical and pronominal forms

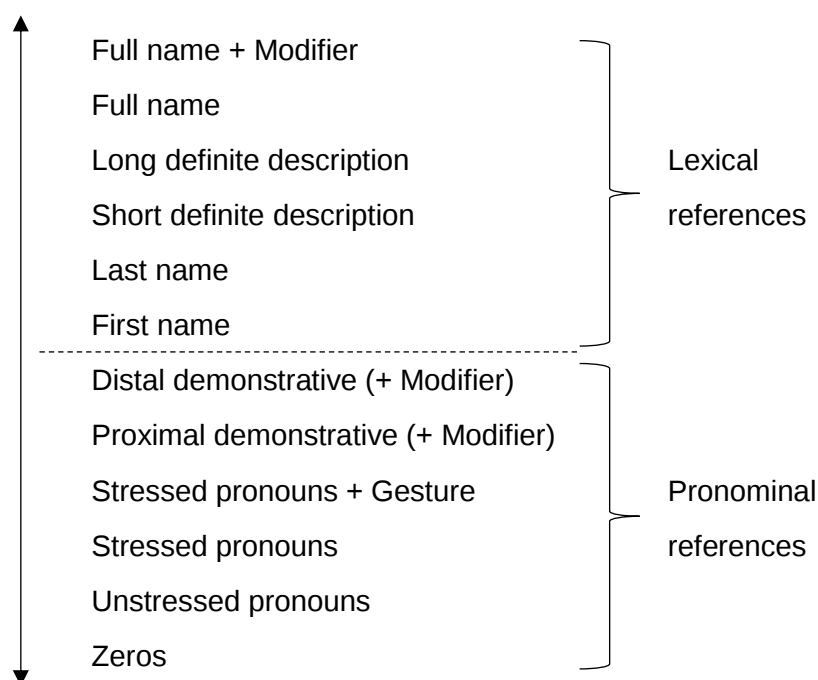
Many existing studies have suggested referential choice relates to information structure, wherein the information varies in *givenness* and *salience*. In discourse, speakers typically presuppose a set of information shared between the interlocutors, known as *common ground* (Stalnaker, 2002). Both factual and discourse-specific information may be stored in the common ground (Krifka, 2008, p. 16). Since common ground is changing continuously, there is naturally a distinction between what is *new* and what has been *given* in discourse. There is a cross-linguistic trend for less given referents to be overtly marked (Ariel, 1991, 2001; Chafe, 1994; Givón, 1983c; Greenfield & Smith, 1976; Gundel, 1996, 2010; Gundel et al., 1993 *inter alia*). In addition, since speakers tend to balance the need to be informative and the desire to be efficient when referencing (Arnold & Zerkle, 2019, p. 1162), the overt expression of a highly salient referent may be considered discourse-redundant⁴ (Arnold & Zerkle, 2019; Gudmestad & Geeslin, 2010). Hierarchical analyses have been proposed in Givón (1983b), Ariel (1991, 2001) and Gundel (1996, 2010; Gundel et al., 1993) in similar manners, based on the *predictability*, *accessibility* or *givenness* to account for the choice among referential expressions ranging from long descriptive lexical NPs to fully reduced zero forms. Zero forms are commonly regarded as recoverable silent pronominals in current formal analyses (e.g. Carnie, 2021, p. 446; Sportiche et al., 2013, p. 244 *inter alia*), and this thesis

⁴ This is different from *semantic* redundancy in the following subsection (see § 2.3).

follows this tradition in the categorisation. For example, the accessibility hierarchy adapted from Ariel (1991, p. 449) is shown in (1) below.

(1)

LOW ACCESSIBILITY



HIGH ACCESSIBILITY

Existing studies across languages also support the existence of the accessibility hierarchy. The choice between lexical and pronominal references mostly pertains to third-person referents, since reference tracking for speech-act participants⁵ (SAPs) is much easier (Shin & Cairns, 2009). Mithun (1990) summarises that only lexical NPs rather than general third-person pronouns can refer to entirely new referential entities outside of the consciousness of both interlocutors in Central Pomo, a Pomoan language spoken in North America.

⁵ i.e. first- and second-persons, also referred to as SAP(s) henceforth (*cf.* Goffman, 1974, 1981; McCawley, 1999 *inter alia*).

New referents are also less likely to have overt pronominal references than re-introduced references in Spanish, and instances are usually associated with first- and second-person referents, whereas third-person new referents favour lexical realisation (Carvalho & Child, 2011). Givón's (1983a) survey across nine languages and varieties from different language families also demonstrates that continued referents favour pronominal references across languages beyond Romance languages. In addition, switch referents⁶ also prefer overt pronominal references to zero forms in Spanish and Portuguese (Amaral & Schwenter, 2005; Cameron, 1995; Jacobson, 2016; Lastra & Butragueño, 2015; Posio, 2011, 2013; Torres Cacoullos & Travis, 2019; Travis & Torres Cacoullos, 2012 *inter alia*). Accessibility also affects referential choice in Austronesian languages. Arka & Sedeng (2018, pp. 155–158) note switch referents tend to be referenced with lexical NPs in Sembiran Balinese, while highly salient referents prefer pronominal references, either zero (67%) or overt (15%).

In addition, Beaver (2004) suggests such hierarchical proposals are developed from the relation between form and topicality. In Beaver's (2004) summary, topical referents tend to be pronominalised along with their incoming discourse salience (*cf.* Bresnan, 2001 *inter alia*). Frascarelli and Hinterhölzl (2007, p. 89) identify three types of topics: *shifting topic*, *contrastive topic*, and *familiar topic*. Once a (re-)introduced referent has been established as a shifting topic with an overt expression, it becomes a familiar topic and typically receives least prominence in pronominal forms, unless the speaker wants to contrast it against a semantically compatible alternative (Frascarelli & Hinterhölzl, 2007; Givón, 1983b; Krifka & Musan, 2012; Pesetsky, 1987 *inter alia*). Although the

⁶ Switch referents include both new and re-introduced referents.

identification of topics and their types is idiosyncratic across languages, referents are universally more likely to be topics according to the *topicality/animacy hierarchy*. Initially proposed by Hawkinson and Hyman (1974) for Shona, a Bantu language spoken in Zimbabwe, the topicality hierarchy has since been extended to other Bantu (e.g. Duranti, 1979; Kidima, 1987) and non-Bantu languages (e.g. Dunbar, 1983 on German; Iemmolo, 2010 on Romance languages). Selected relevant elements relating to the topicality hierarchy is shown in (2) below, adapted from Dunbar (1983, p. 142), where referents to the left-hand side are more topical and more likely to be pronominalised:

- (2) 1st-person > 2nd-person > 3rd-person
 Singular > Plural
 Human > Animate > Inanimate
 Definite > Indefinite
 Proper > Common
 Discourse anaphoric (given) > Discourse non-anaphoric (new)

Compared with the accessibility hierarchy, the topicality hierarchy additionally highlights the effects of person, number, and animacy. The effect of person is instead a binary division between SAP and third-person referents, as mentioned above. As for number, multiple studies in Spanish show plural pronouns are rarer in discourse and there is less variability between their presence and topic continuity (Cameron, 1992; Flores-Ferrán, 2004; Shin & Cairns, 2009). In addition, the preference for pronominal references with animate referents has been cross-linguistically attested. Genetti & Crain (2003) note pronominal references are highly associated with animacy in Nepali, an Indo-Aryan language of Nepal, based on which they propose the *animate pronoun*

constraint. Similar patterns have also been observed in Gurindji and Bilinarra, two Australian languages (Meakins, 2015).

However, more topical referents may prefer overt pronominal references to zero forms. Genetti & Crain (2003) mention that inanimate referents are referenced with either lexical NPs or zero forms in Nepali, and this is also the case in Spanish and Portuguese (Schwenter, 2006, 2014). Among animate or human referents, null pronouns are more associated with third-person than SAPs (C. Huang, 2011; Orozco, 2015; Torres Cacoullos & Travis, 2019). Torres Cacoullos & Travis (2018, pp. 88–89) also mention that free subject pronouns are preferred in direct speech in Spanish, especially for 1sg referents compared with 3sg referents (though the rare occurrence of direct speech in the data does not allow its effect to be systematically examined). Schnell (2018b) also suggests SAP referents are highly associated with direct speech in narrative texts. Based on this suggestion, it may be worth examining the effect of direct speech and SAP across genres. The preference for overt pronouns with topical referents, despite their higher discourse accessibility, is also attested in Oceanic languages. In Vera'a, an Oceanic language of northern Vanuatu, animate or SAP referents prefer overt pronouns to zero, although singular referents are more likely to have zero reference than plural ones (Schnell, 2018b; Schnell & Barth, 2020). A human referent also takes an independent pronoun reference while inanimates are referenced by zero forms in Roviana, spoken in the Solomon Islands (Corston-Oliver, 2003), and Nêlêmwa (Bril, 2004), spoken in New Caledonia. These studies all suggest that topicality and discourse accessibility predict referential choice differently, and that there may be other factors that condition the choice between pronominal and zero forms.

2.2.3 Amongst pronominal references

Languages vary in their choice of not only referential forms but also the realisations of these. Krahmer and van Deemter (2012, p. 204) note that the choice between lexical and pronominal references takes place at the first stage, and then the content or realisation is determined. For pronominal references, the differential realisation at the second stage is between full and reduced forms based on Krahmer and van Deemter (2012, p. 204). The omission of pronominal references is usually termed *pro-drop*. This subsection discusses the influencing factors for *pro-drop* across languages.

Pro-drop is a widely adapted approach to avoid redundancy. The term redundancy is used in two different scenarios: either the overt expression of a highly predictable referent (*cf.* Gudmestad & Geeslin, 2010 *inter alia*), or the duplication of information by more than one source (Horning, 1979, p. 315). The former one is termed *discourse redundancy* in this thesis, and the latter one is termed *semantic redundancy*. Semantic redundancy can be grammatical or contextual. Grammatical redundancy is the obligatory double expression required by language-internal rules, such as the English clause 'he runs', where *he* and *-s* express the person and number of the same referent but neither can be deleted (Wit & Gillette, 1999, p. 5). Contextual redundancy includes all grammatically non-obligatory instances (Wit & Gillette, 1999, p. 9). Heltai (2005) comments that contextual⁷ redundancy can be either *functional*, which helps comprehend the proposition, or *disfunctional*, which hinders understanding. In languages with rich agreement morphology, pro-drop avoids semantic

⁷ Originally termed *optional redundancy* but with the same definition as *contextual redundancy*.

redundancy. Leufkens (2020, pp. 87, 95) suggests a partially redundant lexical item may be used for unambiguity when the information provided by a grammatical marker is limited, otherwise it is omitted. This is clearly exemplified in the Oceanic language Vera'a, which has no dedicated subject proclitics but retains agreement in some TAMP markers. In Vera'a, subject pronouns are commonly present, but tend to be zero when the TAMP marker expresses number agreement, so that the referent number is not redundantly expressed (Schnell & Barth, 2020). However, Ruda (2015) indicates it is also possible for an agreement-rich language to manifest agreement-drop in the presence of an overt subject. Nichols (2018 *inter alia*) also notes that overt pronouns tend to be present when the argument index (see also § 2.3.3) is absent in the *complementarity hypothesis*. Therefore, it is not mandatory for an overt pronoun to be dropped in agreement-rich languages. Moreover, in Vollmer's (2023) recent investigation across eight languages, pro-drop is also observed in languages with no or limited agreement morphology, suggesting that pro-drop also avoids discourse redundancy, and that more accessible referents tend to attract pro-drop. Furthermore, Torres Cacoullos & Travis (2019) find that the same discourse constraints for pro-drop in Spanish also apply to the rare instances of pro-drop in English, a non-pro-drop language.

In pro-drop languages, overt pronouns naturally indicate emphasis and have pragmatic functions. Free subject pronouns are commonly attested as *floor-shifting* devices in conversations in pro-drop languages, whose agreement morphology may be rich, like Spanish (Davidson, 1996; Travis & Torres Cacoullos, 2012) or limited, like Japanese (D.-Y. Lee & Yonezawa, 2008), Korean (N. Lee, 2021), and Javanese (Ewing, 2014). As defined by Hayashi (1987, 1991), *floor* in conversation structure refers to a strategy of

communicative awareness orientation between interlocutors, and a floor shift indicates a previous addressee has taken over the role as a speaker. Pronouns marking floor-shifts may undergo further grammaticalisation and become attention-getters, as termed by Trillo (1997). For instance, in Longxi Qiang, a Tibeto-Burman language of China, speakers tend to use the first-person singular topic pronoun *qá mà* as a pragmatic marker to attract the collocutor's attention to their thoughts or response (Zheng, 2020). The pronoun *qá mà* as a pragmatic marker in Longxi Qiang has been further grammaticalised and become non-truth-conditional, hence they are also used in cases where the topic is not first-person singular (Zheng, 2020).

Although most cases of pro-drop also qualify as *null anaphora* (Hartmann & Stork, 1972, pp. 258–259), the omission of an expletive subject is an exception. Expletive subjects are non-argumental or non-referential subjects, which cannot be replaced by any other NP (e.g. Biberauer, 2010). Weather-related events and existential clauses are usually expressed with expletive subjects across languages (Bleotu, 2012; Cardinaletti, 1997; Krejci, 2014; Levin & Krejci, 2017; Pescarini, 2022). For example, the pronoun *it* in the English sentence 'it rains' is not anaphoric to any entity but fills the subject position. A null expletive subject is also believed to be a feature of pro-drop languages (e.g. LaFond, 2001). Like other zero pronouns, the null expletive subject is also analysed as a covert argument *pro* and the argument index on the verb is regarded as an agreement marker, which agrees with either the associate⁸, as in Standard Italian, or the silent *pro*, as in some Northern Italian varieties (Cardinaletti, 1997; Pescarini,

⁸ For example, *three girls* in the English existential sentence *there come three girls* (Cardinaletti, 1997, p. 521).

2022). Overall, overt pronouns are less preferred across languages but are commonly used for pragmatic purposes, and languages which allow zero pronouns tend to also allow zero expletives.

2.3 Subject expressions in Austronesian languages

2.3.1 Overview

Pronominal references for subjects distribute and function differently across Austronesian languages. Free personal pronouns are characterised by their syntactic and pragmatic significance in Western Austronesian (i.e. Formosan and Western Malayo-Polynesian/WMP) languages, whereas in Oceanic languages they may be thought of as optional (Neeleman & Szendrői, 2007). Apart from free pronouns, there are usually a range of bound forms that encode person and number, which may be regarded as either agreement markers or bound pronouns in different languages. This section presents the morphosyntactic diversity of subject pronominals in Austronesian languages, with § 2.3.2 focusing on free personal pronouns, and § 2.3.3 on bound person markers.

2.3.2 Free personal pronouns

Free personal pronouns are morphologically diverse across Western Austronesian languages. In a number of Formosan languages (e.g. L. M. Huang, 1995 on Atayal; Li, 2000 on Pazeh; Teng, 2005, 2008 on Puyuma), there are distinct sets of neutral/topic free pronouns⁹ apart from nominative

⁹ The topic pronouns can reference not only subjects but also other grammatical roles.

forms. In these languages, neutral/topic pronouns are coreferential with a bound pronoun for emphasis. Some Formosan languages also possess two or multiple sets of third-person free pronouns based on *visibility* and *distance*. In languages where there are two sets of third-person pronouns such as Rukai (Zeitoun, 1997, 2007) and Tsou (Zeitoun, 2000, 2005), the distinction is based on whether the referent is *visible* or not, whereas in languages like Pazeh (Li, 2000), third-person pronouns exhibit a tripartite categorisation, where the *visible* pronouns are further divided into *proximate* and *distant*. A few other Formosan languages like Saisiyat (X. Huang & Hsieh, 2006; Yeh, 2003) and Thao (Blust, 2003) have also developed more complete grammatical case systems. For example, in Saisiyat, six grammatical cases are attested in free pronouns (Yeh, 2003, p. 17). Case morphology is also attested in free pronouns in some Austronesian languages in Philippine (e.g. Rubino, 1997, p. 52 on Ilocano; Schachter & Otnes, 1983, p. 88 on Tagalog) and Indonesia (Klamer, 1998 on Kambera from the Sumba Island). Free pronouns in a few Indonesian languages additionally feature speech levels (Suwarno & Yanwar, 2020), such as Balinese (Arka, 2003), Indonesian (Sneddon et al., 2012; Sneddon & Neil, 2010), and Javanese (Poedjosoedarmo, 1968), whose speakers use different sets of free pronouns to express intimacy, politeness, or respect to a higher social status, respectively. Free personal pronouns also interact with verbal morphology and constituent order to indicate discourse focus in Western Austronesian languages, and arguments with various semantic roles can be brought to the subject position by different verbal voice markers, depending on the focus of the clause (Chen & McDonnell, 2019).

As for Oceanic languages (see also Figure 1-1 in § 1.2.1 for a classification), free pronouns appear to be less informative than in Western Austronesian

languages, and are commonly deleted (Neeleman & Szendrői, 2007). Neeleman & Szendrői (2005) suggest that languages with fusional case morphology on free pronouns like English do not allow pro-drop whereas languages with agglutinating or zero case morphology like Japanese and Mandarin do. Based on the fact that most Oceanic languages either do not exhibit case morphology at all, or only have agglutinated case-markers, as in many Eastern Oceanic languages (R. Clark, 1973), Neeleman & Szendrői (2007) additionally comment that the optional nature of Oceanic free pronouns is not likely a result of rich agreement, since free pronouns are commonly deleted in Oceanic languages with both rich agreement, such as Niuean (Massam et al., 2012), and limited agreement, such as Cheke Holo spoken in the Solomon islands (Palmer, 2009).

2.3.3 Subject indexes

The status of bound forms, or *argument indexes* (Haspelmath, 2013) is complex across languages, both within and beyond the Austronesian family. Research in this area has used a range of different terms and criteria to distinguish agreement markers and bound pronouns. A common approach to distinguish different argument indexes is based on the obligatoriness of coreferential NPs, or *conominals* (Haspelmath, 2011, 2013; Siewierska, 1999; Wegener, 2012), as presented in Table 2-1 below. Corbett (2003, pp. 188–189) additionally comments that an argument index should be considered as an agreement marker if it occurs more than once in a clause.

Table 2-1: Typology of argument indexes based on the obligatoriness of coreferential NPs (conominals)

Conominal	Type	Reference(s)
Obligatory	Agreement marker	(Haspelmath, 2013; Siewierska, 1999)
Optional	Bound pronoun with cross-reference	(Bloomfield, 1984; Hockett, 1958; Ordóñez & Treviño, 1999)
Optional	Agreement marker with pro-drop	(Barbosa, 2019)
Prohibited	Bound pronoun	(Haspelmath, 2013; Siewierska, 1999)

Argument indexes are not analogous between Western Austronesian and Oceanic languages (see also Figure 1-1 in § 1.2.1). In Western Austronesian languages, subject indexes mostly fall into the bound pronoun category, since they are in complementary distribution with free pronouns or nominals, unless the conominal becomes a topic which is then external to the clause's argument structure (e.g. L. M. Huang, 1995; Li, 2000; Pan, 2012; Teng, 2008; Zeitoun, 2000, 2007). In Oceanic languages, subject indexes are commonly obligatory with optional conominals which are only used to indicate emphasis (Lynch, 2016). While the exact status of subject indexes across and within different Oceanic languages is unclear, many existing studies analyse them as agreement markers (e.g. Yamada, 2006 *inter alia*). The agreement marker analysis is also supported by languages which only display agreement in some of the TAM markers, such as Vera'a (Schnell & Barth, 2020 *inter alia*), or languages whose subject indexes are optional except for in specific environments, such as Marovo from the Solomon Islands (B. Evans, 2008).

2.4 Subject expressions in Lelepa

This section summarises the description of subject expressions in Lelepa from the reference grammar (Lacrampe, 2014). In Lelepa, the basic constituent order is SV/AVO/AVRT (Lacrampe, 2014, p. 252). Subjects are expressed through two preverbal elements: an obligatory subject proclitic at the left edge of the verb complex, and an optional subject NP in clause-initial position (Lacrampe, 2014, pp. 243–245).

The most common subject expression in Lelepa is the subject proclitics. Lacrampe (2014, p. 244) describes the subject proclitics as anaphoric agreement markers when conominated, and as bound pronouns elsewhere. Table 2-2 below illustrates subject proclitics in Lelepa and Nafsan, a closely related language from the dialect network (adapted from Lacrampe, 2014, p. 243; Thieberger, 2006, p. 105). The only non-analysable irrealis form in both languages is the second-person singular $\tilde{p}a=^{10}$. As for other proclitics, a particle ga^{11} is attached to the realis forms to indicate irrealis mood in Lelepa. In Nafsan, an irrealis affix k is suffixed to proclitics with consonant onsets and affixed to proclitics with zero onset. In Lelepa, there are also two 1pl.excl subject proclitics: $ur=$ and $au=$. Only some older speakers occasionally use $au=$ in speech, whereas other speakers use $ur=$ all the time (Lacrampe, 2014, p. 243). This has led to the syncretism between 1pl.excl and 3pl proclitics in Lelepa. Based on an older source (Miller, 1945) which lists $ur=$ as 3pl only and $au=$ as 1pl.excl, Lacrampe (2014, p. 243) infers that this alternation in 1pl.excl subject

¹⁰ The orthographic < $\tilde{p}$ > corresponds to the phoneme / $\widehat{k}p$ / in Lelepa/

¹¹ The orthographic < g > corresponds to the phoneme / η / in Lelepa.

proclitics is a nearly completed change in progress within the last five to six decades.

Table 2-2: Realis and irrealis subject proclitics in Lelepa and Nafsan

	Lelepa		Nafsan	
	realis	irrealis	realis	irrealis
1sg	a=	aga=	a=	ka=
1du.incl	ta=	taga=	ta=	tak=
1du.excl	ar=	arga=	ra=	rak=
1pl.incl	tu=	tuga=	tu=	tuk=
1pl.excl	ur=, au=	urga=, auga=	u=	ko=
2sg	ku=	ṗa=	ku=	ṗa=
2du	kar=	karga=	ra=	rak=
2pl	kur=	kurga=	u=	ko=
3sg	e=	ega=	i=	ke=
3du	ar=	arga=	ra=	rak=
3pl	ur=	urga=	ru=	ruk=

Although subject proclitics are described as obligatory constituents, proclitic omission in the presence of a subject NP has also been noted in the reference grammar (Lacrampe, 2014, pp. 241–212, 351). When the subject NP in a copular clause is headed by *nagi* ‘name’, either the proclitic or both the proclitic and the copula can be unconditionally omitted. However, the subject proclitic cannot stand alone with only the copula being omitted (Lacrampe, 2014, pp. 241–242). The constraint on proclitic omission elsewhere is nonetheless not described in detail. Subject proclitics are also obligatory in closely related Nafsan and Nguna, although proclitic omission is also observed in both languages. Schütz (1969, p. 3) mentions ‘the elision of verbal-pronoun markers’ in Nguna which is nonetheless not accounted for in the sketch grammar. In

Nafsan, subject proclitics are optional in clause-chains (Thieberger, 2006, pp. 324–327). A clause-chain is a group of clauses where the tense-aspect-mood (TAM) or argument(s) are not overtly expressed but rather inferred from the fully inflected clause (Longacre, 1999, 2007). Thieberger (2006, pp. 324–327) remarks that only the first verb in a clause-chain obligatorily takes a subject proclitic, while the following verb(s) can optionally take no overt subject marking, as the subject can still be understood through the context. Apart from clause-chains, in various Oceanic languages of Vanuatu, information flow is also commonly manipulated through tail-head linkage, which is the repetition of the final part of a clause at the beginning of a subsequent one, for example, in Abma (Schneider, 2009, p. 24), Erromangan (Crowley, 1998, p. 282), Lewo (Early, 1994, p. 454), Lolovoli (Hyslop, 2001, p. 426), and Nafsan (Thieberger, 2006, pp. 327–328). Unlike in clause chains, subject proclitics in Nafsan tail-head linkage are not deleted (Thieberger, 2006, pp. 327–328). Lacrampe (2014, pp. 166–168) notes the existence of clause-chaining in Lelepa, but the obligatoriness or otherwise of subject proclitics in clause-chains has not been studied in detail. An example of a subject proclitic in a clause extracted from the reference grammar is presented in example (3) below:

- (3) **a=maturu**
 1sg.NOM=sleep
 ‘I slept.’ (Lacrampe, 2014, p. 91)

While the obligatoriness of subject proclitics are unclear, subject NPs in Lelepa are generally absent when the referent is accessible (Lacrampe, 2014, p. 270), in accordance with the cross-linguistic trend mentioned in § 2.2 above. NP constituents can also be present and fronted (or *left-dislocated*) to mark contrast

in Lelepa (Lacrampe, 2014, pp. 215, 272). However, unlike verb-initial languages like Samoan (Calhoun, 2015) and Gela (Simard & Wegener, 2017), left-dislocated and clause-initial subject NPs in Lelepa are in formally identical positions in clauses. Lacrampe (2014, pp. 271–272) notes that left-dislocated subjects differ from clause-initial subjects in that they form their own prosodic phrase with a pause after. A perceptible cesura separating a pre-clausal construction and the main clause into two prosodic phrases is also widely attested across Oceanic languages (Macdonald, 2017, pp. 105–130).

Subject NPs in Lelepa can have either a lexical or a pronominal head. Most commonly, the subject position is filled with a lexical NP to introduce a switch referent which is either new or re-introduced to the discourse (Lacrampe, 2014, pp. 212–214, 270–273). Lexical NPs can also reference continuous referents for topic continuity to a lesser extent (Lacrampe, 2014, p. 214). Personal pronouns can also head subject NPs to establish new or re-introduced referents as switch topics, especially human referents (Lacrampe, 2014, p. 214). As shown in Table 2-3 below, adapted from Lacrampe (2014, p. 146), free personal pronouns in Lelepa are morphologically simple compared with both subject proclitics in Lelepa and free pronouns in Western Austronesian languages. Pearce (2012) observes that either free subject pronouns or subject proclitics can encode a dual distinction in number across Vanuatu languages. Comparing Table 2-2 and Table 2-3, it is obvious that Lelepa has a more elaborated number system on subject proclitics than on free pronouns, hence free pronouns are less informative and may be considered redundant (Leufkens, 2020).

Table 2-3: Free personal pronouns in Lelepa

1sg	1pl.incl	1pl.excl	2sg	2pl	3sg	3pl
konou	kinta	kenem	nag	kumu	nae	naara

In addition, since Lelepa does not distinguish topic/neutral and nominative pronouns, pronominal subjects and topics have identical forms in Lelepa, except for contrastive topics which are either left-dislocated, or marked by a demonstrative *na* or a topic marker *wei* (Lacrampe, 2014, pp. 120, 146, 191, 215–216). An example of free subject pronoun as a left-dislocated contrastive topic is illustrated in example (4) below:

- (4) $\bar{p}a=to.$ ***konou,***
 2sg.IRR.NOM=stay **1sg**
a=ga *kat* *sak* *pan.*
 1sg.NOM=IRR CERT go.up GO
 ‘You will stay. As for me, I will go up.’ (Lacrampe, 2014, p. 146)

2.5 Summary

To sum up, subject expressions in Lelepa need further investigation. Compared with other Austronesian languages, the functions of free pronouns in Lelepa are highly limited in that they do not indicate grammatical relations, topicality, proximity, speech levels, and additional grammatical number. Current descriptions regard Lelepa free pronouns as completely optional and only used to mark (re-)introduced or contrastive referents. However, how frequently they are used in Lelepa speech, and what other functions they may have when present is yet unclear. For subject proclitics, since existing evidence has

suggested they are occasionally deleted, it is also worth examining what factors condition their deletion in Lelepa. Findings of a systematic investigation may also suggest whether subject proclitics are agreement markers or bound pronouns, and to what extent referential choice in Lelepa accords with or differs from existing studies across languages.

Chapter 3 Methodology

3.1 Overview

This chapter presents the data, methods, and procedure used in the study. Section 3.2 offers an overview of the recording selection, the size of the corpus, and speaker metadata. Section 3.3 introduces the annotation schema I follow during the data re-annotation process. Section 3.4 outlines the total number of tokens in the corpus as well as their extraction and coding process, and Section 3.5 details the subsets of data used in each study and their selection criteria. Section 3.6 summarises the statistical models, variables, and procedure of the four studies.

3.2 Data selection

This study draws on a 57-minute sample of narrative data produced by 13 Lelepa speakers, 5 female and 8 male, collected from 2006 to 2010 and archived in ELAR as part of a larger open-access corpus (Lacrampe, 2017)¹². Each of these narratives comes with both an audio recording in .wav format and a transcription file in .eaf format (ELAN, 2023) with full translation of the text¹³. The genres include folktales, and personal, traditional, descriptive, and historical narratives. Procedural narratives are speakers' description of their routine activities such as cooking or fishing (Lacrampe, 2014, p. 10). A special type of folktales is *naluokia*, a form of oral tradition in Lelepa like a proverb that

¹² See also § 1.3.1 for more detail of the corpus.

¹³ The full-sentence translation also comes with word-by-word glossing in most transcriptions.

is usually didactic and metaphorical (Lacrampe, 2014, pp. 10, 501). The speaker metadata, genre¹⁴, recording date, duration and number of tokens for analysis in each recording chosen are summarised in Table 3-1 below:

Table 3-1: Metadata for the recordings chosen from Lacrampe (2017)

Item	Speaker ID	Sex	Genre	Date	Duration (mm.ss)	Number of tokens
SL1- 20060402Toms enFrance-TNIs	TSN	M	Narrative	2006- 04-02	9.52	310
SL1- 20070420Harrie rFalcon-GMmp	GKM	M	Folktale	2007- 04-20	3.55	84
SL1- 20080401DevilA rtok-CMmut	CKM	M	Folktale	2008- 04-01	12.14	370
SL1- 20080401Mega pode-LLJmal	LLJ	F	Narrative	2008- 04-01	1.49	34
SL1- 20080404Fishin g-BPatg	BPK	M	Narrative	2008- 04-04	2.32	57
SL1- 20080411Nalua ki-ETnt	ENT	F	Folktale	2008- 04-11	3.27	95
SL1- 20080424Huntin g-BTn	BLT	M	Procedural Narrative	2008- 04-24	4.57	137
SL1- 20080424MatW eaving-MRnn	MRS	F	Procedural Narrative	2008- 04-24	2.14	75
SL1- 20080424Peace Ceremony- MLnat	MSL	M	Folktale	2008- 04-24	3.23	71
SL1- 20090625Canoe Making2-TMrar	MTM	M	Description	2009- 06-25	0.43	11
SL2- 20101806Tuktu kChief-TMmnt	TMM	F	Folktale	2010- 19-06	6.14	219
SL2- 20101906MatUs es-LMnn	LTM	F	Narrative	2010- 19-06	2.23	89

¹⁴ The genre of recordings is as specified in the ELAR collection.

SL2- 20102007Ances tors-WManc	WLM	M	Historical Narrative	2010- 20-07	2.5	64
-------------------------------------	-----	---	-------------------------	----------------	-----	----

While the full set of recordings available in the archival collection is larger than this sample, a subset was identified as suitable for this study. Only spontaneous speech data was used, and only for narratives, which are the majority of the natural speech data in the corpus. A wide range of corpora and corpus-based studies exclusively based on natural speech data in Oceanic languages are available (e.g. Aznar & Gala, 2020; Billington et al., 2021; Davey, 2021; Duhamel, 2019; Schnell & Barth, 2020), so it is advisable for this study to also concentrate on spontaneous speech rather than elicited speech to allow for potential comparative corpus-based studies between Lelepa and other Oceanic languages in the future. Moreover, within natural speech data, the usage preference for free pronouns still varies to some extent between speakers. Since an investigation of the usage frequency of free pronouns across speakers at the individual level is beyond the scope of this thesis, the recordings where the speakers produce no free subject pronoun at all are not appropriate for this study, hence they were also excluded. In addition, only recordings with transcription and translation in the original files were chosen, in order to avoid misinterpretation of the texts, since the original transcription and translation were done with the assistance of Lelepa speakers (Lacrampe, 2014, p. 11).

3.3 Data re-annotation

Because the original glossing in the transcription (see § 2.4 above for examples) was not specifically intended for this corpus-based study on person references, subjects in the recordings were re-annotated following the GRAID annotation

schema (Haig & Schnell, 2014) to code additional information. GRAID annotations, in full *Grammatical Relations and Animacy in Discourse*, though initially designed for cross-corpus study in linguistic typology, are also helpful in various quantitative analyses of natural discourse. Compared with traditional glossing schema, GRAID annotations not only register morphological features such as forms, functions, and parts of speech, but also syntactic functions like grammatical relations, prosodic realisations like <w> for phonetically lighter forms, and extralinguistic factors like animacy or humanness, which turns out to be an important predictor of the phenomenon on which this study focuses (see § 4.2.3 and § 4.4.5). Hence, GRAID offers a consistent way of tokenising the texts with a larger amount of discourse information. Moreover, GRAID has been proved successful in previous studies on Vanuatu languages (e.g. Schnell, 2018a) and referential choice (e.g. Vollmer, 2023). During the re-annotation process in this study, an additional tier of GRAID annotations was added to the .eaf file in ELAN (ELAN, 2023), as exemplified in Figure 3-1 below.



Figure 3-1: A screenshot of re-annotation using GRAID in ELAN with the additional annotation tier named *GRAID@TSN*

However, the re-annotation focuses only on the conventions relevant for subjects, given that grammatical relations other than subjects and agents are beyond the scope of this thesis. The following table lists all the GRAID symbols used in this study, adapted from Haig & Schnell (2014, pp. 54–55):

Table 3-2: GRAID symbols used in this thesis

Symbol	Meaning
0	'zero': argument position not filled by an overt referring expression
1	argument with first person referent(s)
2	argument with second person referent(s)
A	transitive subject
cc	complement clause
Ds	direct speech
Dt	dislocated topic
H	NP has human referent(s), or refers to anthropomorphised referents
Ncs	non-canonical subject: An argument that lacks some or all of the morphological properties associated with subjects in the language, but commands most of the syntactic properties associated with subjects in the language concerned
Np	lexical NP
Pro	free pronoun in its full form (in contrast to <-pro> or <=pro>)
=pro	'weak' clitic pronoun
-pro	pronominal affix
Rc	relative clause
s	intransitive subject
W	'weak': Indicates phonologically lighter form of a particular element (e.g. pronoun) that may, under certain conditions, be realized as clitic. Simply precedes regular gloss, e.g. <wpro>
##	boundary of independent clause, inserted at left edge
#	boundary of dependent clause, inserted at left edge, further specified
%	end of a dependent clause (if not coinciding with the end of its main clause)

3.4 Token extraction and coding

During the token extraction process, 1616 eligible tokens were identified¹⁵, including not only subject referents but also expletive subjects (see also § 3.5). Essentially, every verb, copula, (Lacrampe, 2014, p. 136) or numeral (Lacrampe, 2014, pp. 171–172) in Lelepa is expected to take a syntactic subject, except for prepositional verbs, which have lost most of their verb properties and cannot take any subject proclitics (Lacrampe, 2014, p. 167). Since only eight verbs related to direction or manner can be used as prepositional verbs and they behave very differently from predicate verbs in that they do not take any TAM particles, auxiliaries, or negation markers (Lacrampe, 2014, pp. 165, 167–169), instances of prepositional verbs were easily identified and excluded from the dataset.

Then, in addition to person, number, animacy, and humanness¹⁶ tagged during the re-annotation process, each token was coded for the presence and type of NP and the presence of subject proclitics, to inform the choice of dependent variables in each model (see § 3.6). The presence of NP and subject proclitics was assigned binary Y/N¹⁷ values. Types of NP include null NP, lexical NP, free pronoun, demonstrative pronoun, both lexical NP and free pronoun. In addition, according to Lacrampe (2014, pp. 213–214), the givenness of a subject referent

¹⁵ See also Table 3-1 for the number of tokens in each recording

¹⁶ According to the GRAID annotation schema, humanised animals and spirits in folktales that talk and behave like human beings are also tagged *human* or in fact *human-like* referents in this category, and Lacrampe (2014, p. 145) also mentions human-like or anthropomorphised referents are treated like real human referents.

¹⁷ Y stands for yes/present and N stands for no/absent, henceforth.

relates to the presence of an overt NP, including free subject pronouns, hence, tokens were coded for whether they are continued, re-introduced, or new subjects. Whether the token is anaphoric, deictic, or expletive is also coded as an initial variable. Compared with anaphoric subjects, expletive subjects¹⁸ tend to have null NP references in pro-drop languages (e.g. LaFond, 2001), while deictic subjects prefer overt NP references (e.g. Ivanova-Sullivan, 2014).

3.5 Sample selection

3.5.1 Overview

In order to address the research questions for this study, two subsets of the data were compiled, one relating to free subject pronouns and one relating to subject proclitics, following consideration of additional criteria for inclusion. These two subsets of data were drawn on for four quantitative analyses, discussed in § 3.6.

3.5.2 Free subject pronouns

To model the presence of free subject pronouns, a subset of data comprising 1440 tokens, where free subject pronouns are either present or absent, was selected for analyses of free subject pronoun variations. Referents with free pronoun references in the subject NP (preverbal) or left-dislocated slots were extracted from an automatic corpus search. However, not every pre-clausal NP is a dislocated subject, and special criteria are necessary to distinguish other left-dislocated NPs functioning as objects, obliques, or possessors. The most

¹⁸ See (9) and (10) in § 3.5.

distinctive feature of a left-dislocated subject is that it is coreferential with the subject proclitic on the verb, whereas dislocated objects are reflected on the post-verbal object markers. The oblique enclitics at the end of the verb complex agree with dislocated obliques. In addition, although not overtly expressed in the verb complex, the left-dislocated possessor always leave a pronominal trace on the possessed NP within the basic clause (Lacrampe, 2014, p. 273), in the form of either a bound form or a free word¹⁹. Examples (5) to (8) from the reference grammar (Lacrampe, 2014) show the left-dislocated subject, object, oblique, and possessor in Lelepa respectively, following the Leipzig glossing rules (Comrie et al., 2015), where the dislocated NPs and their anaphors are in bold:

- (5) *konou, a=ga ras nanu*
 1sg 1sg.NOM=IRR gather coconut
 ‘As for me, I’ll gather the coconuts.’ (Lacrampe, 2014, p. 272)

- (6) *fakna*, *mamei* *nae*
 Fakna father 3sg.POSS
e=rki=nia=s [lag...]
 3sg.NOM=tell=**3sg.ACC**=3OBL [that...]
 ‘As for Fakna, her father told her [that...]’ (Lacrampe, 2014, p. 272)

¹⁹ The possessor of an inalienable possessive construction may be identical to its free form, as in example (8), for more details on possessions, see Lacrampe (2009, 2014, pp. 217–232).

- (7) *lolaapa, patriki e=ti*
Lelepa mosquito 3sg.NOM=NEG
laapa=s mau
 be.many=**3OBL** NEG

‘In Lelepa, there are not many mosquitoes’ (Lacrampe, 2014, p. 273)

- (8) *konou, nagi konou thompson*
1sg name **1sg** Thompson

‘As for me, my name (is) Thompson.’ (Lacrampe, 2014, p. 495)

Nonetheless, instances where free subject pronouns are absent need to be manually selected within subjects which are not coreferential with any overt free pronoun. There are two types of clauses where a subject pronoun cannot fill the subject NP position, hence they do not qualify as having an ‘absent’ free subject pronoun since free pronouns are categorically impossible to be ‘present’ in those cases, hence excluded from the subset of data. The first is when the subject referent is referenced by a demonstrative pronoun such as *tena*, which is in complementary distribution with a free pronoun in the subject position of a clause in the dataset. The other case is in existential or time-expression sentences, where the copulas are marked with non-referential or expletive bound pronouns and the subject NP slots are categorically empty, as exemplified in the following sentences:

- (9) *e=pitlaka naa brij na e=wane sua*
3sg.NOM=have HESIT bridge REL 3sg.NOM=lie PRF

‘There was a bridge that was there.’ (Lacrampe, 2014, p. 495)

- (10) *e=pi* *eit* *oklok*
 3sg.NOM=COP eight o'clock
 'It was eight o'clock.' (Lacrampe, 2014, p. 500)

3.5.3 Subject proclitics

As for subject proclitic variations, a 1574-token subset of data was selected. Instances where no subject proclitic was present in a clause were classified as subject referents with omitted subject proclitics²⁰. 42 tokens were excluded where subject proclitics were either categorically absent or unclear.

As Lacrampe (2014, pp. 241–242) points out, when the subject NP of a clause is headed by *nagi* 'name', both the copula and the subject proclitic are optionally deleted, as shown in (8) above, hence referents in in copula omission constructions were excluded because subject proclitics are attested systematically absent in that case. Nonetheless, the copula is also found occasionally present without a subject proclitic, as in (11) below, which Lacrampe (2014, pp. 241–242) terms *omitted agreement*. Since Study 4 focuses on the variability in the deletion of subject proclitics, only instances of copula omission rather than omitted agreement were excluded from the subset of data.

²⁰ Although subject proclitics are not pronounced in this case, they are still shown in most transcriptions, suggesting that Lelepa speakers assisting with transcription may view constructions are more grammatical/more typical with proclitics than without them.

(11)	<i>konou,</i>	<i>nagi</i>	<i>konou</i>	<i>pi</i>	<i>nagi</i>
	1sg	name	1sg	COP	Nagi

‘Me, my name is Nagi.’ (Lacrampe, 2014, p. 241)

There are also instances where syllable boundaries are unclear, for example, when a third-person singular subject proclitic *e=* directly follows a lexical NP specified by the determiner *nge* at its right edge, sometimes the vowel in *nge e=* is realised as merely a short [e] rather than the expected long [ee], and it is difficult to distinguish whether the word-final vowel or the subject proclitic is deleted in this case. Therefore, these unclear tokens were also excluded.

3.6 Statistical analyses

3.6.1 Overview

In terms of statistical methods, although Bayesian inference has been widely accepted as an alternative to traditional frequentist statistical methods in recent studies (e.g. Davey, 2021; Flego & Forrest, 2021; Levshina, 2016, 2022; MacKenzie, 2020; Nalborczyk et al., 2019; Nicenboim & Vasishth, 2016; Vasishth et al., 2013), this thesis still concentrates on frequentist approaches, since this is the first systematic investigation of free pronouns and subject proclitics in Lelepa, and the most attractive advantage of Bayesian methods, i.e. they enable researchers to draw on the findings of previous studies as priors (Levshina, 2022), is not applicable in this context.

This thesis comprises four different inferential statistical models with varying sample size and number of variables: Study 1 – Study 3 are interrelated, aiming at the variation in the presence of free subject pronouns; Study 4 is an

independent analysis focusing on subject proclitic deletion. All the four models were implemented using different packages in R (R Core Team, 2023). Data pre-processing was done at the beginning of each study using the {tidyverse} package (Wickham, 2023) so that the statistical models could process the coded variables.

3.6.2 Study 1

Study 1 applies a generalised linear mixed model (GLMM) with logit link function to assess the effects of various potential factors on the presence of free subject pronouns using the `glmer()` function in {lme4} (Bates et al., 2023). Whether a free subject pronominal is present was coded as the dependent variable for all tokens. As for independent variables, in addition to the variables tagged in § 3.4 above, whether a subject referent is in direct reported speech was coded as a potential predictor as well, since a preliminary impression of the corpus data suggested that free subject pronouns were used more extensively in quoted speech. The referent's definiteness and specificity were also coded as initial variables, considering the importance of identifiability in referential choice (Gundel et al., 1993). All deictic referents are inherently specific and definite, whereas when an anaphoric referent is introduced to the discourse for the first time, according to Lacrampe (2014), it is:

- Generic (neither specific nor definite) if not modified with any determiner,
- Specific (and indefinite) if modified with an indefinite determiner *skei*,
- Definite (and specific) if modified with a definite determiner *nge* or a demonstrative determiner such as *na*.

After this, a forward selection was performed to decide on the inputs represented in the final model. Forward selection started with a null model and then added individual potential variables into the model, and variables failing to contribute significantly to the model's predictive power or resulting in collinearity were eliminated (Walczak & Massart, 2000). Based on the results of forward selection, the referents' definiteness and specificity turned out to negatively impact the model's prediction, and were therefore excluded. In addition, since the substantial majority of animate referents in this dataset are human-like, and very few animate referents are non-human, animacy and humanness are highly correlated. Consequently, only humanness is coded as an input variable since human referents tend to have free pronoun references based on the reference grammar (Lacrampe, 2014, pp. 145, 214, 190, 244).

After forward selection, five independent variables with binary categorical values were coded as fixed effects in the final model, including whether there is a shift from the previous referent, whether it is the first mention of the referent, whether the referent is in direct reported speech, whether the referent is deictic or anaphoric, and the referent's humanness. In addition, individual speaker was coded as a random effect. Overall, variables are shown as follows, with the reference levels in bold:

- Dependent variable
 - Presence of a free subject pronoun
 - Y, N
- Fixed effects
 - Switch in referent
 - Y, N
 - First occurrence in speech
 - Y, N
 - Humanness
 - Y, N
 - Direct speech
 - Y, N
 - Referent type
 - **Anaphoric**, deictic
- Random effect
 - Speaker

However, since the distribution of free subject pronouns is highly imbalanced (see also § 4.2.1 and § 5.2.1), the model's prediction would have been very likely to bias towards the majority group, i.e. the absence of a free pronoun. Therefore, the data was undersampled in this study to artificially increase the relative frequency of the minority group. Two undersampling methods were performed for comparison, random undersampling and principled undersampling. The random undersampling was automatically implemented using the {caret} package (Kuhn, 2023), wherein the size of the majority group was pruned to only three times that of the minority group. An iteration was

performed to find the random seed from 1 to 1000 with the highest balanced accuracy in prediction, which was then set as the seed in the undersampling process for reproducibility. Additionally, a principled undersampling was manually performed according to Torres Cacoullos & Travis (2019), where only subject referents with free pronoun references and the immediately preceding and following subject referents produced by the same speaker were included in the undersampled dataset. As a result, the random undersampled dataset comprised 396 tokens, and the principled undersampled dataset consisted of 258 tokens.

Then, a series of tests were also executed to assess the adequacy of the models trained with undersampled data. Both models were attested to be significantly better than a null model with only the random effect at explaining the data based on the Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), likelihood-based measures, and a Chi-squared test²¹ through an anova comparison (Hastie & Pregibon, 1992) with the built-in `anova()` function in R (R Core Team, 2023). The lower BIC values also indicate that the current model, trained with undersampled data, has been successful in preventing overfitting (Coupé, 2018). Furthermore, the results of overdispersion and goodness-of-fit tests using the {DHARMA} package (Hartig, 2022) show that both datasets followed a uniform distribution of residuals²², the models'

²¹ The results were $p < 2.2\text{e-}16$ for the random model ('the random' henceforth), and $p < 1.879\text{e-}11$ for the principled.

²² $p = 0.8096$ for the random, and $p = 0.9875$ for the principled.

dispersion was consistent with simulated residuals²³, and no significant outliers were detected²⁴.

While previous quantitative studies of linguistic variation (e.g. Davey, 2021) have demonstrated the effectiveness of these two tests in assessing the model adequacy, due to the random nature of the undersampling process, an additional leave-one-out cross-validation (LOOCV) using the {caret} package (Kuhn, 2023) was implemented to test the performance of the *random* model in predicting unseen data. The model reached a satisfactory overall accuracy of 77.53% in the LOOCV. Furthermore, the overall balanced accuracy was 64.48%, which is acceptable as for unseen data. Although balanced accuracy is not inherently applicable in LOOCV, it could still be computed afterwards using the formula in (12) below (*cf.* Paulo et al., 2021 *inter alia*):

(12)

$$\frac{\frac{TP}{PP} + \frac{TN}{PN}}{2}^{25}$$

In addition, the *random* model achieved a balanced accuracy of 72.73% when predicting the undersampled 396-token training dataset and 71.20% for the full 1440-token dataset, both of which are adequate. The marginal difference between the balanced accuracies also suggests that the undersampling process has not significantly compromised the model's predictive performance

²³ $p = 0.94$ for the random, and $p = 0.816$ for the principled.

²⁴ $p = 1$ for both with an estimated outlier frequency of 0.

²⁵ TD stands for true positives, PD stands for predicted positives, TN stands for true negatives, and PN stands for predicted negatives.

when applied to the full dataset, which could have been a potential risk of undersampling (Fernandez et al., 2018).

However, for better predictive performance of the GLMM, the referents' person and number were excluded from the input variables in this study, due to the multitude of levels within this category, coupled with a pronounced imbalance in their distribution. For example, only six second-person plural subject referents were identified in the 1440-token dataset.

3.6.3 Study 2

Study 2 involves a recursive partitioning analysis for a conditional classification tree with binary splits using `{party}` (Hothorn et al., 2023), to examine the effect of person and number on the presence of free subject pronouns in Lelepa. A conditional classification tree, which separates categorical variables into groups of levels in each split, is more appropriate for a predictor with multiple and imbalanced levels, therefore adopted in this study. The entire 1440-token dataset was used in training the classification tree. The output variable was still the presence of a free subject pronoun. Six input variables were coded in the model: *switch in referent*, *first occurrence in speech*, *humanness*, *direct speech*, *referent type*, and *person and number*. The effect of speaker was not considered in Study 2 since the model did not support the coding of a random variable.

3.6.4 Study 3

While Study 1 and Study 2 both focus on the presence versus absence of free subject pronouns, Study 3 turns to a comparison between the presence of free pronouns, lexical NPs, and null NPs in subject position, aiming to further explore

whether the variables are related to the overall presence of a subject NP or exclusively the presence of free subject pronouns. A multinomial baseline-category logit (mblogit) model was implemented in this study using the mblogit() function in {mclogit} (Elff, 2022). 1422 tokens were presented in the study, as 18 tokens coreferential with both a free pronoun and a lexical NP were excluded to avoid complexity. The dependent variable in the model was the type of subject NP, with three levels: *pronominal/lexical/null*. Since most subject referents did not have an overt NP reference, null NP was chosen as the reference level in this study. Four independent variables were coded as fixed effects: switch in referent, first occurrence in speech, humanness, and direct speech. Referent type was coded as an input variable in Study 1 but was excluded during forward selection in Study 3 since no deictic subject referent was referred to with a lexical NP (see also § 4.2.4). Speaker was coded as a random effect. In sum, the variables in Study 3 are shown below, where reference levels are in bold:

- Dependent variable
 - NP type
 - Lexical, Pronominal, **Null**
- Fixed effects
 - Switch in referent
 - Y, **N**
 - First occurrence in speech
 - Y, **N**
 - Humanness
 - Y, **N**
 - Direct speech
 - Y, **N**
- Random effect
 - Speaker

In terms of model adequacy, the model's prediction achieved a 76.65% accuracy, and its reliability is attested to be higher than a null model according to lower AIC and BIC values from similar tests as presented in § 3.6.2 above.

3.6.5 Study 4

Study 4 focuses on the variability in Lelepa subject proclitic deletion. Since phonotactics can be used as a clue to distinguish bound pronouns and agreement affixes (e.g. Donohue & Sawaki, 2007), an additional set of phonotactic attributes were coded as potential predictors, including the subject

proclitics' syllable structure, preceding phoneme, following environment²⁶, and whether vowel reduction takes place in the preceding and following words. The mood of a clause also associates with subject proclitics in Lelepa (see also § 2.4), therefore, each token was coded for mood to investigate the effects of markedness (realis vs irrealis) and fusionality (2sg irrealis vs other irrealis) in subject proclitic deletion. Finally, since no evidence shows the effect of whether a referent is deictic or anaphoric on the presence of subject proclitics, based on both preliminary observation of Lelepa data and existing studies, the *referentiality* (with binary values Y/N) instead of *subject type* (with ternary values *deictic/anaphoric/expletive or non-referential*) of tokens was coded as an initial predictor in this study.

Study 4 consists of three stages in chronological order. The first part of this study was a conditional random forest analysis with all the initial inputs and the presence of a subject proclitic as an output using `cforest()` in `{party}` (Hothorn et al., 2023). The permutation importance scores of the predictors were then calculated and ranked using `{permimp}` (Debeer et al., 2021) to decide on the predictors in the next model.

Input variables with a permutation importance score over 0.1 were then incorporated in a recursive partitioning analysis for a conditional classification tree with binary splits. In brief, five predictors were presented in the recursive partitioning analysis, including the subject proclitics' syllable structure, preceding phoneme, mood, person and number, and coreferential NP type (see

²⁶ This in fact indicates the following *two* phonemes. *Following phoneme* is not coded because all subject proclitics, both present and deleted, are found preceding a consonant in the dataset.

also § 4.5.1). For better interpretability of the results, only splits with significance values less than 0.001 are generated in the conditional classification tree.

However, a classification tree only allowed for an exploratory look into the variability in subject proclitic deletion, and predictors leading to splits with lower significance were excluded in the results. Furthermore, like Study 2, speaker was not able to be coded as a random effect in the first two models. Hence, additional hypotheses informed by the results and supported with descriptive statistical data (see also § 4.5) were tested with a GLMM using {lme4} (Bates et al., 2023) at the end of this study. The output variable was still the presence of a subject proclitic. As for input variables, four binary variables based on the hypotheses were selected as fixed effects, including mood, the presence of a lexical NP, whether the subject proclitic is singular or non-singular, and whether the proclitic is second-person singular (see also § 4.5.6). Speaker was coded a random effect.

Chapter 4 Results

4.1 Overview

This chapter presents and interprets the results of the statistical analyses. Sections 4.2 – 4.4 incorporate the analyses of free pronouns, including § 4.2 on the GLMM (Study 1), § 4.3 on the recursive partitioning (Study 2), and § 4.4 on the multinomial baseline-category logit (mblogit) model (Study 3). Section 4.5 moves on to the analyses of subject proclitic variations (Study 4). Each subsection in Sections 4.2 – 4.5 comprises an overview of the results, followed by a brief discussion of each significant variable with tables and plots showing the distribution. The plots are generated with the `{ggplot2}` package in R (Wickham et al., 2023). Section 4.6 summarises the findings as to what factors affect the distribution of various subject expressions. Although the GRAID annotation schema was used in the re-annotation process for analytical purposes (see § 3.3), the example sentences in this chapter still follow the Leipzig glossing conventions for convenience (Comrie et al., 2015).

4.2 Study 1: GLMM analysis of free pronouns

4.2.1 Overview

Table 4-1 and Table 4-2 present the results of the GLMM analyses focusing on what relates to the presence of free pronouns. The negative estimates of the intercept in both results suggest free subject pronouns are overall uncommon (6.88%, $n = 99$) in the Lelepa dataset. The zero variance and standard deviation in Table 4-1 is possibly a feature loss due to the undersampling method, as the proportion of free pronoun presence versus absence within each speaker was

adjusted to a (quasi-)constant value, i.e. 1:2. Likewise, the non-significant effect of *direct speech* on free subject pronoun presence (see also § 4.2.5) is possibly because most referents with no free pronoun reference that were pruned during the undersampling process were not in direct speech.

Table 4-1: Results of the GLMM analysis of free subject pronoun presence with principled undersampling with significant predictors in bold

Random effect:					
Name		Variance	Std.Dev.		
Speaker (Intercept)		0	0		
Number of observations: 258, groups by speaker: 13					
Fixed effects:					
Name	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	-4.9264	1.1358	-4.337	1.44E-05	***
switch.in.referentY	1.5891	0.3383	4.697	2.64E-06	***
first.occurrence.in.speechY	0.4024	0.4371	0.921	0.35728	
referent.typeDeictic	1.1874	0.3372	3.521	0.00043	***
direct.speechY	0.1431	0.3637	0.393	0.69403	
humannessY	2.8348	1.0951	2.589	0.00964	**

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1					

In comparison, the model trained with random undersampled data in Table 4-2 displays greater variability for free subject pronoun presence in its results, hence the discussion in this subsection will mainly draw on the results of this model. The *random effects* section shows that there is moderate inter-speaker variance in the use of free subject pronouns. A negative intercept in the *fixed effects* suggest that free subject pronouns are significantly ($p < 0.001$) unlikely to be present ($n = 99$, 6.88% of the entire dataset) overall, conforming with previous impressions that free subject pronouns are highly optional in Lelepa (Lacrampe, 2014). Generally, a subject referent is significantly more likely to be

referred to with a free pronoun when it is a switch referent ($p < 0.001$), human ($p < 0.01$), deictic ($p < 0.05$), or in direct speech ($p < 0.05$).

Table 4-2: Results of the GLMM analysis of free subject pronoun presence with random undersampling with significant predictors in bold

Random effect:					
Name		Variance	Std.Dev.		
Speaker (Intercept)		0.5134	0.7165		
Number of observations: 396, groups by speaker: 13					
Fixed effects:					
Name	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	-5.2611	0.9175	-5.734	9.81E-09	***
switch.in.referentY	1.7939	0.3173	5.653	1.57E-08	***
first.occurrence.in.speechY	0.7101	0.4458	1.593	0.11118	
referent.typeDeictic	0.8487	0.3358	2.527	0.01149	*
direct.speechY	0.979	0.3832	2.555	0.01063	*
humannessY	2.4607	0.851	2.892	0.00383	**

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1					

4.2.2 Switch in referent

Table 4-3: Free subject pronoun presence based on switch in referent

Presence of free subject pronouns (FSP)						
		No (N)		Yes (Y)		Total
Switch in referent	N	744	97.38%	20	2.62%	764
	Y	597	88.31%	79	11.69%	676
	Total	1341	93.13%	99	6.88%	1440

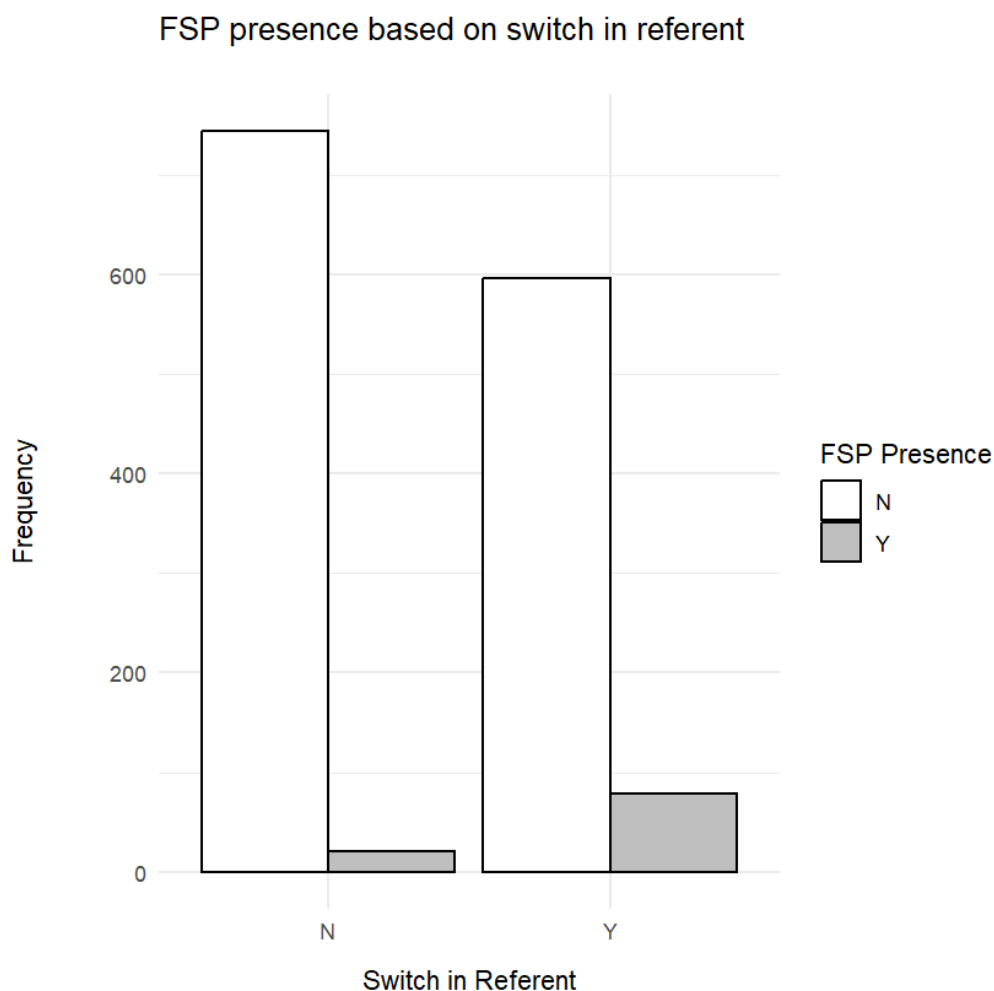


Figure 4-1: Free subject pronoun presence based on switch in referent

As illustrated in Table 4-3 and Figure 4-1, switch subject referents take up a comparable proportion to non-switch subject referents in the data, and 11.69% of them are referred to with free pronouns, which is approximately 4.46 times higher than for non-switch referents. This observation supports Lacrampe's (2014, pp. 269–272) description that an NP, including a free personal pronoun, is overtly expressed when the referent is re-introduced (see also § 2.4), as manifested in examples (13), (14), and (15) below. Along with the occurrence of a free subject pronoun in a clause featuring a switch subject referent, the corresponding English translations in (13) and (14) are also distinguished by

the use of *as for* constructions (Lacrampe, 2014, p. 269), where the subject changes from 2sg to 1sg in the final clause and a free pronoun *konou* is used to indicate contrast instead of simply establishing a topic (see also § 5.2.2). However, the two examples with free subject pronouns as contrastive topics are both in clause-initial position, hence they are realised in the same way as switch topics in (15).

- (13) *ae, nag ku=tae a-sei lag*
 hey 2sg 2sg.NOM=know LOC-who COMP
ḡa=fa-ki wara-e pa? konou
 2sg.IRR=go.IRR-TR there-ADD GO 1sg
a=ti pa-ki wara-e mau
 1sg.NOM=NEG go.R-TR there-ADD NEG
 ‘Hey, do you know where you’ll go? **As for me**, I didn’t go there.’
 (SL1-20060402TomsenFrance-TNIs_016)

- (14) *nag ḡa=fsru-k bislama,*
 2sg 2sg.IRR.NOM=speak.IRR-TR Bislama
konou ga fsru-k franis
 1sg IRR speak.IRR-TR French
 ‘As for you, you will speak Bislama, and **as for me**, I will speak French.’
 (SL1-20060402TomsenFrance-TNIs_044)

- (15) *go naar ur=to pea,*
 and **3pl** 3pl.NOM=IPFV be.first.R
se konou a=rak go
 while **1sg** 1sg.NOM=follow and
a=ta gor naa
 1sg.NOM=cut block HESIT
 ‘And they were first, while I would follow, I would cut and block hum...’
 (SL1-20080424Hunting-BTn_010)

Nonetheless, non-switch referents can also be referenced with free subject pronouns, although all the 16 instances (only 11.11% of the data) of continued referents with a sole²⁷ free subject pronoun reference are all *human(-like)* referents, and mostly *deictic* and/or in *direct speech* (see also § 4.2.3, § 4.2.4, and § 4.2.5), as exemplified in example (16) below:

- (16) *ṗa=faam ntai konou,*
2sg.IRR.NOM=eat.IRR excrement 1sg
nag ṗa=ti paam ntai konou,
2sg 2sg.IRR.NOM=NEG eat.R excrement 1sg
mes ga²⁸ faam-ko nuk to
 today IRR eat.IRR-2sg.ACC as.is STAT
 ‘Eat my shit, if you don’t eat my shit today, I’ll eat you as you are.’
 (SL1-20080401DevilArtok-CMmut_110)

The correlation between switch referents and the occurrence of free subject

²⁷ i.e. not coreferential with both a free pronoun and a lexical NP.

²⁸ This is an instance of subject proclitic omission, which is to be discussed in detail in § 4.5.

pronouns is significant. However, as a special subtype of referent switch, whether a subject referent is in its first occurrence in speech does not have a significant effect on the presence of free subject pronouns. This observation differs somewhat from Lacrampe's (2014, pp. 210, 270) description, that subject referents are referenced with an NP when they are either re-introduced or introduced to the discourse for the first time. However, this could be due to the fact that a new referent can be referred to with either a lexical NP or a pronominal NP, which is to be discussed in more detail in § 4.4 below.

4.2.3 Humanness

Table 4-4: Free subject pronoun presence according to humanness

		Presence of FSP				
		N		Y		Total
Humanness	N	177	99.44%	1	0.56%	178
	Y	1164	92.23%	98	7.77%	1262
	Total	1341	93.13%	99	6.88%	1440

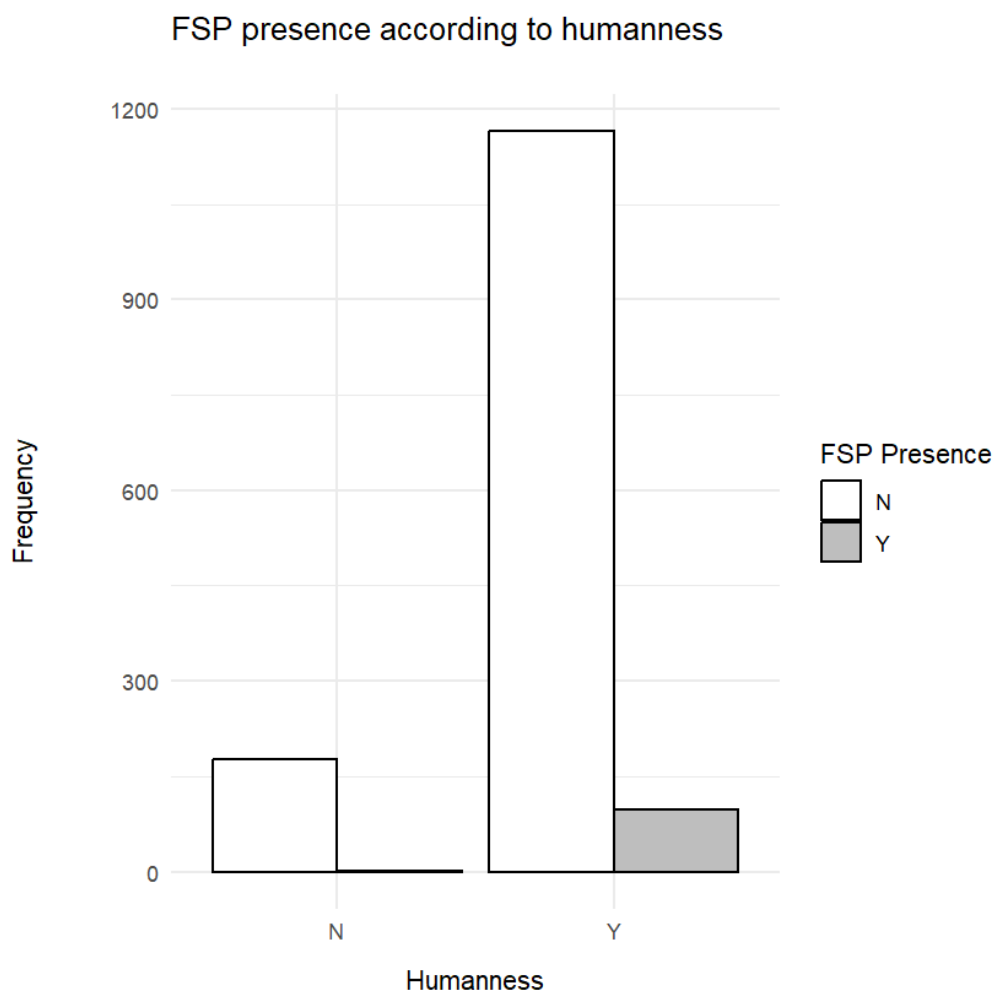


Figure 4-2: Free subject pronoun presence according to humanness

It has been observed in multiple studies that human or animate referents are more frequently mentioned and more likely to be clausal subjects (e.g. H. H. Clark & Begun, 1971; Givón, 1983b; Itagaki & Prideaux, 1985; Pearson et al., 2001), and similarly most referents in this corpus are also human or human-like. However, Table 4-4 and Figure 4-2 show that despite the many more instances of human(-like) referents overall (around 87.64% of the 1440 tokens), free subject pronouns are still significantly likely to associate with human(-like) referents, as in example (17) below, where the 3sg referent is coreferential with both the free pronoun *nae* and the bound pronoun *e=*:

- (17) *nae e=pkate marou pan pa,*
3sg 3sg.NOM=too.much be.thirsty GO GO
naa e=kat mat wane naa napua.
 HESIT 3sg.NOM=CERT die lie HESIT road
 ‘He was too thirsty, (and)...he died on...the road.’
 (SL1-20080411Naluaki-ETnt_ETnt.017)

Although unbalanced, this distribution accords with Lacrampe’s (2014, p. 214) observation, that free subject pronouns are present mostly when the referent is human. Furthermore, the only non-human subject referent found referred to with a free subject pronoun in the dataset is coreferential with both a pronominal NP and a lexical NP, and the free subject pronoun may be a left-dislocated topic characterised by the rising pitch contour with a short pause after (Lacrampe, 2014, p. 271), as in example (18) below:

- (18) *nae, na-lega-n nagna,*
3sg NMLZ-sing-NMLZ 3sg.ASS.POSS
e=pan tapla
3sg.NOM=go.R like.this
 ‘It, its song, goes like this.’ (SL1-20070420HarrierFalcon-GMmp_031)

4.2.4 Referent type

Table 4-5: Free subject pronoun presence based on referent type

		Presence of FSP				
		N		Y		Total
Referent type	Anaphoric	912	96.10%	37	3.90%	949
	Deictic	429	87.37%	62	12.63%	491
	Total	1341	93.13%	99	6.88%	1440

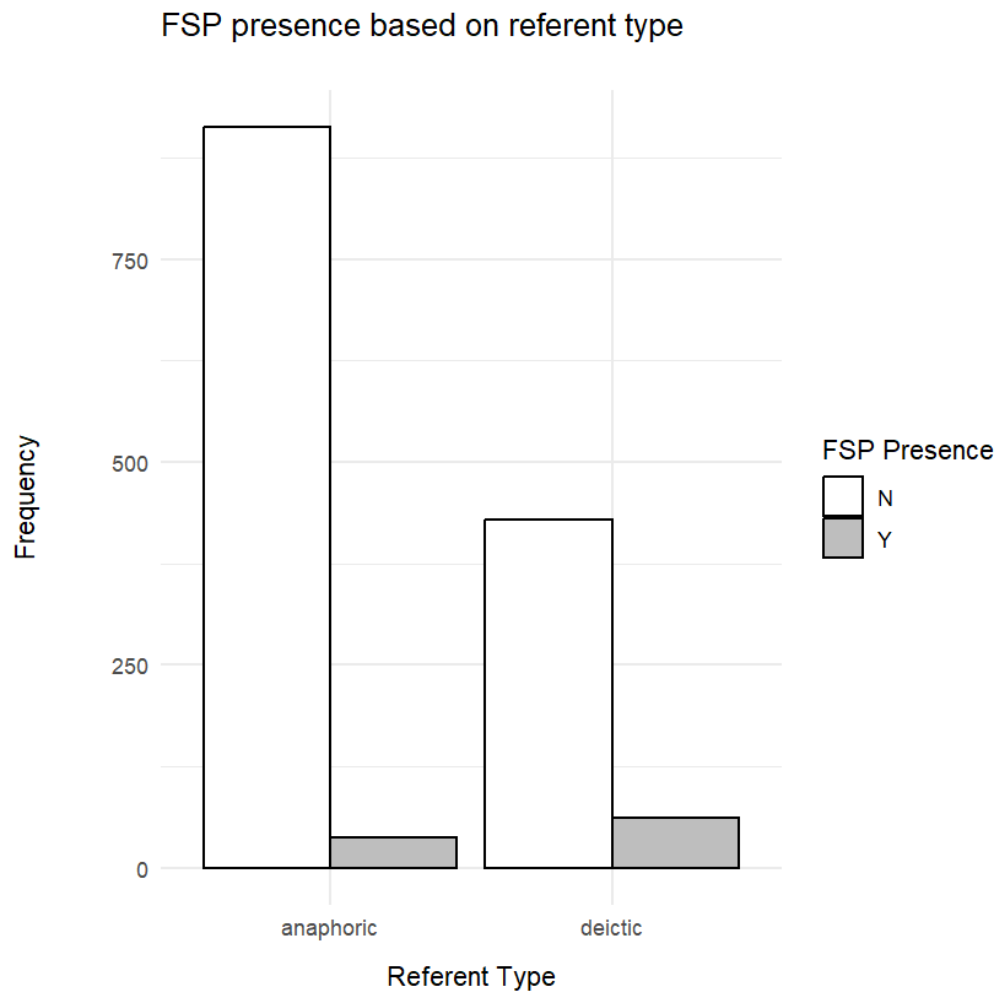


Figure 4-3: Free subject pronoun presence according to referent type

Unlike switch in referent where the two categories are almost evenly distributed, anaphoric subject referents are 1.93 times more frequently found than deictic referents in the dataset. This might nonetheless be explained by the large proportion of traditional narratives and folk stories, in which most referents are anaphoric. As shown in Table 4-5 and Figure 4-3, deictic subject referents exhibit a significant correlation with the occurrence of free pronouns, as 12.63% of deictic referents are coreferential with free subject pronouns, 3.24 times higher than anaphoric ones. Example (19) below is an instance of the speaker questioning another person beside him, where the deictic 2sg referent is

coreferential with a proclitic *ku=*, a lexical NP (the referent's name) and a free pronoun *nag*:

- (19) *nafet, nag ku=ti maroa-k*
Naviti 2sg 2sg.R.NOM=NEG think-TR
lag p̃a=skar=i mou
COMP 2sg.IRR.NOM=join=3sg.ACC NEG
 'Naviti, don't you think that you can complete it?'
 (SL2-20102007Ancestors-WManc_035)

A possible explanation for this effect is that deictic referents are mostly speech-act participants, which may have inherent salience in discourse and hence may be more likely to be topicalised with free subject pronouns (see also § 2.2.2 and § 5.2.1). This phenomenon might also relate to the fact that a deictic subject referent is almost exclusively referenced by a free pronoun when overtly referred to with an NP, whereas an anaphoric referent can also be referred to with a lexical NP. This does not necessarily suggest that an anaphoric referent with a lexical NP reference cannot be further coreferential with a free pronoun, nor that a deictic subject referent cannot be referenced by a free pronoun. Example (18) above comprises an anaphoric subject referent coreferential with two NPs, one lexical and one pronominal. Example (20) below is an instance where the 1sg deictic referent is coreferential with both a free subject pronoun and a lexical NP:

- (20) *konou, ritait elda poikiiki*
1sg retired elder Poikiiki
a=mro msou=n magmu til
1sg.NOM=REP want=3sg.ACC 2pl.BEN tell

<i>nalati</i>	<i>skei</i>	<i>a=pa</i>	<i>tagau</i>
day	one	1sg.NOM=go.R	fish.with.hook

'Me, retired elder Poikiiki, I want to tell you again about one day I went fishing.' (SL1-20080404Fishing-BPatg 002)

In addition, an anaphoric non-singular subject referent can also be coreferential with both a free subject pronoun and a lexical NP through the so-called *split-antecedent* anaphora (Berman & Hestvik, 1997). For instance, in example (21) below, *ar* = '1du.EXCL.NOM' is coreferential with both the independent pronoun *kenem* '1pl.EXCL' and partially the lexical NP *tat masok* 'paternal great uncle Masok', while the *1sg*, the other member of the dual referents, was introduced to the discourse before:

- | | | | | |
|------|-----------------------------|----------------|----------------------|----------------------|
| (21) | <i>kasem</i> ²⁹ | <i>naleati</i> | <i>skei,</i> | <i>kenem,</i> |
| | to | day | INDEF | 1pl.EXCL |
| | <i>tat</i> | | <i>masok,</i> | |
| | paternal.great.uncle | | Masok, | |
| | <i>ar=to</i> | <i>wok</i> | <i>wuru</i> | <i>aaa...</i> |
| | 1du.EXCL.NOM=IPFV | work | pass | HESIT |
- ‘Until one day, we – (me and) my paternal great-uncle Masok – we were working at...’ (SL1-20060402TomsenFrance-TNls 005)

However, this binary division might not be accurate enough to capture the distribution of free subject pronouns, as deictic referents are, to a large extent, limited to SAPs, as mentioned before. Hence, the results of a subsequent

²⁹ This *kasem* is a Bislama word, and the Lelepa word for ‘today’ is *mesa* (e.g. Lacrampe, 2014, p. 163).

analysis taking person (and number) into consideration will be presented in Section 4.3. The only exception takes place in one descriptive recording where the speaker is pointing at other people and referring to them using *ur*= '3pl.NOM' when describing what they are doing, as in example (22) below, and no free subject pronoun is used in this case:

- (22) *ur=to* *uḗa* *nakon=os*
 3pl.NOM=IPFV cut shape=3OBL
 ‘They are cutting its shape.’ (SL1-20090625CanoeMaking2-TMrar_002)

4.2.5 Direct speech

Table 4-6: Free subject pronoun presence based on whether in direct speech

		Presence of FSP				
		N		Y		Total
Direct speech	N	1167	94.88%	63	5.12%	1230
	Y	174	82.86%	36	17.14%	210
	Total	1341	93.13%	99	6.88%	1440

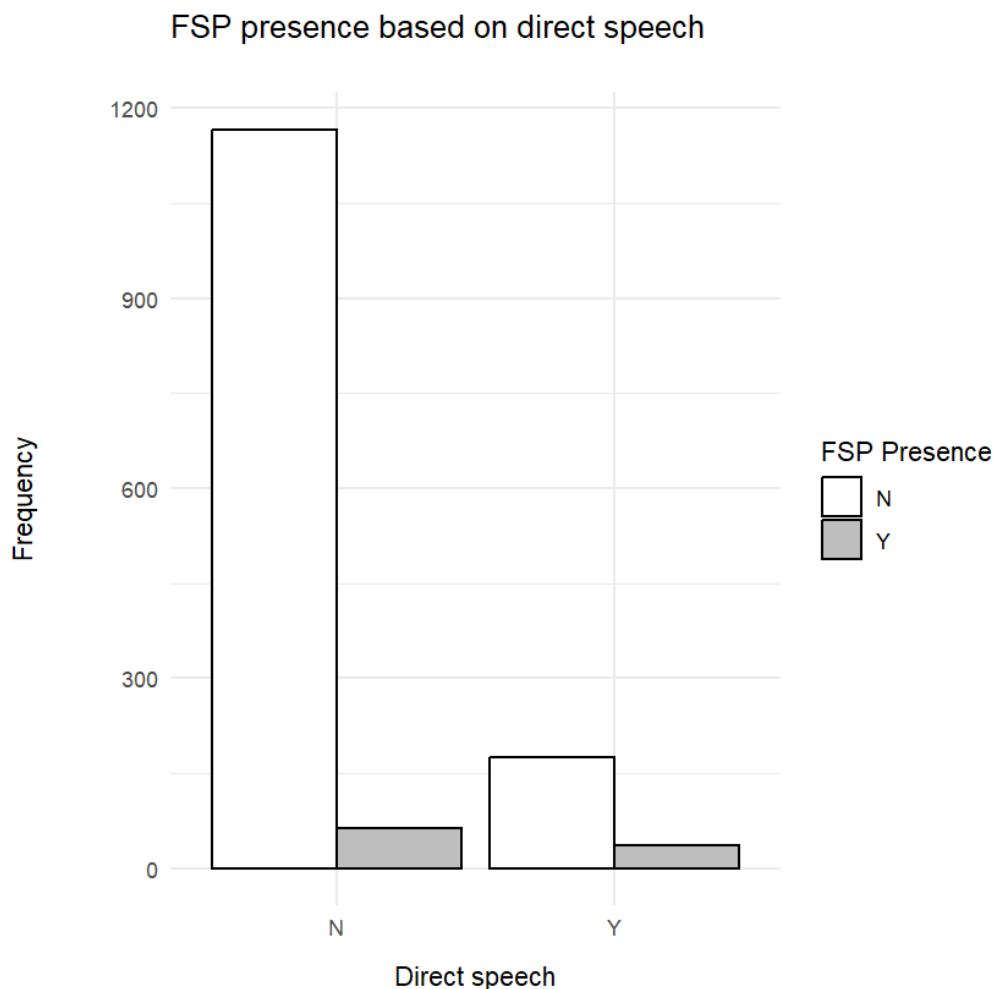


Figure 4-4: Free subject pronoun presence based on whether in direct speech

By virtue of the narratives on personal history and folklores, the dataset comprises 210 instances of direct speech, i.e. quotation, where the speaker might want to accurately reproduce the context or to tell the story in a more vivid manner. In example (23) below, the speaker uses the free pronoun *kon(ou)* ‘1sg’ twice in the direct speech, perhaps as an attention-getter (see also § 5.2.2) or to intensify his emotion:

- (23) *a=rki=nia=s* *lag,*
 1sg.NOM=tell=3sg.ACC=3OBL COMP
'kon a=mal na-fan-ona,
 1sg 1sg.NOM=not.want NMLZ-go.IRR-NMLZ
konou ga to.'
 1sg IRR STAT
 'I told him, "I don't want to go! I will stay."
 (SL1-20060402TomsenFrance-TNIs_072)

However, very few instances of indirect reported speech are found in the dataset, hence it is as yet impossible to examine the distribution of free subject pronouns in indirect reported speech and compare it with direct quotations in this study. Examples of indirect speech are all complement clauses of verbs like *rki* 'to tell', and free subject pronouns are typically absent, as shown in (24) below:

- (24) *e=rk konou=s lag*
 3sg.NOM=tell 1sg=3sg.ACC COMP
(a=)ga wus fok
(1sg.NOM=)IRR get fork
 'He (my boss) told me to get the fork.'
 (SL1-20060402TomsenFrance-TNIs_050)

As depicted in Table 4-6 and Figure 4-4, direct speech closely relates to the occurrence of a free pronoun, since more than one-third of free subject pronouns are found in direct speech, while only 14.58% of the subject referents in the entire dataset are in direct speech. For example, in example (25) below,

the addressee in the direct speech is coreferential with both a proclitic *ku=* and a free pronoun *nag*:

- (25) *tetei nae mamei nae*
 mother 3sg.POSS father 3sg.POSS
ar=rki=nia=s lag
 3du.NOM=tell=3sg.ACC=3sg.ACC COMP
 ‘*nag ku=fia rak mamei*
 2sg 2sg.R.NOM=fear towards father
naaram tetei nae [...]’
 and mother 3sg.POSS
 ‘Her mother and father told her “You’re afraid of his father and mother
 [...]”’ (SL2-20101806TuktukChief-TMmnt_022)

This could arguably be regarded as a special subtype of referent switch based on the *common ground* theory (see also § 2.2.2), and there may be a different common ground within the direct speech, as it takes place outside the current speech act. Therefore, the referent could possibly be treated as an unfamiliar referent regardless, as in example (26) below, where a free pronoun is used although there is no switch in subject referent between the reporting clause and the content of direct speech, compared with example (25) above. However, a comparison between the effects of direct speech and switch in referent on lexical NP is necessary to test the common ground hypothesis (see also Section 4.4.4).

- (26) *a=lag, 'ae, konou a=msau=na*
1sg.NOM=say hey 1sg 1sg.NOM=want=3sg.ACC
lag ga fa.'
 COMP IRR go.IRR
 'I said, "Hey, I want to go."
 (SL1-20060402TomsenFrance-TNIs_009)³⁰

Alternatively, this distribution may relate to SAPs as well, as most (85.24%) subject referents in direct speech are SAP in the dataset, and only one non-SAP referent is referenced by a free subject pronoun, which is also coreferential with a lexical NP, as in example (27) below:

- (27) *kan nlakan te=ru kik, naar*
 but because **SBST=two be.small 3pl**
ar=pat maan kik panei, [...]
3du.NOM=hit bird be.small COME
 'But because the two little ones, they shot a few birds, [...]'
 (SL1-20080401DevilArtok-CMmut_065)

Furthermore, in Schnell's (2018b) study, reference to SAP are found only in direct quoted speech between characters in a story, due to the limitation in the genres of narratives. Schnell (2018b) points out that a speaker may treat *real* interlocutors differently from interlocutors in direct speech. Since this study draws on more diverse genres of narratives and includes both real and direct-speech interlocutors, it may be able to answer whether Lelepa speakers

³⁰ However, the use of free subject pronouns in this case may also be a form of emphasis (Lacrampe, 2014, p. 146).

distinguish real and direct-speech interlocutors in the use of free subject pronouns. Figure 4-5 is the effect plot of the probability of free subject pronoun presence based on referent type and direct speech drawn using {effects} package (Fox et al., 2022), and the near parallel slopes (slope ratio ≈ 1.62 which is still close to 1) indicate that there is only a minor difference in the speakers' use of free subject pronouns based on referent type between direct speech and non-direct speech referents.

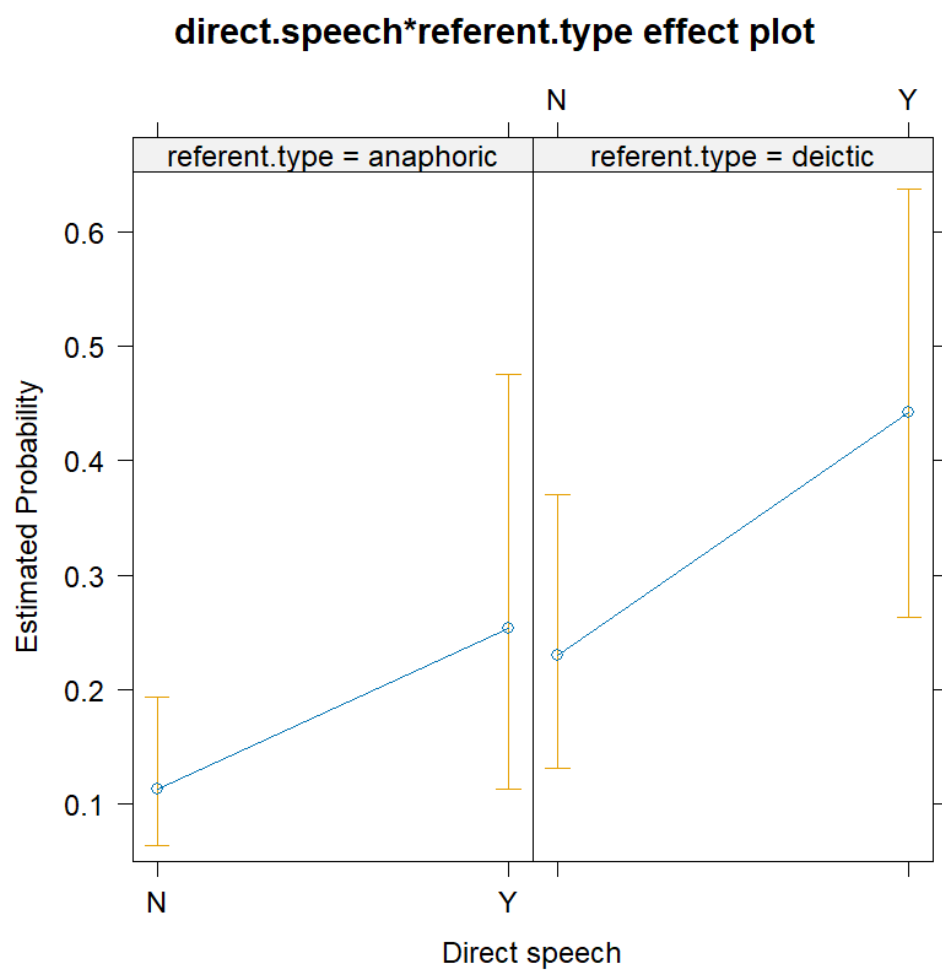


Figure 4-5: Probability of free subject pronoun presence based on referent type and direct speech

However, a more fine-grained analysis introducing person and number as an additional predictor is necessary to examine whether free subject pronouns are in fact more likely to occur when they are SAP or simply when they are deictic. Therefore, the effect of person and number will be examined and discussed in the following section.

4.3 Study 2: Recursive partitioning analysis of free pronouns

4.3.1 Overview

Figure 4-6 presents the results of the recursive partitioning analysis which includes the effect of person and number on the presence of free subject pronouns, building on the GLMM results in § 4.2. Table 4-7 and Figure 4-7 illustrate the count and percentage of free subject pronoun presence in each leaf of the classification tree. As is shown in the classification tree, two predictors out of six show significant ($p < 0.001$) effects on the presence of free subject pronouns: whether there is a *switch in referent* and a subject referent's *person and number*. The four splits demonstrate how these two inputs interact and influence the presence of free subject pronouns, and the five leaves on the bottom show the proportion of free subject pronoun presence based on the binary splits. The exclusion of humanness, referent type, and direct speech in the classification tree indicates that their correlations with free subject pronouns are of lower significance after person and number have been considered. Additionally, the exclusion of direct speech in the results also suggests that Lelepa speakers do not significantly distinguish direct-speech interlocutors with real interlocutors in the use of free pronouns, since the effect of person and

number is much more important (see also § 4.2.5). The effects of switch in referent and person and number will be discussed in the following subsections.

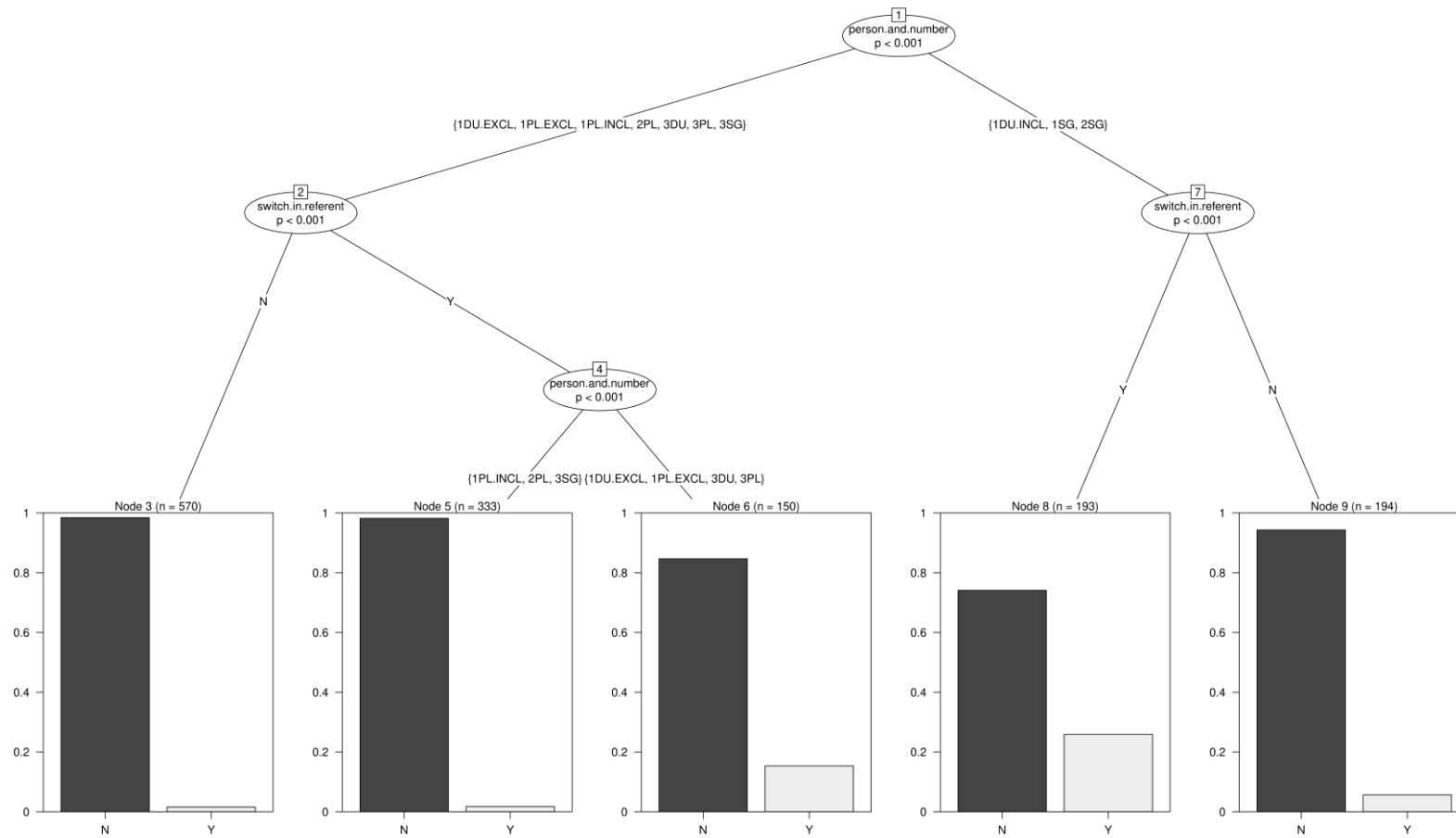
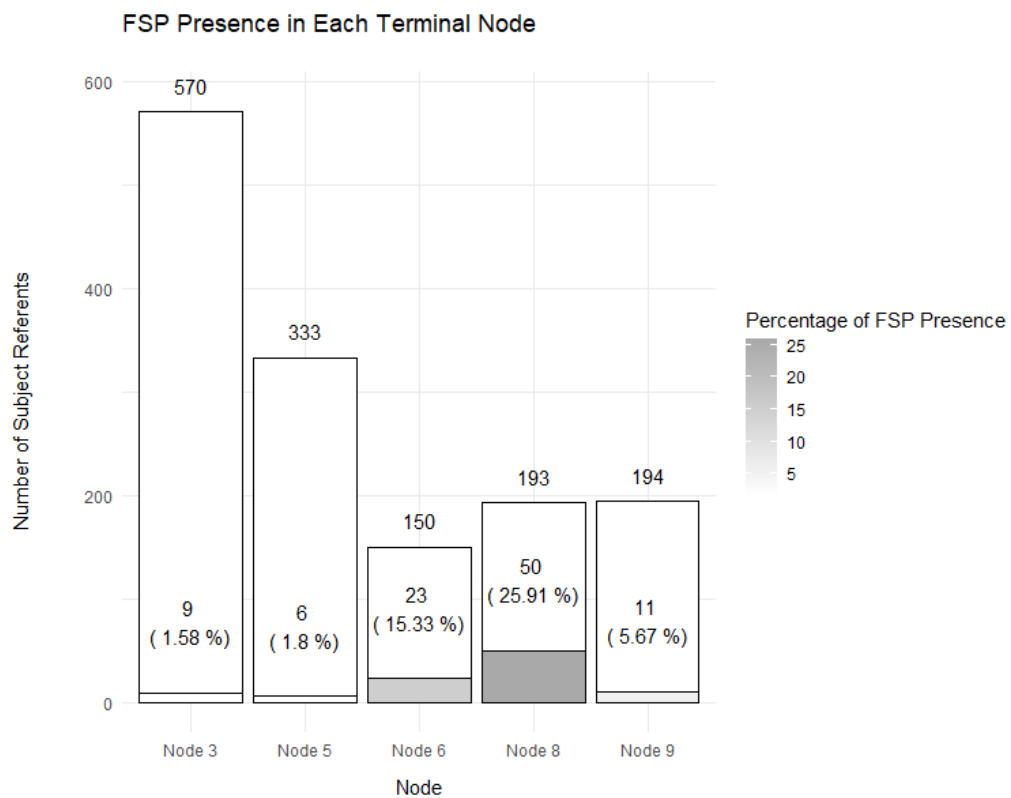


Figure 4-6: Conditional classification tree of free subject pronoun presence

Table 4-7: Free subject pronoun presence in terminal nodes

Node in the tree	Presence of FSP		
	N	Y	Total
Node 3	561	9	570
	98.42%	1.58%	-
Node 5	327	6	333
	98.20%	1.80%	-
Node 6	127	23	150
	84.67%	15.33%	-
Node 8	143	50	193
	74.09%	25.91%	-
Node 9	183	11	194
	94.33%	5.67%	-

**Figure 4-7: Free subject pronoun presence in terminal nodes**

4.3.2 Switch in referent

The significant effect of switch in referent is expected since it also shows the highest significance among the five regressors of the GLMM in § 4.2 above. At splits 2 and 7, the dataset is divided based on switch in referent. Terminal nodes 3 and 9 represent non-switch referents, and leaves 5, 6, and 8 consist of all switch referents. Specifically, comparing nodes 3 and 9 with nodes 6 and 8, the apparently higher proportion of free subject pronoun presence among switch referents is as anticipated in the regression model. On the other hand, the low frequency of free subject pronouns in node 5, despite the switch in referent, is in fact a result of interaction with the referents' person and number (see also Section 4.3.3). The result also suggests that although switch in referent plays an important role in affecting free subject pronoun presence, it mainly fine-tunes the decision as it only leads to secondary splits in the classification tree, while person and number have evidently provided stronger discriminatory power.

4.3.3 Person and number

Table 4-8: Free subject pronoun presence under the first split

		Presence of FSP				
		N		Y		Total
Node in	Nodes 3-6	1015	96.39%	38	3.61%	1053
the tree	Nodes 8-9	326	84.24%	61	15.76%	387

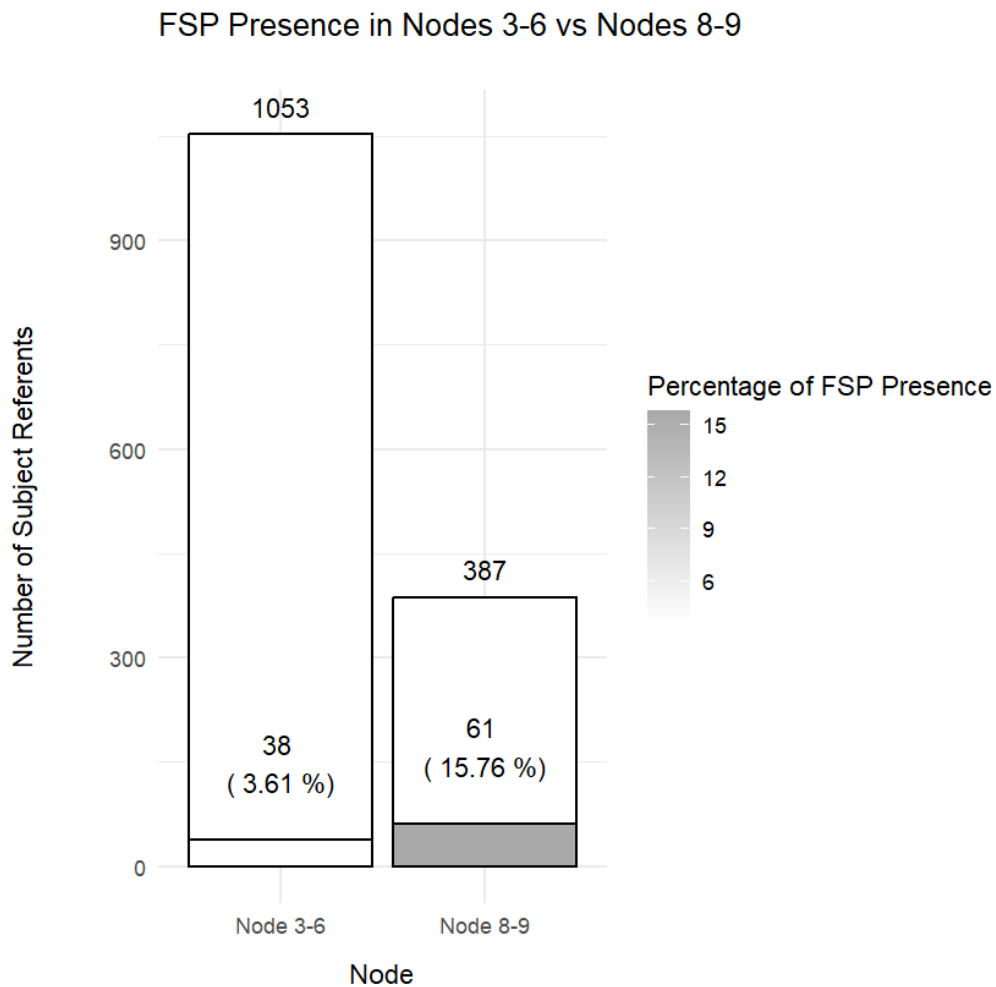


Figure 4-8: Free subject pronoun presence under the first split

According to Figure 4-6, the subject referents' person and number, where the first split comes from, in fact play the most important role in deciding on whether free subject pronouns are present. At the first split, 1sg, 1du.incl, and 2sg subject referents are separated from the rest. Table 4-8 and Figure 4-8 demonstrate that 1sg, 1du.incl, and 2sg subject referents (nodes 8-9) are much more frequently referenced by free subject pronouns. In addition to examples like (19) and (20) above showing 1sg and 2sg subject referents being referenced by free subject pronouns, example (28) below is an instance of free subject pronoun also referencing a 1du.incl referent:

- (28) *ae* *kinta* *tag=to* *psa-k*
 hey **1pl.INCL** **1du.INCL.NOM=IPFV** speak.R-TR
 bislama, *nag* *ḡa=kano* *tae* *franīs*
 Bislama 2sg 2sg.IRR.NOM=be.unable know French
 ‘Hey, (if) we (two) speak Bislama, you will not know French.’
 (SL1-20060402TomsenFrance-TNIs_043)

Moreover, apart from all belonging to speech-act participants, 1sg, 1du.incl, and 2sg subject referents are all deictic referents, and they refer to either the sole speaker (1sg), the sole hearer (2sg), or both (1du.incl). In response to the discussion in § 4.2 above, this finding suggests that, rather than the sole effect of either SAP or referent type, it is indeed the interaction between both that relates to the presence of free subject pronouns, and a deictic speech-act participant is most likely to be referenced with a free subject pronoun. Furthermore, split 7 in Figure 4-6 indicates within deictic SAPs, switch referents are significantly more likely to have free pronoun references.

The next split based on person and number in Figure 4-6 is split 4, illustrating the presence of free subject pronouns is also related to 1du.excl, 1pl.excl, 3du, and 3pl referents when there is a referent switch. As is shown and discussed in the introduction and literature review sections, 1du.excl and 3du subject proclitics in Lelepa share the identical form *ar=*, and the traditional 1pl.excl proclitic form *au=* has been generally replaced by *ur=* in an almost-completed syncretism (Lacrampe, 2014, p. 243), resulting in homonymy between 1pl.excl and 3pl subject proclitics. Table 4-9 and Figure 4-9 illustrate that free subject pronouns in nodes 3-6 are only likely to be present when both the proclitics are ambiguous and there is a referent switch.

Table 4-9: Free subject pronoun presence in nodes 3-6 based on morphological ambiguity and referent switch

		Morphological ambiguity			
		Y		N	
		Switch in referent			
		Y	N	Y	N
Presence of FSP	Y	23	3	6	6
		15.33%	1.44%	1.80%	1.66%
	N	127	205	327	356
		84.67%	98.56%	98.20%	98.34%
	Total		150	208	333

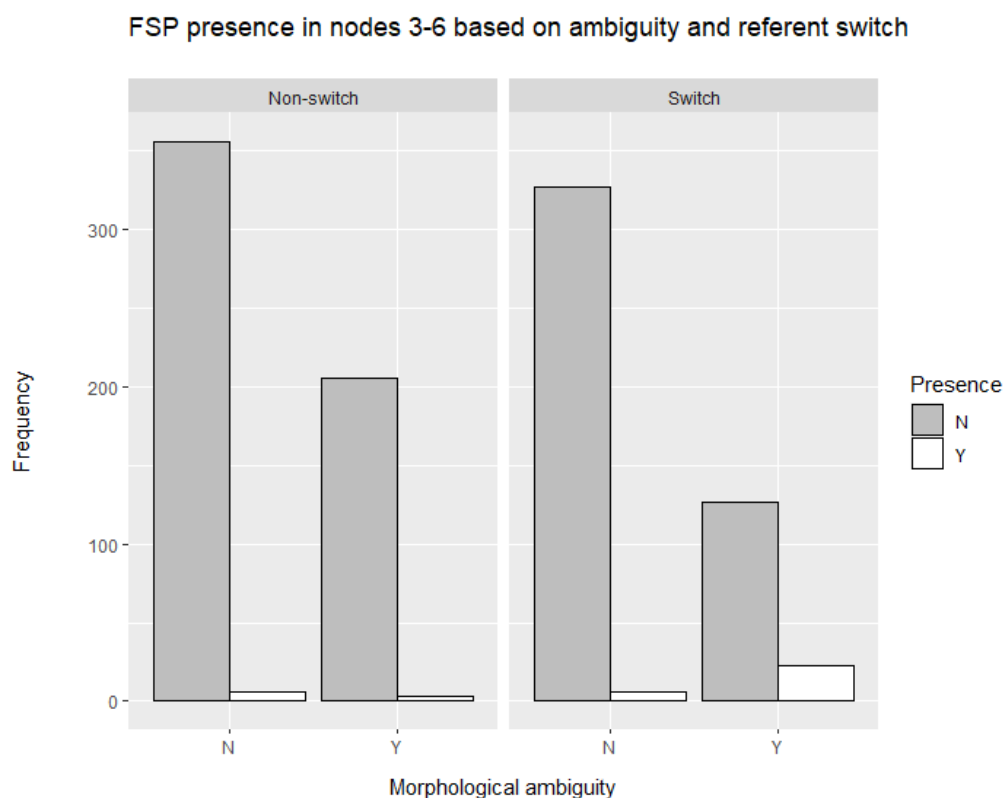


Figure 4-9: Free subject pronoun presence in nodes 3-6 based on morphological ambiguity and referent switch

Therefore, the higher chance of free subject pronouns being present in this category is possibly a way to specify the person and number of these referents

when they are new or re-introduced, since the forms of their corresponding free subject pronouns are different (see also § 5.2.1). Examples (29) – (32) below exhibit how the use of free pronouns helps resolve the ambiguity due to the identical subject proclitic forms of 1pl.excl and 3pl. However, since all the 1pl.excl referents in the corpus are referenced with *ur=* instead of *au=*, it would be necessary for a follow-up study to delve into the use of 1pl.excl and 3pl free subject pronouns in older sources or among older speakers where one may expect to find more instances of *au=* in order to test this hypothesis.

- (29) *kenem* *ur=tu* *naur* *tu [...]*
 1pl.EXCL **1pl.EXCL.NOM=IPFV** island STAT
 ‘We live on this small island [...]’
 (SL2-20102007Ancestors-WManc_027)

- (30) *naara,* *ur=kut* *panei*
 3pl **3pl.NOM=CERT** come
 to *sal* *natm̃at*³¹
 IPFV dance peace.ceremony
 ‘(As for) them, they had come back to dance for peace.’
 (SL1-20080424PeaceCeremony-MLnat_014)

³¹ The orthographic <ṁ> refers to the phoneme /ŋm̃w/ in Lelepa.

- (31) *go naar ur=to pea,*
 and **3pl** **3pl.NOM=IPFV** be.first.R
se konou a=rak go
 while 1sg 1sg.NOM=follow and
a=ta gor naa
 1sg.NOM=cut block HESIT
 ‘And they were first, while I would follow, I would cut and block hum...’
 (SL1-20080424Hunting-BTn_010)

- (32) *napu na kenem malange ur=pa*
 road REL **1pl.EXCL** then **1pl.EXCL.NOM=go.R**
 ‘The road that we go by at that time.’ (SL1-20080424Hunting-BTn_011)

In addition, as is briefly introduced in § 4.3.2 above, node 5 in Figure 4-6 suggests that free subject pronouns are the least likely to be present when the referent is 1pl.incl, 2pl, or 3sg. Although prototypically 1pl.incl and 2pl also mostly refer to deictic referents, they do not show the strongest correlation with the presence of free subject pronouns. As for 2pl referents, it is very likely simply due to the imbalance of the data, since there are very few second-person non-singular referents overall ($n = 6$, with no 2du referent at all in the dataset), and none is referenced with a free subject pronoun. Example (33) below exemplifies a 2pl referent not being referenced by a free subject pronoun:

- (33) [...] *se a=ga magmu*
 as 1sg.NOM=IRR 2pl.BEN
lega=s, kur=gu rog=e
 sing=3sg.ACC 2pl.NOM=IRR hear=3sg.ACC
 '[...]As I will sing it, you (folks) will hear it.'
- (SL1-20070420HarrierFalcon-GMmp_031)

The case for 1pl.incl is nevertheless rather complex, as the 1pl.incl proclitics are in fact sometimes used exclusively rather than inclusively, and/or anaphorically, in multiple recordings (see also Section 5.3.1). This is seen in examples (34) and (35) below, where the proclitic *ur=* and *tu=* refer to the same referent excluding the hearer:

- (34) *mala ur=to sak*
 when 1pl.EXCL.NOM=IPFV go.up
naa ntaaf lag uru=g
 HESIT hill PURP 1pl.EXCL.NOM=IRR
fa-k namlas
 go.IRR-TR forest
 'Once we were going up hum...the hill in order for us to go into the bush.'
- (SL1-20080424Hunting-BTn_009)

- (35) *wan naa tu=pan=s, nat skei,*
 when HESIT 1pl.INCL.NOM=go=3OBL thing one
nat skei e=maeto-keni to
 thing one 3sg.NOM=be.angry-TR STAT
 'When we go there, one thing, one thing would be angry (at something).'
- (SL1-20080424Hunting-BTn_013)

As for 3sg referents, they are less likely to be referenced by free subject pronouns. A possible explanation is that they are much more commonly referenced by lexical NPs while it is very uncommon for a subject referent to be coreferential with two overt NPs, as illustrated in Table 4-10 and Figure 4-10. This will be investigated in more detail in the subsequent analysis in § 4.4 below.

Table 4-10: Types of NP referencing 3sg subject referents

NP Type	Both	Lexical	Null	Pronominal	Total
Number	5	159	389	5	558
Percentage	0.90%	28.49%	69.71%	0.90%	-

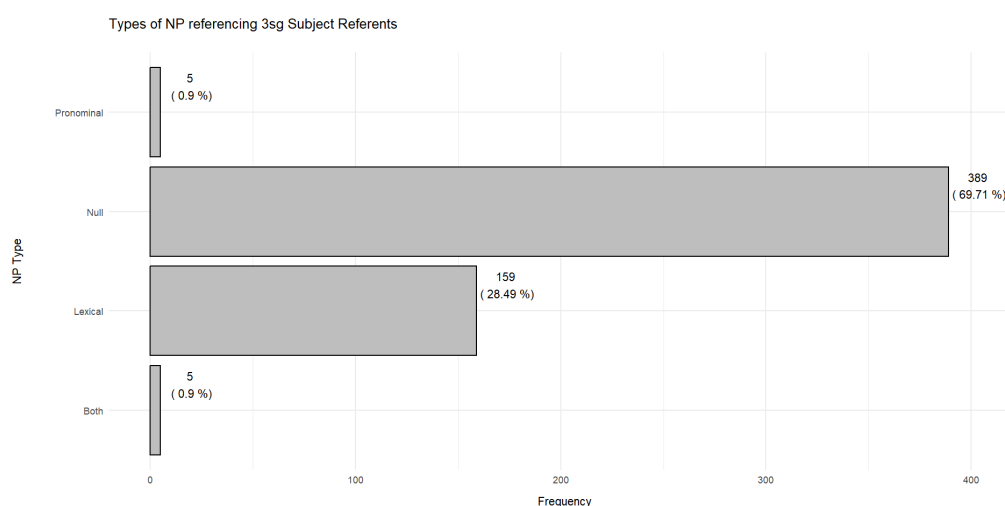


Figure 4-10: Types of NP referencing 3sg subject referents

In brief, the effects of person and number show that free subject pronouns are most frequently present when the referent is a deictic speech-act participant, and it is also likely to be used when the referent is referenced by an ambiguous subject proclitic.

4.4 Study 3: Mblogit analysis of free pronouns

4.4.1 Overview

Table 4-11: Results of the mblogit analysis with reference levels and significant predictors in bold

Lexical NP vs Null NP					
	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	-2.4592	0.3544	-6.939	3.94E-12	***
switch.in.referentY	2.5796	0.3401	7.586	3.31E-14	***
first.occurrence.in.speechY	2.7854	0.2728	10.211	< 2e-16	***
direct.speechY	-1.0894	0.3834	-2.841	0.00449	**
humannessY	-2.0691	0.2642	-7.832	4.80E-15	***
Pronominal NP vs Null NP					
	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	-5.6926	1.0377	-5.486	4.11E-08	***
switch.in.referentY	1.6133	0.2996	5.385	7.24E-08	***
first.occurrence.in.speechY	1.1428	0.4258	2.684	0.00727	**
direct.speechY	1.1378	0.2539	4.481	7.44E-06	***
humannessY	1.7676	1.0215	1.73	0.08357	.

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1					
(Co-)Variances					
Grouping level: speaker					
	Estimate	Std. Error			
Lexical NP vs Null NP	0.0113768	4.10E-07			
Pronominal NP vs Null NP	0.0004456	1.62E-08			
Number of observations					
Groups by speaker: 13					
Individual observations: 1422					

Table 4-11 summarises the results of the mblogit model using *Null NP* as the reference level in order to examine whether the variables are related exclusively to the presence of free subject pronouns or to the overall presence of an overt

subject NP. The estimates of the random effect show that there is only minimal inter-speaker variance in the use of overt NP versus null NP. The negative intercepts ($p < 0.001$) in both lexical and pronominal NP suggest that the presence of an overt NP is uncommon overall (275 overt NP vs 1147 null NP), and pronominal NP is even less likely to be present compared to lexical NP (81 pronominal NP vs 194 lexical NP). According to the fixed effects, a lexical NP is more likely to be present than null NP when there is a switch in referent ($p < 0.001$) or it is the first occurrence in speech of a referent ($p < 0.001$), and it is less likely to be present when the referent is in direct speech ($p < 0.01$) or is human(-like) ($p < 0.001$). In comparison, a significantly higher chance of pronominal NP presence is observed when there is a switch in referent ($p < 0.001$), it is the first occurrence in speech of a referent ($p < 0.01$), or the referent is in direct speech ($p < 0.001$). There is also a trend for pronominal NP to favour human-like referents, but the effect is not significant ($p < 0.01$). Each predictor will also be compared with the results in § 4.2 above in the following subsections.

4.4.2 Switch in referent

Table 4-12: Types of NP based on switch in referent

	Switch in referent		
	N	Y	Total
NP type	Null (N)	733	414
			1147
		63.91%	36.09%
	Pronominal	16	65
	(PRO)	19.75%	80.25%
	Lexical (LEX)	11	183
		5.67%	94.33%
	Total	760	662
		53.45%	46.55%

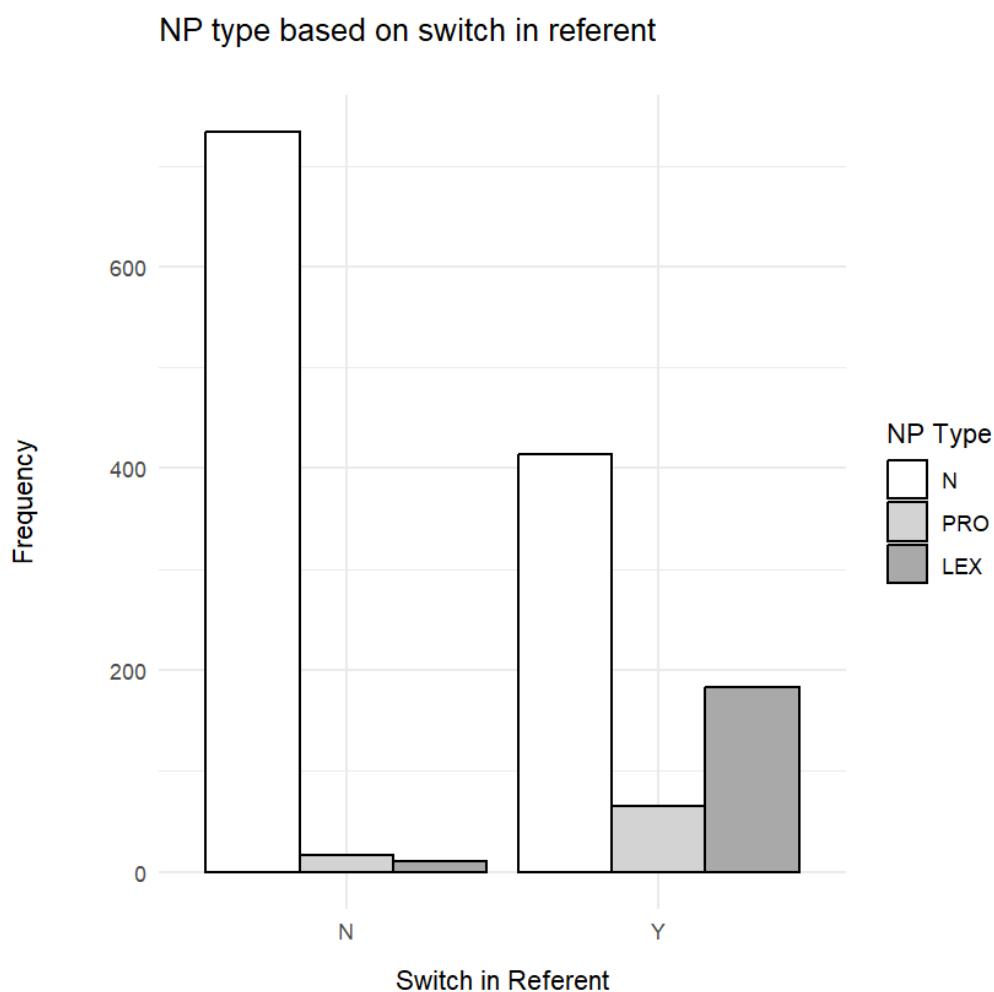


Figure 4-11: Types of NP based on switch in referent

Table 4-12 and Figure 4-11 illustrate that 37.46% of switch referents in this dataset are referenced with an overt NP. 80.25% ($n = 65$) of free subject pronouns and 94.33% of lexical NPs refer to switch referents. In addition, all lexical NPs are found referring to third person referents in this subset of data, especially 3sg referents ($n = 158$). Recalling the results in § 4.3.3 above, 3sg referents also display the lowest probability of being referenced through free subject pronouns. An exploration into the corpus data suggests that lexical NPs rather than pronominal NPs are usually used to specify the subject referents

when there is a switch in referent but no change in the pronominal form, neither bound nor free, which is most commonly found in switch between two 3sg subject referents. Example (36) below exemplifies a switch from a 3sg subject referent to another in two consecutive clauses, where the referents' proclitic forms are identical and hence ambiguous:

- (36) *kanokik* *e=kis* *suk*
 boy **3sg.NOM=press** tight
 n̄pou *soūp̄oumila*.
 head red.headed.honeyeater
 male **sou** *e=kai* *tapla*
 when **honeyeater** **3sg.NOM=cry** like.this
 nina, *ur=ursrapori=s [...]*
 then 3pl.NOM=be.surprised.at=3sg.ACC
 'The little boy squeezed the head of the red-headed bird, when the bird
 cried like this, they were surprised at it [...]'
 (SL1-20080424PeaceCeremony-MLnat_021)

On the other hand, a slightly larger number of continued subject referents prefer free pronouns ($n = 16$, 19.75% among all free pronouns,) to lexical NPs ($n = 11$, 5.67% among all lexical NPs), despite the smaller number of free pronouns overall. For example, in example (37) below, the free pronoun *kono* '1sg' is used while there is no switch in referent:

- (37) *a=to* *pan pa mala*
1sg.NOM=stay GO GO when
ntau e=kat nou,
 year 3sg.NOM=CERT finish
a=kat tae psruk tapla,
1sg.NOM=CERT can speak.R like.this
kono a=kat msau=n lag
1sg 1sg.NOM=CERT want=3sg.ACC COMP
(a=)ga ti to suña to mau
 (1sg.NOM=)IRR NEG stay house STAT NEG
 'I stayed on, at the end of the year, I could speak (French) like this, so I
 wanted not to stay at home.'
 (SL1-20060402TomsenFrance-TNIs_059)

4.4.3 First occurrence in speech

Table 4-13: Types of NP based on first occurrence in speech

		First occurrence in speech		
		N	Y	Total
NP type	N	1121	26	1147
		97.73%	2.27%	-
	PRO	72	9	81
		88.89%	11.11%	-
	LEX	79	115	194
		40.72%	59.28%	-
	Total	1272	150	1422
		89.45%	10.55%	-

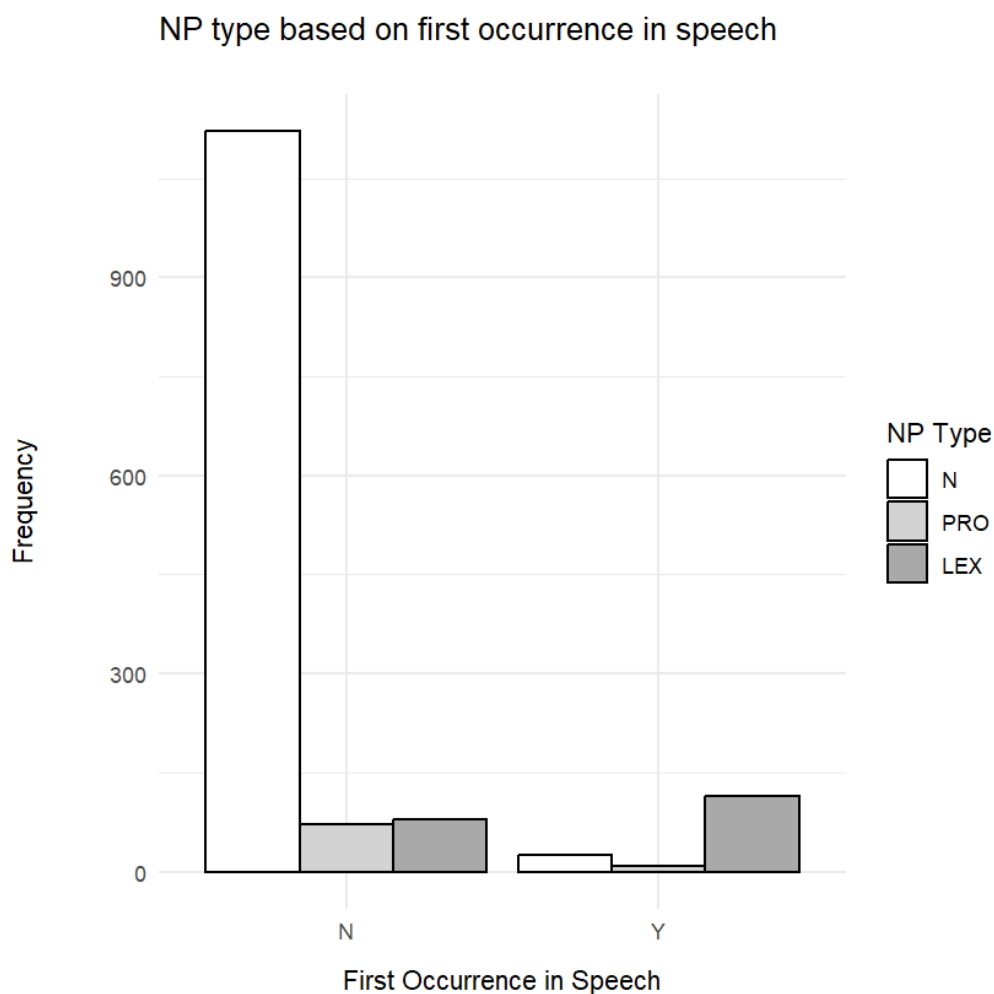


Figure 4-12: Types of NP based on first occurrence in speech

As is depicted in Table 4-13 and Figure 4-12, an overt NP fills the preverbal subject slot 82.67% of the time when the referent is mentioned for the first time in the discourse, in accordance with Lacrampe (2014, pp. 269–272). Furthermore, lexical NPs are much more likely to be present than pronominal NPs when the referent is new, although first occurrence in speech still associates with the presence of free pronouns compared with null NP only. Therefore, its effect is not significant in § 4.2 where lexical and null NP are not distinguished within the *free pronoun absence* category. Like the case in § 4.4.2 above, lexical NPs are extensively used to single out new referents when there

is ambiguity in the possible interpretation of the subject proclitic. However, free pronouns are still commonly found at the beginning of narratives when the speakers introduce themselves, as shown in examples (38) and (39) below. However, example (40) below suggests the use of free pronoun *konou* is optional when it comes to 1sg referents, since there is only one possible interpretation of the subject proclitic *a=*. Hence, while the semantic information can still be equally conveyed in the absence of *konou* ‘1sg’, the use of free subject pronouns for 1sg subject referents in this case is pragmatic (see also § 5.2.2). By contrast, in example (41) below, the speaker wants to ‘individuate the referent’ (Lacrampe, 2014, p. 129) amongst other people who would have all been referenced by the same third-person singular proclitic, hence, the speaker uses a lexical NP with a demonstrative *na* while also pointing at the referent:

- (38) *mais, konou a=to magaliliu to*
 today 1sg 1sg.NOM=stay Magaliliu STAT
 ‘Today, I live in Magaliliu.’ (SL1-20080401Megapode-LLJmal_002)

- (39) *konou, a=pi naota*
 1sg 1sg.NOM=COP chief
 ‘I am the chief.’ (SL1-20080401DevilArtok-CMmut_001)

- (40) *a=ga fsruk malmauna*
 1sg.NOM=IRR speak.IRR now
 ‘I will speak now.’ (SL1-20080424MatWeaving-MRnn.001)

- (41) *narop na e=uḗanakono=s*
 Narop DEM 3sg.NOM=shape=3sg.ACC
 ‘Narop here is shaping it.’ (SL1-20090625CanoeMaking2-TMrar_003)

Moreover, non-singular new referents are sometimes not referenced by any overt NP, instead, they are expressed through *split-antecedent anaphora* (see also § 4.2.4), where each member of the non-singular referent has been mentioned before and the referent is hence assumed based on the context, as in (42) below:

- (42) **e=msug** **konou** **ar=kat**
3sg.NOM=take **1sg** **1du.EXCL.NOM=CERT**
pa-ki *bellevue* *pa*
go.R-TR Bellevue GO
‘He took me and we (two) went to Bellevue.’
(SL1-20060402TomsenFrance-TNIs_012)

4.4.4 Direct speech

Table 4-14: Types of NP based on direct speech

		Direct speech		
		N	Y	Total
NP type	N	984	163	1147
		85.79%	14.21%	-
	PRO	46	35	81
		56.79%	43.21%	-
	LEX	183	11	194
		94.33%	5.67%	-
	Total	1213	209	1422
		85.30%	14.70%	-

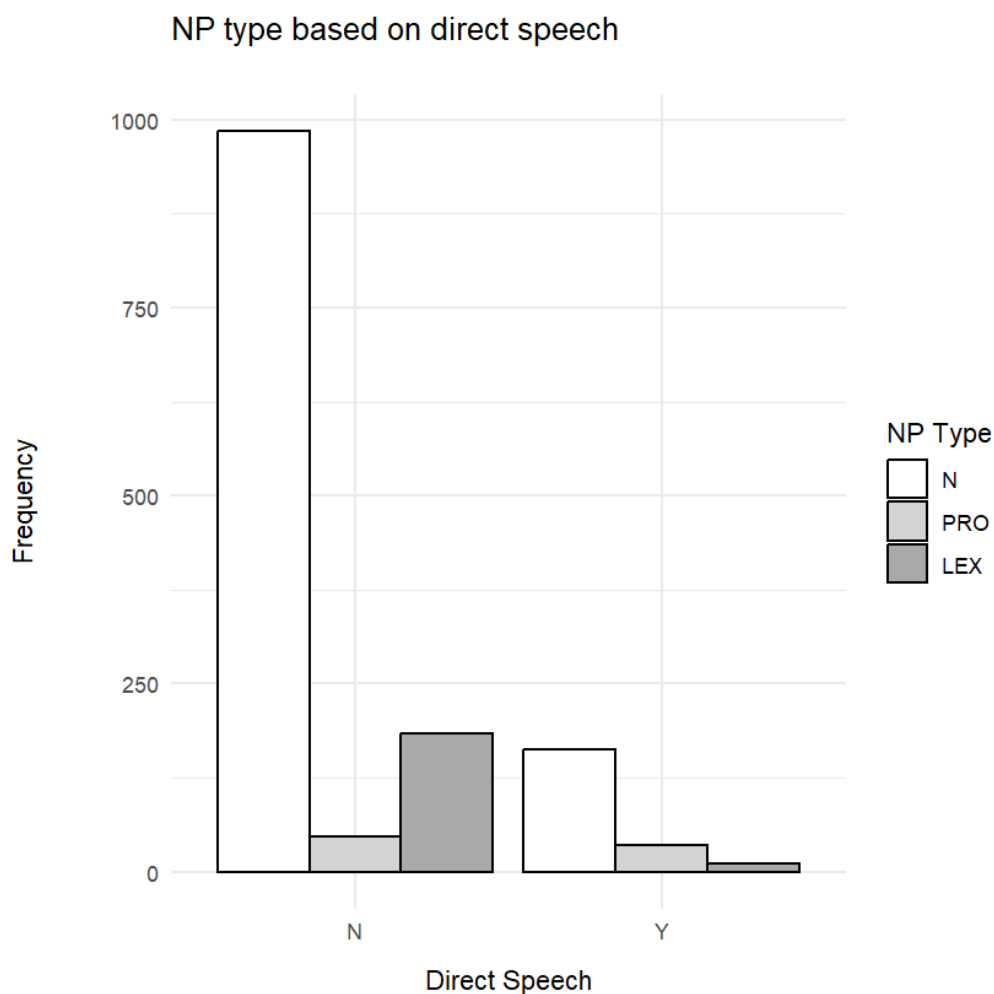


Figure 4-13: Types of NP based on direct speech

§ 4.4.2 and § 4.4.3 suggest that switch in referent and first occurrence in speech relate to the presence of both lexical and pronominal NP compared with null NP. However, as is shown in Table 4-14 and Figure 4-13, subject referents in direct speech are much more likely to be referenced with free pronouns, while they are also significantly related to the absence of lexical NPs. This result suggests that the common ground hypothesis in § 4.2.5 above is incorrect, since direct-speech referents are treated differently from regular switch referents. Moreover, Table 4-15 and Figure 4-14 depict the clear distinction in the use between lexical and pronominal NPs based on speech-act participant, that SAPs in direct

speech are consistently devoid of lexical NP references, while 19.55% of them are referenced by free subject pronouns. Consequently, the fact that most (85.65%) subject referents in direct speech are SAPs may account for the higher frequency of free pronouns in direct speech.

Table 4-15: Types of NP in direct speech based on person

		Person			
		First	Second	Third	Total
NP type	N	68	76	19	163
		41.72%	46.63%	11.66%	-
	PRO	21	14	0	35
		60.00%	40.00%	0.00%	-
	LEX	0	0	11	11
		0.00%	0.00%	100.00%	-
	Total	89	90	30	209
		42.58%	43.06%	14.35%	-

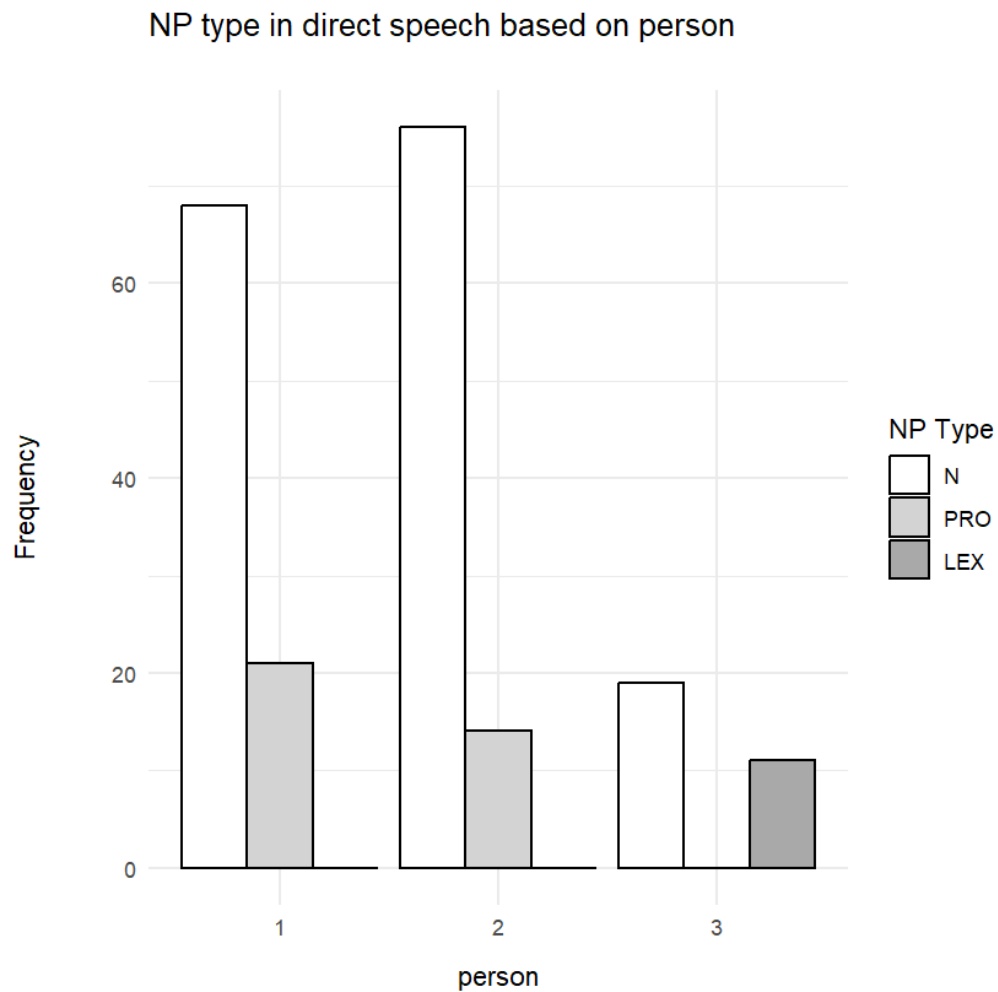


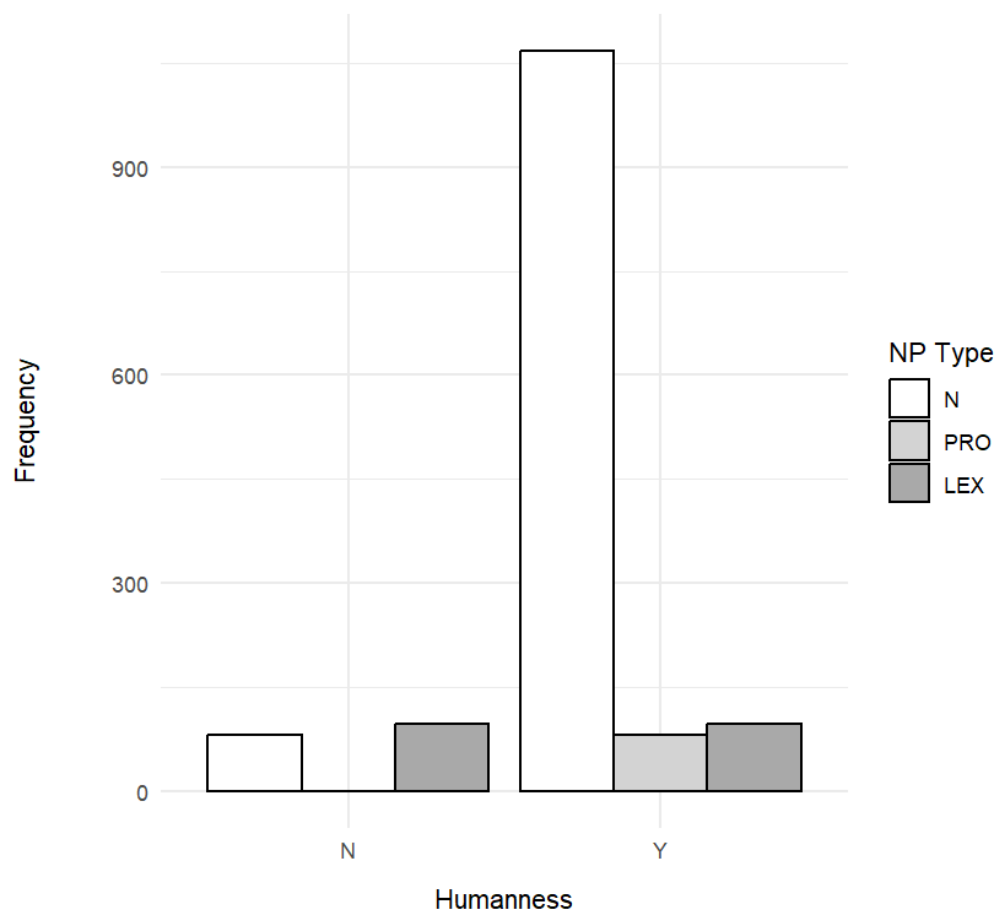
Figure 4-14: Types of NP in direct speech based on person

4.4.5 Humanness

In light of the distribution outlined in Table 4-16 and Figure 4-15, a substantial majority of non-human-like subject referents have lexical NP references, while none is referenced by a free pronoun. Examples (43) and (44) below show lexical NPs referring to non-human-like and human-like referents respectively.

Table 4-16: Types of NP according to humanness

		Humanness		
		N	Y	Total
NP type	N	80	1067	1147
		6.97%	93.03%	-
	PRO	0	81	81
		0.00%	100.00%	-
	LEX	97	97	194
		50.00%	50.00%	-
	Total	177	1245	1422
		12.45%	87.55%	-

NP type according to humanness**Figure 4-15: Types of NP according to humanness**

- (43) *naki* *e=ga* *ĩmit* *tapla*
 pandanus **3sg.NOM=IRR** be.short like.this

‘The pandanus will be short like that.’

(SL1-20080424MatWeaving-MRnn.015)

- (44) *manuwia* *e=to* *pagtof* *neik*
 Manuwia **3sg.NOM=IPFV** buy fish

‘Manuwia used to buy fish.’ (SL1-20080404Fishing-BPatg_019)

There is a seemingly minor disparity between the results of the multinomial model here and the binomial model in § 4.2.3 above in that the correlation between humanness and presence of free subject pronouns is no longer significant in the multinomial model. However, these findings are in fact not contradictory. The results in the binomial model indicate that a pronominal NP is more likely to occur, compared with any other types of NP, i.e. lexical or null NP, when the referent is human-like. However, since bound pronouns in Lelepa are obligatory (Lacrampe, 2014, p. 243), all (or at least most, see § 4.5 below for a more detailed picture) subject referents with null NP still have pronominal references. Hence, the results of this multinomial model demonstrate that human-like referents tend to favour pronominal references over lexical NPs, while there is no significant preference for either free or bound pronouns over the other, despite a slight trend towards using free pronouns in the results. Considering the fact that all speech-act participants are human(-like), the effect of SAP (see § 4.3.3) may account for the preference for overt realisation.

4.5 Study 4: Subject proclitic deletion based on multiple models

4.5.1 Overview

This section presents the results of multiple models implemented in study 4 investigating what variables relate to the presence of subject proclitics in Lelepa. Additional observations supported by descriptive statistical data, which are informed by the inferential statistical analysis but not presented in the results, will also be covered in this section through a more thorough exploration into the corpus.

Figure 4-16 shows the permutation importance scores of the variables' effects on subject proclitic presence from a preliminary random forest analysis using `{party}` (Hothorn et al., 2023) and `{permimp}` (Debeer et al., 2021) in R (R Core Team, 2023). The results suggest that the correlation between the presence of subject proclitics and *switch in referent*, *first occurrence in speech*, and *humanness* is not significant, unlike free subject pronouns. The five variables scoring over 0.1 in their importance scores are employed in the final model.

Overall, unlike free subject pronouns, subject proclitics are unlikely to be absent (16.52%, $n = 260$). As illustrated in Figure 4-17, four inputs are presented in the classification tree, which will be discussed in the following subsections according to the results. Furthermore, although the variable *person and number* is excluded in the results, an in-depth exploration of the effect of person and number will also be included in this section, elucidated with data, plots, and examples.

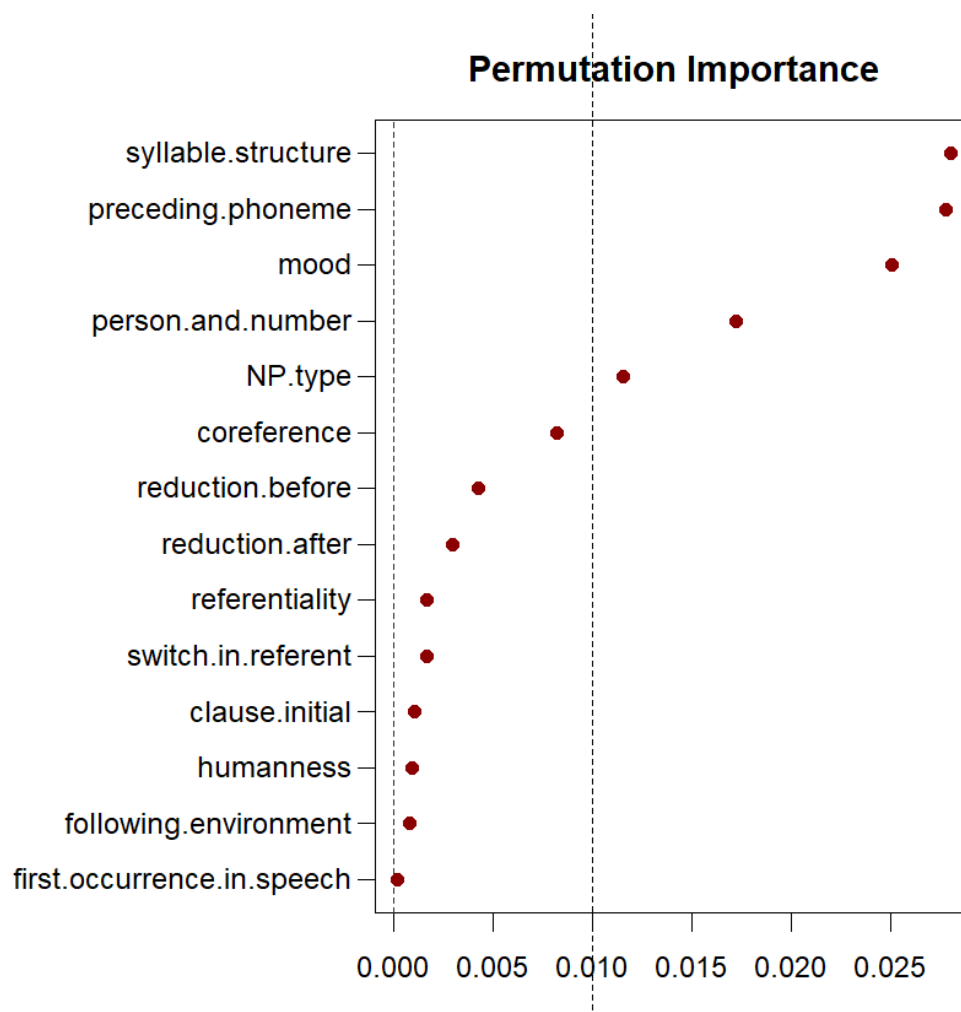


Figure 4-16: Importance scores of predictors of subject proclitic presence

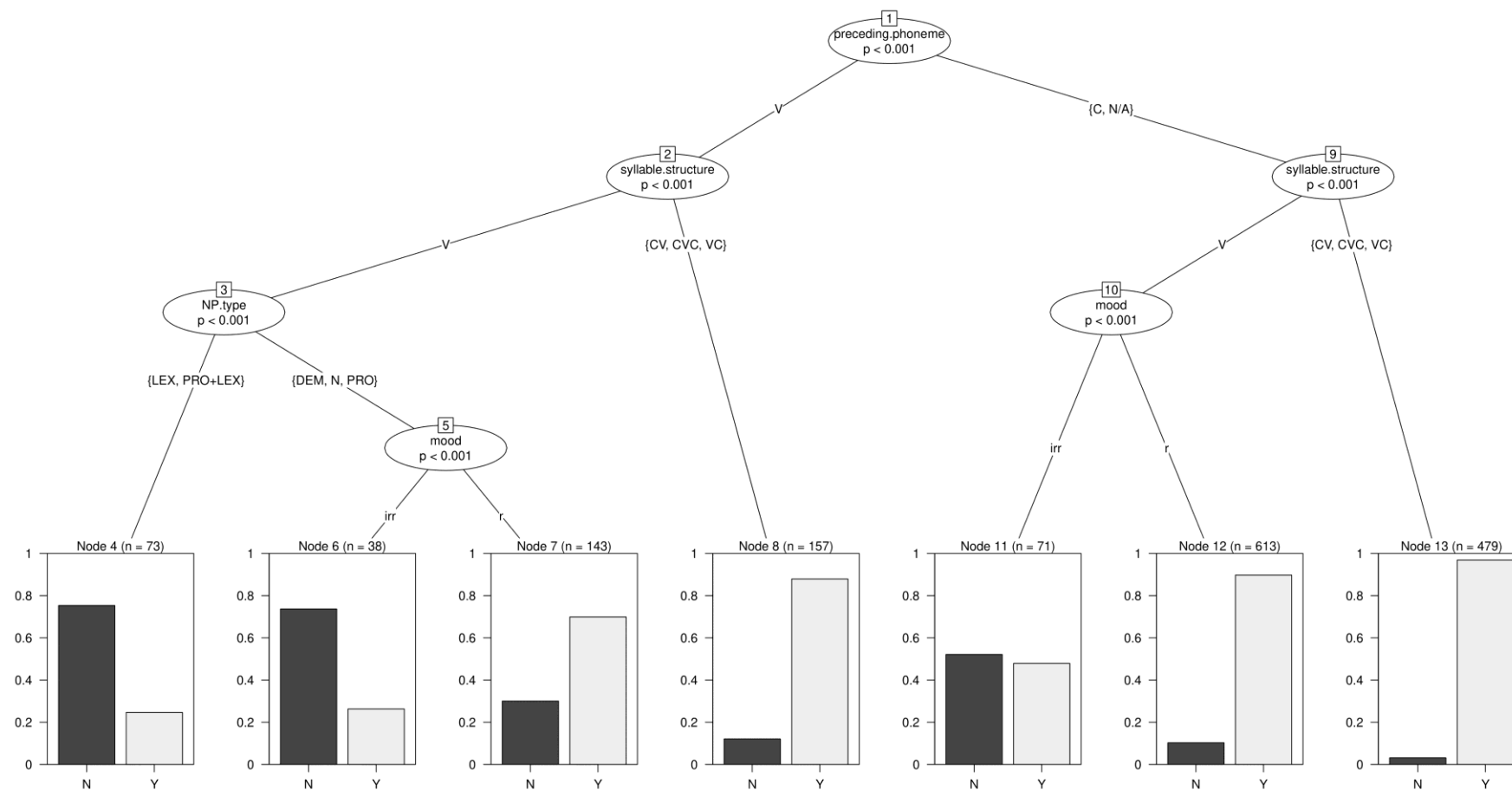


Figure 4-17: Conditional classification tree of subject proclitic presence

4.5.2 Preceding phoneme

Table 4-17: Presence of subject proclitics based on preceding phoneme (1)

		Preceding phoneme						
		Consonant (C)		Pause (N/A)		Vowel (V)		Total
Subject proclitic presence	N	58	22.31%	57	21.92%	145	55.77%	260
	Y	511	38.89%	537	40.87%	266	20.24%	1314
	Total	569	36.15%	594	37.74%	411	26.11%	1574

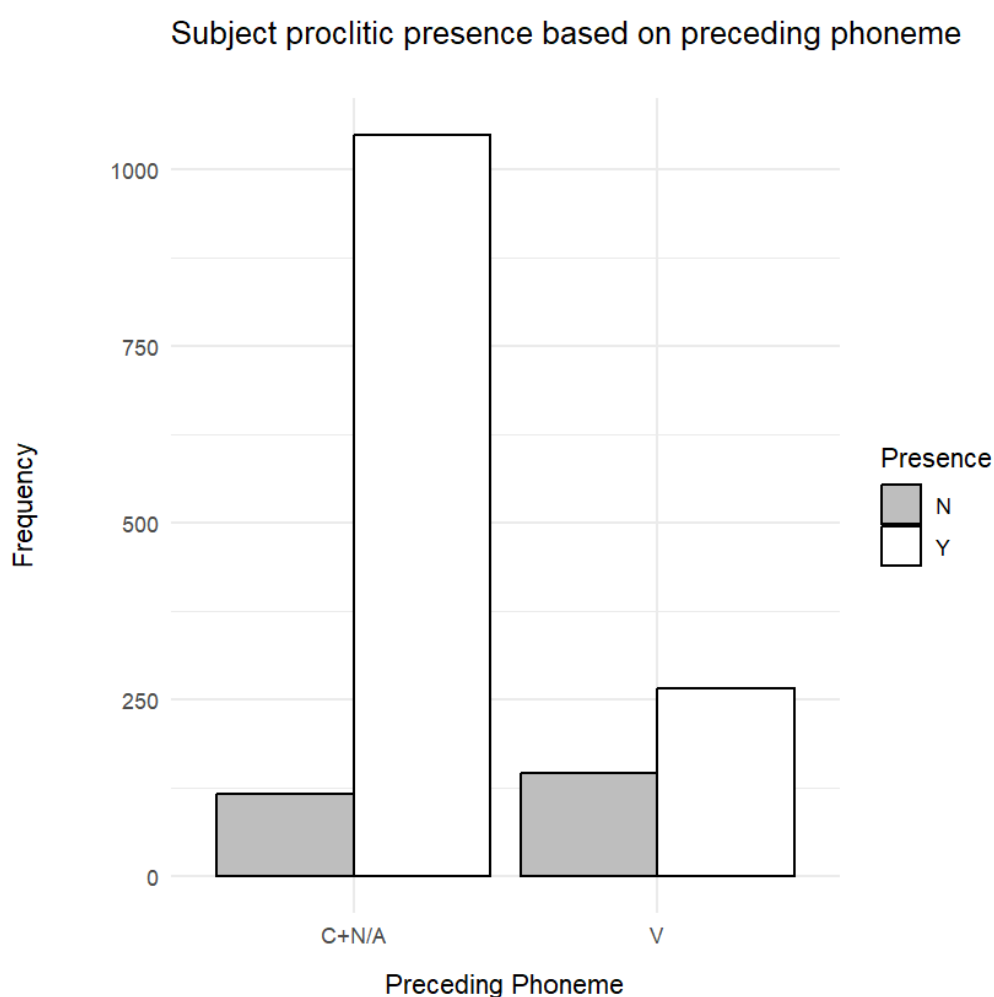
Preceding phoneme leads the first split which results in nodes 4-8 versus nodes 11-15 in the classification tree, demonstrating a subject proclitic is significantly more commonly deleted when following a vowel, while Table 4-17 points out that there is little variation between the overall presence of post-consonant and clause-initial subject proclitics. The distribution of subject proclitics following vowels versus non-vowels is illustrated in Table 4-18 and Figure 4-18. Example (45) below includes an omitted subject proclitic³²:

- (45) *kan a=kat ti tae takanei*
 but **1sg.NOM=CERT** NEG know how
(a=)ga fat=i mou
(1sg.NOM=)IRR make.IRR=3sg.ACC NEG
 'But I didn't know what I would do.'
 (SL1-20060402TomsenFrance-TNIs_041)

³² Henceforth, morphemes in parentheses are those which are not pronounced but which appear in the transcription of the recording, henceforth (see also § 3.5.3).

Table 4-18: Presence of subject proclitics based on preceding phoneme (2)

		Preceding phoneme				
		C + N/A		V		Total
Subject	N	115	44.23%	145	55.77%	260
proclitic	Y	1048	79.76%	266	20.24%	1314
presence	Total	1163	73.89%	411	26.11%	1574

**Figure 4-18: Presence of subject proclitics based on preceding phoneme**

In terms of exceptions, subject proclitics following a consonant are also often found deleted (74.14%, $n = 43$) when the preceding and following consonants are homorganic or share the same manner of articulation. Example (46) below

exemplifies a proclitic omission where the following irrealis marker *ga* is bound to the preceding lexeme *nag*, forming *nag=ga*:

- (46) *ḡa=pea* *pu* *lwa* *fefe*
 2sg.IRR.NOM=be.first.R pull.R remove oven.blanket
nag *(e=)ga* *skimau,* *(a=)ga*
 2sg.POSS (3sg.NOM=)IRR one (1sg.NOM=)IRR
loḡa=e *s* *a=g* *kat* *pa*
 see=3sg.ACC then 1sg.NOM=IRR CERT go
 ‘First take out one of your oven blankets, I will look then I will go.’
 (SL1-20080401DevilArtok-CMmut 070)

In addition, the 1sg subject proclitics *a=* between a diphthong and an irrealis marker *ga* are both deleted in examples (45) and (46) above, while the other 1sg subject proclitic *a=* in (46) is retained (see also § 4.5.5). Since vowel sequences can be realised over syllable boundaries and sequences of over two vowels are uncommon and unlikely to be fully realised (Lacrampe, 2014, pp. 56–58), the zero-onset proclitic *a=* is deleted possibly in order to avoid an overly complex vowel sequence. On the other hand, the vowels in the lexemes around the retained 1sg proclitic *a=* in (46) are both deleted, resulting in a prosodic word *sag* [saŋ] with *a=* being the nucleus. The presence of subject proclitic *a=* in this case is virtually mandatory, since otherwise the /s/ and /ŋ/ would have had to form a prosodic word either on their own as *[sŋ] or with the following *kat* as *[sŋkat], which are both impossible since nasals can only be syllabified word-initially (Lacrampe, 2014, pp. 21–22, 48, 71).

4.5.3 Syllable structure

Table 4-19: Presence of subject proclitics based on syllable structure

		Syllable structure				
		Vowel-only (V)		Others (Non-V)		Total
Subject proclitic presence	N	226	86.92%	34	13.08%	260
	Y	712	54.19%	602	45.81%	1314
	Total	938	59.59%	636	40.41%	1574

Subject proclitic presence based on syllable structure

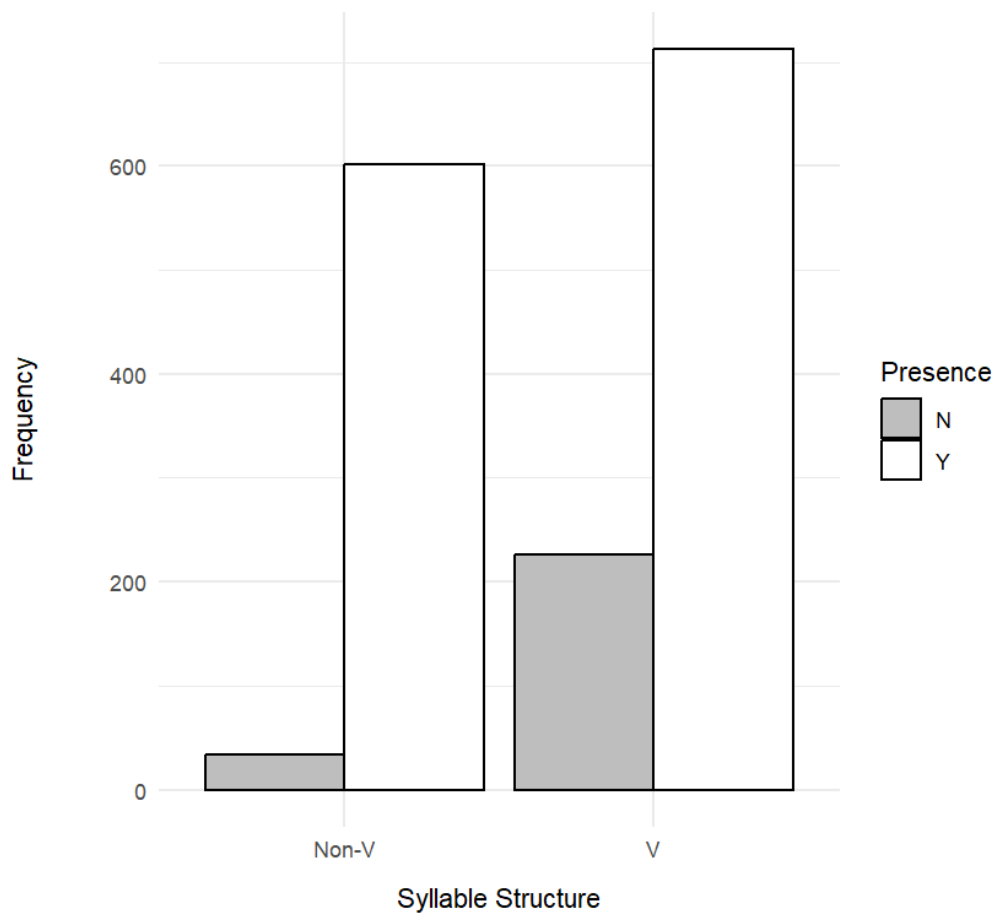


Figure 4-19: Presence of subject proclitics based on syllable structure

According to splits 2 and 9 in Figure 4-17, subject proclitics with a nucleus-only syllable structure are separated from the rest. As Table 4-19 and Figure 4-19 further demonstrate, 86.92% (n = 226) of the deleted proclitics are nucleus-only, i.e. 1sg or 3sg. Furthermore, instances where the vowel of a non-nucleus-only proclitic is dropped while the consonant is retained are extremely rare (1.42%, n = 9), suggesting this division is not just another effect of the extensive vowel reduction in Lelepa. In productive vowel deletion processes in Nafsan, CV subject proclitics are also an environment where vowel deletion that otherwise would occur is prevented from occurring (Billington et al., 2020, pp. 140–142). This may nonetheless relate to the high frequency of 1sg and 3sg proclitics in Lelepa speech, as Bybee (2006, p. 714) summarises that ‘high-frequency words and phrases undergo phonetic reduction at a faster rate’. However, examples (47) – (49) below exemplify a series of peculiar cases where the 3pl proclitic with a VC syllable structure is also deleted:

- (47) *ur=slat=i* (ur=)kat
 3pl.NOM=carry=3sg.ACC (3pl.NOM=)CERT
mato *lega~lega-ki=ni* *pa*
 IPFV sing~RED-TR=3sg.ACC GO
 ‘They carried him, (they were) singing to him.’
 (SL1-20080424PeaceCeremony-MLnat_017)
- (48) *ru=panei* (ur=)slat=i
 3pl.NOM=come (3pl.NOM=)carry=3sg.ACC
 ‘They came carrying him.’
 (SL1-20080424PeaceCeremony-MLnat_022)

- (49) **(ur=)***slat=i* *se*
(3pl.NOM=)carry=3sg.ACC while
(ur=)*lega~lga-ki=ni*
(3pl.NOM=)sing~RED-TR=3sg.ACC
 ‘They carried him while they were singing to him.’
 (SL1-20080424PeaceCeremony-MLnat_023)

In fact, examples (47) and (48) are instances of clause-chaining in Lelepa, which is not discussed in detail in the reference grammar (Lacrampe, 2014, pp. 166–168). Based on the corpus data, in Lelepa clause-chains, any subject proclitic in the medial and final clauses can be deleted while the referent can still be understood based on prosodic and contextual cues (see also § 5.3.1). On the other hand, the deletion of subject proclitics in (49) is possibly a free variation as the referents are highly salient in the context. In addition, the mechanisms driving the higher probability of 1sg and 3sg subject proclitic deletion may not be analogous, which will be discussed in Sections 4.5.4 and 4.5.5 below.

4.5.4 NP type

Table 4-20: Presence of subject proclitics based on NP type

		Subject proclitic presence				
<i>(DEM = demonstrative)</i>		N		Y		Total
NP type	DEM & N & PRO	173	12.88%	1170	87.12%	1343
	LEX & PRO+LEX	87	37.66%	144	62.34%	231
	Total	260	16.52%	1314	83.48%	1574

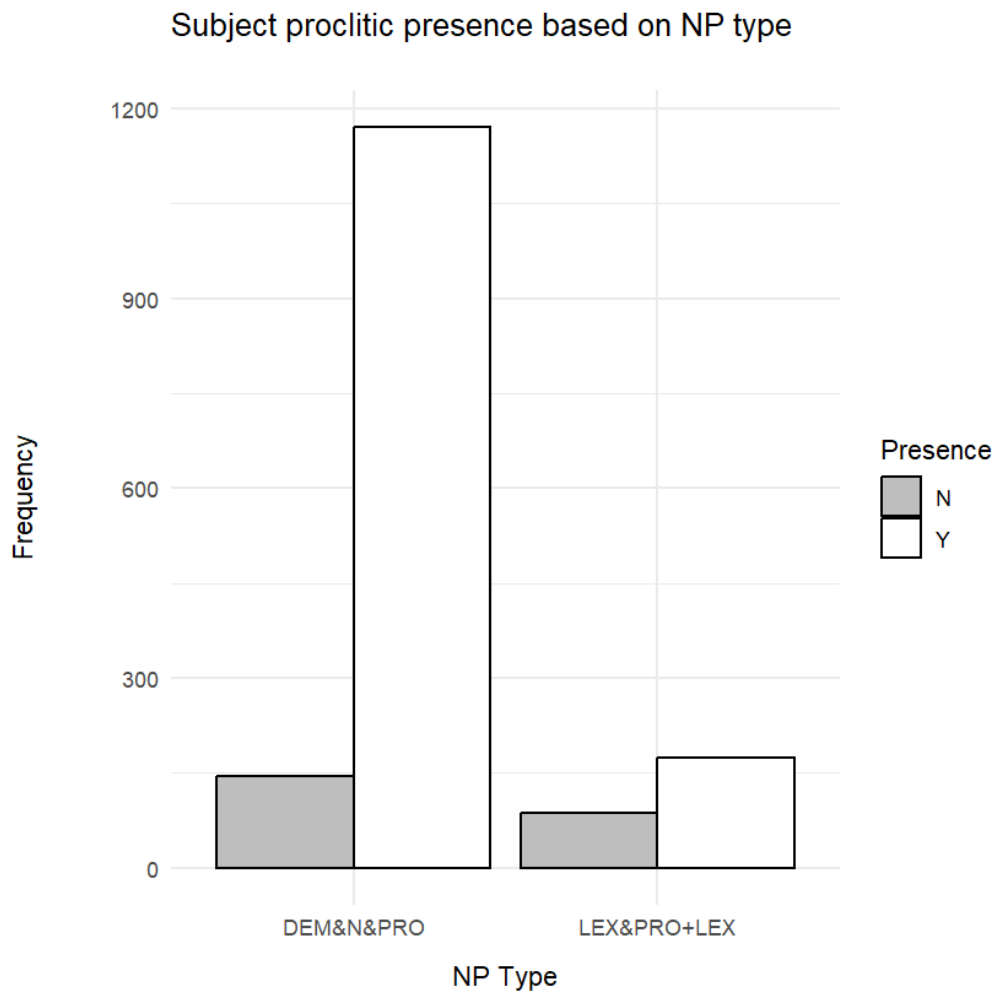


Figure 4-20: Presence of subject proclitics based on NP type

Split 3 comes from NP type in Figure 4-17, separating a subset of data exclusively comprising nucleus-only proclitics in the third level of the tree based on whether a lexical NP reference is present, suggesting that the presence of 1sg and 3sg subject proclitics is the most sensitive to the presence of a lexical NP. Building on Lacrampe's (2014, p. 351) comment that subject proclitics are sometimes omitted when a coreferential NP is present, the effect of coreferential NP type on subject proclitic variation will be discussed in this subsection. Overall, as is shown in Table 4-20 and Figure 4-20, over one-third of subject proclitics are deleted when the referent is referenced by a lexical NP. In addition,

1sg and 3sg referents take up 97.70% (n = 85) in this category. The following sentences illustrate how the subject proclitics distribute differently with different NP types being present, where (50), (51), and (52) includes a demonstrative pronoun, a lexical NP, and a free pronoun respectively:

- (50) *mala tu=pan pa pa-k namlas,*
 when 1pl.INCL.NOM=go GO go.TR forest
te-n e=pi te-fea
SBST-DEM 3sg.NOM=COP SBST-be.first.IRR
na tu=...
 REL 1pl.INCL.NOM=
 ‘When we go to the forest, this is the first thing that we...’
 (SL1-20080424Hunting-BTn_017)

- (51) *a=pat=i, masta agnou*
 1sg.NOM=make=3sg.ACC **boss 1sg.POSS**
(e=)kat rog maeto-ki konou
(3sg.NOM=)CERT feel be.angry-TR 1sg
 ‘I did it, (and) my boss got angry at me.’
 (SL1-20060402TomsenFrance-TNIs_061)

- (52) *e=lag ‘ae, kon*
 3sg.NOM=say hey 1sg
a=lag nag[...]
1sg.NOM=say 2sg
 ‘She said, “hey, I told you [...]”’
 (SL1-20080401DevilArtok-CMmut_073)

In the examples above, the final vowel(s) of the demonstrative pronoun *tena* ‘this’ and free pronoun *konou* ‘1sg’ are deleted and the corresponding proclitics are retained and form a prosodic word with the coreferential pronoun, whereas the final vowel in the lexical NP is retained, instead, the proclitic is deleted. However, similar proclitic deletion is also observed with free pronouns being present, as in example (53) below:

- (53) *a=rki=nia=s* *lag,*
 1sg.NOM=tell=3sg.ACC=3OBL COMP
‘kon a=mal na-fan-ona,
 1sg 1sg.NOM=not.want NMLZ-go.IRR-NMLZ
konou (a=)ga to’
 1sg (1sg.NOM=)IRR stay
 ‘I told him, “I don’t want to go! I will stay.”’
 (SL1-20060402TomsenFrance-TNIs_072)

In (53), the final diphthong of the first *konou* is deleted and the proclitic is retained, while the second *konou* is not reduced and the proclitic is omitted. Despite the higher likelihood of 1sg proclitics being deleted when in irrealis mood (see also Section 4.5.5), the deletion of proclitics after *agnou* in (51) and *konou* in (53) also prevents three-vowel sequences in the prosodic phrase formation process. However, 3sg proclitic omission also takes place in (54) where the final vowel of the diphthong in the lexical NP *kerua* ‘the second’ is deleted and in (55) where the lexical NP *mala nag* ‘your time’ ends with a consonant. By contrast, the subject proclitic is still present when only the final /u/ of the diphthong /ou/ in *konou* is deleted in (56) below, suggesting that the

deletion of subject proclitics is related to lexical NP rather than word-final diphthongs in the coreferential NP:

- (54) *ke-ru* *(e=)lag* ‘*ee*,
ORD-two **(3sg.NOM=)**say no
e=ti *pi* *taati* *mou[...]*’
3sg.NOM=NEG COP maternal.grandma NEG
‘The second one says, “no, this is not grandma[...].”’
(SL1-20080401DevilArtok-CMmut_128)

- (55) *mala nag* *(e=)kat* *nou*
time **2sg.POSS** **(3sg.NOM=)**CERT finish
‘Your time is finished.’ (SL1-20060402TomsenFrance-TNIs_073)

- (56) *kono a=ti* *pa-k* *warae* *mau*
1sg **1sg.NOM=NEG** go.R-TR there NEG
‘As for me, I didn’t go there.’ (SL1-20060402TomsenFrance-TNIs_016)

The examples above, however, all focus on instances in which the subject NPs are clause-initial and directly precede the proclitics. When it comes to clear³³ examples of proclitics being coreferential with left-dislocated subject NPs, which are separated from the main clause by a pause (Lacrampe, 2014, p. 272), all of them are present (n = 18). Examples (57) and (58) below show that both the subject proclitics and coreferential NPs are not deleted:

³³ What I mean by *clear* is that whether the subject NP is clause-initial or left-dislocated is unclear in some other cases.

- (57) ***konou,*** *[pause]* ***a=pi*** ***naota***
 1sg 1sg.NOM=COP chief
 'I am the chief.' (SL1-20080401DevilArtok-CMmut_001)

- (58) ***nkas*** ***kiksa*** ***nge,*** *[pause]*
 wood very.small DEM
e=ga ***latsa***
 3sg.NOM=IRR six
 '(As for) the little pieces of wood, they should be six.'
 (SL1-20080424Hunting-BTn_040)

Furthermore, the correlation between the presence of non-1sg or -3sg proclitics and NP type is not significant in the classification tree, possibly due to the extremely low percentage of these proclitics being absent (5.35%, n = 34), which may be multiple effects of person and number (see also § 4.5.6). Example (59) below is an instance where a 3pl subject proclitic is still present although it directly follows a coreferential lexical NP:

- (59) ***pa=liko*** ***suk~suk*** ***konou***
 2sg.IRR.NOM=hold tight~RED 1sg
nlakan ***natañol*** ***ur=laapa***
 because people 3pl.NOM=plenty
 'Hold on to me tight, because there are a lot of people.'
 (SL1-20060402TomsenFrance-TNIs_034)

Finally, since third-person singular subject referents have the most lexical NP references (72.69%, n = 181), this possibly accounts for the high proportion of

3sg subject proclitics being deleted rather than simply having a nucleus-only syllable structure, as is briefly discussed in § 4.5.3 above.

4.5.5 Mood

The effect of mood in Figure 4-17 is rather vague, and merely suggests that 1sg and 3sg referents are more likely to be deleted when in irrealis mood, which is, nonetheless, also true in terms of the whole dataset, as in Table 4-21 and Figure 4-21:

Table 4-21: Presence of subject proclitics based on mood

		Mood				
		Irrealis (IRR)		Realis (R)		Total
Subject proclitic presence	N	81	31.15%	179	68.85%	260
	Y	226	17.20%	1088	82.80%	1314
	Total	307	19.50%	1267	80.50%	1574

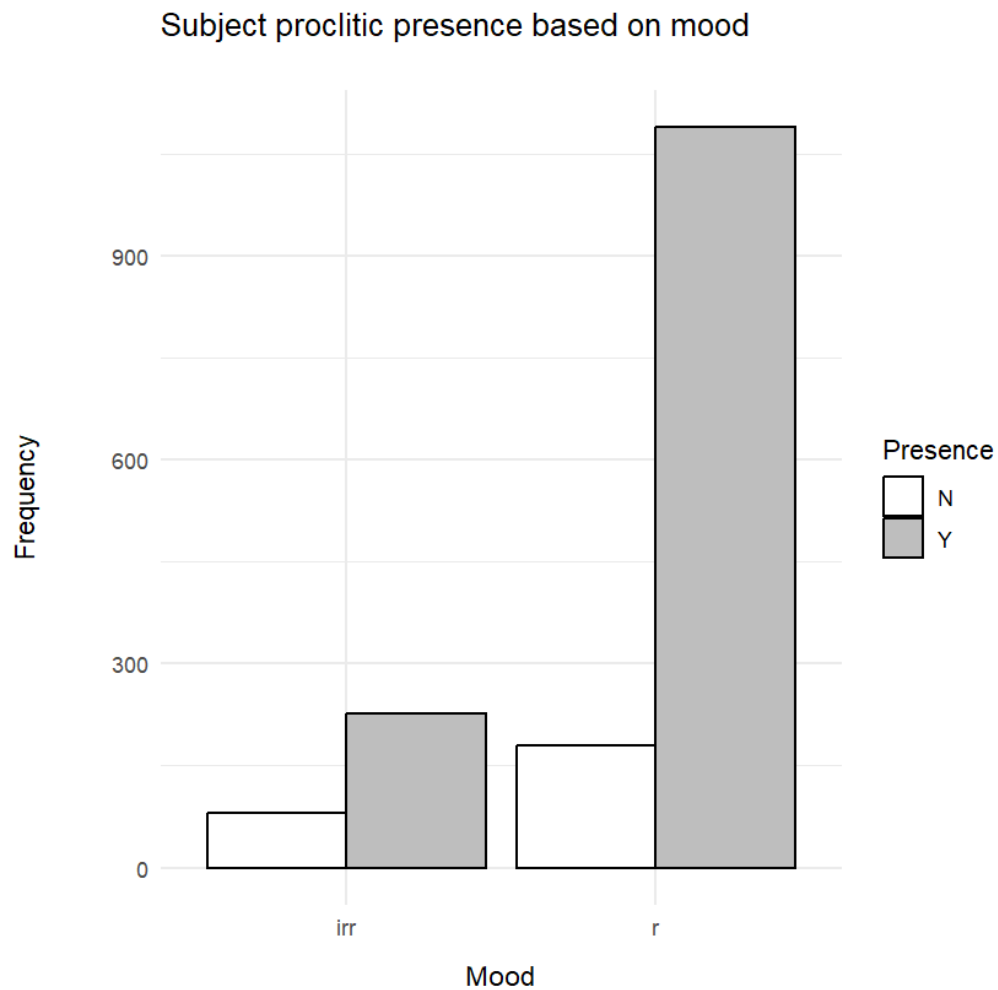
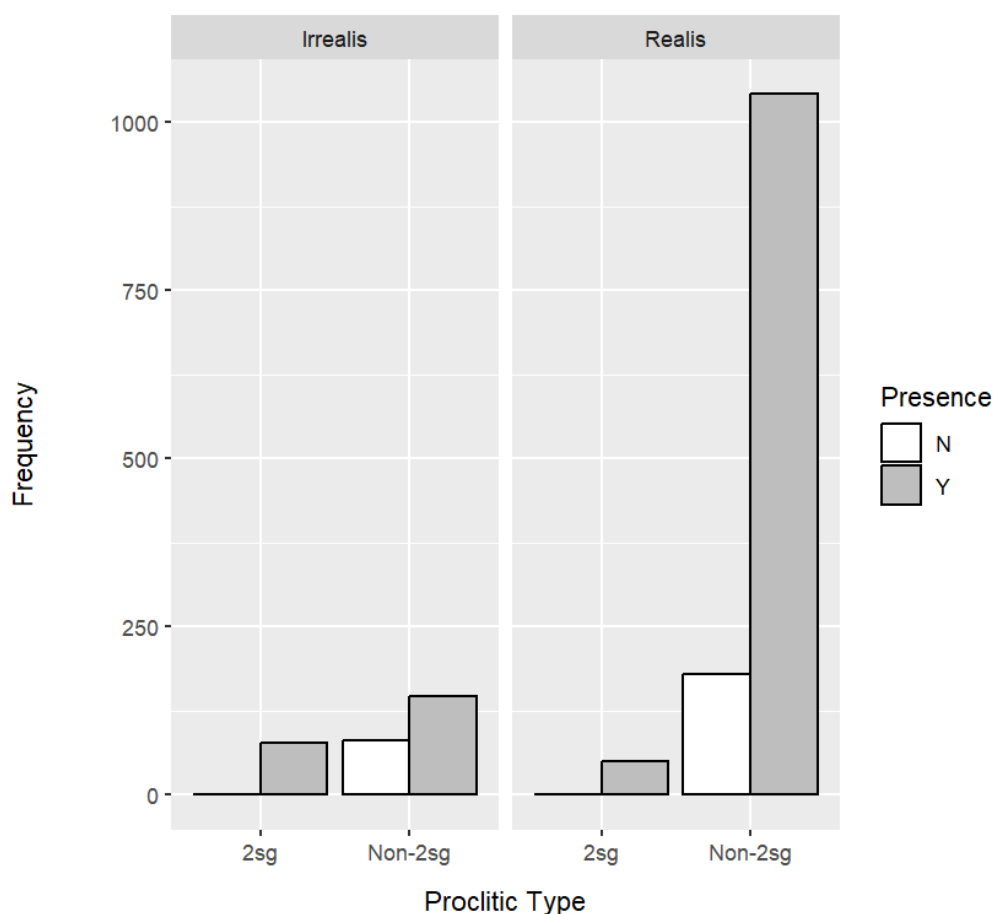


Figure 4-21: Presence of subject proclitics based on mood

However, the results do not reveal whether the higher frequency of proclitic deletion is the effect of mood or the irrealis particle *ga*. Drawing on the asymmetry in the mood of Lelepa subject proclitics, where only 2sg proclitics have distinct realis and irrealis forms while others use the irrealis marker *ga* to encode mood, a comparison between the presence and mood of 2sg and non-2sg is performed to test the effect.

Table 4-22: Presence of subject proclitics based on mood and proclitic type

Subject proclitic presence							
			N		Y		Total
Proclitic type and mood	2sg	IRR	1	1.28%	77	98.72%	78
	2sg	R	0	0.00%	49	100.00%	49
	Non-2sg	IRR	80	35.40%	146	64.60%	226
	Non-2sg	R	179	14.66%	1042	85.34%	1221
	Total	IRR	80	26.14%	226	73.86%	306
	Total	R	180	14.20%	1088	85.80%	1268
							1574

Proclitic presence based on mood and proclitic type**Figure 4-22: Presence of subject proclitics based on mood and proclitic type**

As is shown in Table 4-22 and Figure 4-22, proclitic deletion is very rare among 2sg subject referents regardless of their mood (see also § 4.5.6), moreover, the only instance of 2sg proclitic deletion is in a clause-chain, where the subject proclitic in the non-initial clause is inevitably deleted, as in example (60) below. Lacrampe (2014, p. 351) also mentions subject proclitic deletion takes place mostly when there is a preverbal particle. In addition, there is one instance where the irrealis marker *ga* is absent and the mood is indicated by the irrealis form of the verb, while 3du subject proclitic is still present, as in (61) below:

- (60) *ṗa=pea* *safe* *ga* *skei*
 2sg.IRR.NOM=be.first.R pick.up IRR one
panei, **(ṗa=)***trus=i* *to* *tan*
 COME **(2sg.IRR.NOM=)**put=3sg.ACC STAT down
 ‘Go and pick one up, put it on the ground.’
 (SL1-20080401DevilArtok-CMmut_106)
- (61) *ar=to* *pat=i*
 3du.NOM=IPFV make=3sg.ACC
tkani *lag* ***ar=(ga)*** ***fan [...]***
 like.this PURP **3du.NOM=(IRR)** **go.IRR**
 ‘They were doing this like this in order for them to go...’
 (SL1-20080401DevilArtok-CMmut_137)

Therefore, it is very likely the irrealis marker *ga* that relates to subject proclitic deletion. Furthermore, it is common (66.67%, n = 150) for subject proclitics hosted by an irrealis marker in the corpus that either the proclitic is deleted (34.22%, n = 77) or the irrealis particle *ga* is reduced to *g* (32.44%, n = 73). However, when it comes to *a=ga* '1sg.NOM=IRR' (n = 51), proclitic deletion

occurs much more often (72.55%, $n = 37$) than particle reduction (13.73%, $n = 7$). The notably higher frequency of 1sg proclitic deletion before irrealis marker *ga* possibly accounts for the lower proportion of 1sg proclitic presence mentioned in § 4.5.3 above. Example (62) below comprises an instance of 1sg proclitic deletion before the particle *ga*, by contrast, the non-1sg subject proclitics in (63) are not deleted regardless of whether the particle is produced as *ga* or *g*:

- (62) **(a=)***ga* *pe* *pa* *loṗa=e*
 (1sg.NOM=)IRR be.first.R go see=3sg.ACC
 'I will go look first!' (SL1-20080401DevilArtok-CMmut_027)

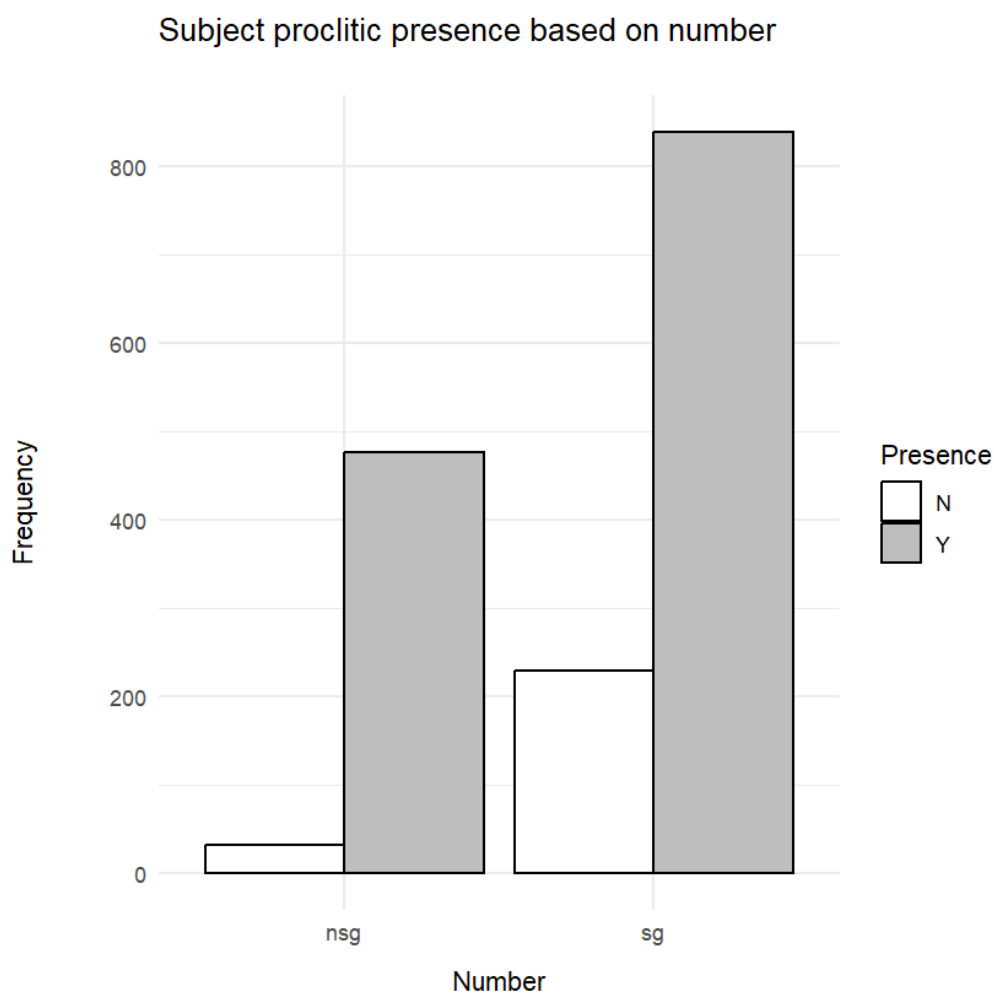
- (63) ***tu=g*** *mro* *pou*
 1pl.INCL.NOM=IRR REP weave.R
 narṗan *ke-ru* *pan* *pan* *pa,*
 side ORD-two GO GO GO
 e=ga *fa* *nou*
 3sg.NOM=IRR go.IRR finish
 'We will weave the second side on and on, (and) it will be finished.'
 (SL1-20080424MatWeaving-MRnn.020)

4.5.6 Person and number

The predictor *person and number* is ranked the fourth in terms of importance scores in the random forest analysis. Despite being excluded in the classification tree, descriptive statistical data shows that singular subject proclitics are more likely to be deleted (21.46%, $n = 229$) than dual and plural ones (6.11%, $n = 31$), as illustrated in Table 4-23 and Figure 4-23:

Table 4-23: Presence of subject proclitics based on number

<i>(nsg = non-singular)</i>		Subject proclitic presence				
<i>(sg = singular)</i>		N		Y		Total
Number	nsg	31	6.11%	476	93.89%	507
	sg	229	21.46%	838	78.54%	1067
	Total	260	16.52%	1314	83.48%	1574

**Figure 4-23: Presence of subject proclitics based on number**

Non-singular subject proclitic deletion also takes place very infrequently even when a coreferential lexical NP is present (4.55%, $n = 2$) compared with singular referents (85%, $n = 85$). Since Lelepa does not encode number marking on

nominals, plural proclitics may be used by speakers to quantify the nominal where necessary (see also § 5.3.1). For example, the 3pl subject proclitic *ur=* is present in (64) below when the coreferential lexical NP *mutuam* ‘devil’ denotes a plural referent while the 3sg subject proclitic *e=* is omitted in (65) where the quantity of the referent *mutuam* is clearly singular through the context and does not need to be specified:

- (64) ***mutuam*** *ur=mato*
 devil **3pl.NOM=IPFV**
 lega~lega-ki=ni *tapla*
 sing~RED-TR=3sg.ACC like.this
 ‘The devils kept on singing to him like this.’
 (SL1-20080424PeaceCeremony-MLnat_019)
- (65) *ar=pan* *pan* *pan* *se*
 3du.NOM=go GO GO while
 mutuam **(*e=*)*wan*** *matur* *tapla [...]*
 devil **(3sg.NOM=)IPFV** sleep like.this
 ‘They went on and on while the devil was sleeping like this [...].’
 (SL1-20080401DevilArtok-CMmut_139)

In addition, 52.38% (n = 11) of subject proclitics are deleted in verb complexes headed by a numeral, as exemplified in (66) below. Although the sample size was too small to test the hypothesis in this study, the proclitics are deleted possibly also in order to avoid redundancy arising from double expression of number. On the other hand, in cases where a subject proclitic is present when the basic clause requires a subject argument, as in (58) above, the proclitic is

usually singular even if the host numeral is plural³⁴, since the number has already been specified by the numeral.

- (66) *tu=g* *mter* *ṁulṁul*
 1pl.INCL.NOM=IRR draw circle
(e=)ga *skei* *e=ga*
 (3sg.NOM=)IRR one 3sg.NOM=IRR
to warange na tu=sra
 stay place REL 1pl.INCL.NOM=brush
psaki=ni to
 clean=3sg.ACC STAT
 ‘we’ll draw one circle in this place which we brushed clean.’
 (SL1-20080424Hunting-BTn_036)

However, when it comes to singular subject proclitics, as is mentioned in § 4.5.3, § 4.5.4, and § 4.5.5 above, deletion takes place predominantly in 1sg and 3sg referents (99.56, n = 227), while 2sg proclitics are virtually always present (99.21%, n = 126) except for the only instance in a clause-chain in (60) above. Furthermore, the proportion of 2sg proclitic deletion is even lower than plural referents. This also supports the hypothesis in § 4.5.5 above that a proclitic is in fact more likely to be deleted when it is hosted by an irrealis marker *ga* rather than simply in irrealis mood. Table 4-24 and Figure 4-24 present a comparison of subject proclitic deletion based between 2sg and non-2sg referents.

³⁴ See also § 5.4.2.

Table 4-24: Presence of subject proclitics based on whether the referent is 2sg

		Subject proclitic presence				
		N		Y		Total
Person and number	Non-2sg	259	17.90%	1188	82.10%	1447
	2sg	1	0.79%	126	99.21%	127
	Total	260	16.52%	1314	83.48%	1574

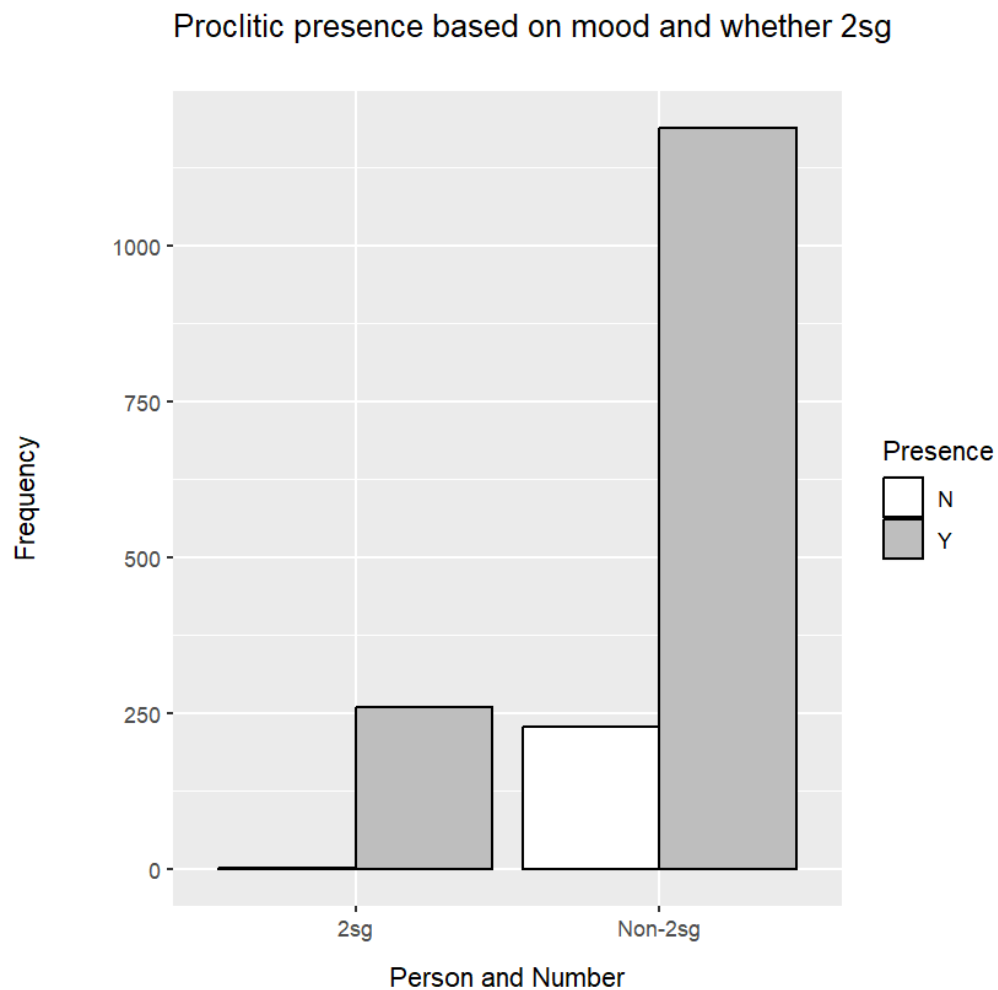


Figure 4-24: Presence of subject proclitics based on whether the referent is 2sg

Finally, to test the validity of these two observation hypotheses, an additional GLMM is performed with four binary predictors as fixed effects and speaker as a random effect. The four predictors include: mood (*r/irr*), number (*sg/ns*), the

presence of a lexical NP (lex.presence, Y/**N**), and whether the referent is second-person singular (2sg, Y/**N**)³⁵. As the results in Table 4-25 suggest, subject proclitics are significantly less likely to be deleted when the referent is 2sg or plural.

Table 4-25: Results of the hypothesis-testing GLMM analysis with significant predictors in bold

Random effect:					
	Variance	Std.Dev.			
Speaker (Intercept)	0.09973	0.3158			
Number of observations: 1574, groups by speaker: 13					
Fixed effects:					
	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	1.8892	0.2384	7.925	2.27E-15	***
moodR	1.6662	0.1988	8.382	< 2e-16	***
2sgY	3.29	0.6089	5.403	6.56E-08	***
numberSG	-1.6577	0.2201	-7.53	5.06E-14	***
lex.presenceY	-1.3043	0.1784	-7.31	2.67E-13	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1					

4.6 Summary

In conclusion, § 4.2 demonstrates that a free subject pronoun rarely occurs in Lelepa, but is more likely to be present when there is a switch in referent, or when the referent is human-like, deictic, or in direct speech. § 4.3 complements the findings in § 4.2 in that it shows a free pronoun is more likely to be present when it is a deictic speech-act participant, or when the corresponding proclitic

³⁵ The reference levels are in bold.

of a switch referent has an ambiguous form. Furthermore, no deictic referent is found having a sole lexical NP reference in the dataset, and instances of coreference with both a free subject pronoun and a lexical NP are very rare. In addition, the comparison between lexical and pronominal subject NPs in § 4.4 finds that an overt subject NP is more likely to occur when there is a re-introduced or new referent, although free pronouns are less common in this category than lexical NPs. In comparison, human-like and direct-speech subject referents prefer pronominal references to lexical NPs. Direct-speech subject referents are also exclusively related to the presence of free subject pronouns, whereas there is only a trend for human-like subject referents to favour pronominal NPs over subject proclitics. When it comes to subject proclitics, they are used much more frequently than free subject pronouns, while results in § 4.5 suggests that a subject proclitic is more likely to be deleted when it follows a vowel, it is first- or third-person singular, a coreferential lexical NP is present, or when it is hosted by an irrealis particle *ga*, especially when it is also 1sg. Moreover, 2sg subject proclitics are the most unlikely to be deleted.

Chapter 5 Discussion and conclusion

5.1 Overview

In addressing the research questions, this chapter discusses the implications of the findings in Chapter 4 for differential subject expressions in Lelepa, with reference to existing studies on referential choice across languages, and also discussing how the results fit into the cross-linguistic typology. Section 5.2 focuses on answering the first research questions about the distribution and functions of free subject pronouns in Lelepa, with § 5.2.1 on distribution and § 5.2.2 on functions. Section 5.3 turns to the second research questions on the distribution (§ 5.3.1) and syntactic status (§ 5.3.2) of subject proclitics in Lelepa. Based on both the statistical analyses and the discussion in Sections 5.2 and 5.3, a unified approach to referential choice for subjects in Lelepa in relation to information structure (§ 5.4.1) and redundancy (§ 5.4.2) is proposed in Section 5.4. Section 5.5 offers a concluding statement of the thesis with a summary of the findings, the significance and limitation of the study, and also suggestions for future research directions.

5.2 Free subject pronouns

5.2.1 Distribution

As detailed in Chapter 4, the results of quantitative analyses support Lacrampe's (2014) observation that free subject pronouns are overall very unlikely to be present in Lelepa. A targeted investigation has also revealed more nuanced patterns in the distribution of free subject pronouns in addition to Lacrampe's (2014) description. This subsection discusses the distribution of

free subject pronouns and suggests that the overall distribution of free subject pronouns in Lelepa accords well with the patterns identified in many other languages.

Lacrampe (2014, pp. 213–215) finds that lexical NPs in Lelepa, such as *neika* ‘fish’, are used to reference either new, re-introduced, or continued referents. Results of a systematic investigation in this thesis have revealed that free subject pronouns like *konou* ‘1sg’ are also used in all three scenarios. As outlined in § 4.2.2, § 4.3.2, and § 4.4.2, free subject pronouns are most commonly used when the subject switches to a referent that has been introduced to the discourse already. This is similar to findings for Spanish and Portuguese, where switch reference also conditions the presence of free subject pronouns (Amaral & Schwenter, 2005; Cameron, 1995; Jacobson, 2016; Lastra & Butragueño, 2015; Posio, 2011, 2013; Torres Cacoullos & Travis, 2019; Travis & Torres Cacoullos, 2012 *inter alia*). In addition, results of Study 2 show that free subject pronouns are significantly more likely to be present when there is a referent switch and the corresponding proclitic has an ambiguous form, for example, *ur*= ‘1pl.excl/3pl’. In multiple varieties of Spanish, free pronouns are also unlikely to be present when the verbal agreement marker is morphologically ambiguous unless there is a referent switch, since the referent can mostly be easily understood despite morphological ambiguity in natural speech (Carvalho & Child, 2011; Lastra & Butragueño, 2015; Shin, 2014 *inter alia*). Free pronouns are also likely to be present when the referent is mentioned for the first time in discourse, although this usage is relatively uncommon in the present dataset and is typically limited to speech-act participants, whereas lexical NPs are preferred by third-person new referents (see also § 4.4.3). This pattern is strikingly similar to Carvalho & Child’s (2011

inter alia) observation of Spanish referential choice. As Mithun (1990) points out, the use of third-person pronouns is avoided when the referent is out of an interlocutor's consciousness. Recalling the results in Studies 1 to 3, this constraint accounts for both the higher likelihood of free subject pronoun presence for deictic referents, of which interlocutors are always conscious, and the preference of new subject referents for lexical NP references. Furthermore, while Lacrampe (2014, p. 214) finds that lexical NPs can also encode a continued referent for topic in Lelepa, results of Study 3 have extended this proposal and demonstrate that free pronouns are also allowed for continued subject referents, as shown in example (37) above in § 4.4.2. Moreover, continued subject referents even prefer pronominal over lexical noun phrase references according to the corpus data, although null NPs or bare subject proclitics are still the most chosen referential expression for continued subjects in Lelepa (see also § 4.4.2). This also accords with the results of Givón's (1983a) cross-linguistic survey, as well as Arka & Sedeng's (2018) observation in Sembiran Balinese, an Austronesian language like Lelepa.

In addition, the correlation between pronominal references and human(-like) referents in Lelepa (see also § 4.2.3 and § 4.4.5) clearly demonstrates the cross-linguistic validity of Hawkinson & Hyman's (1974) topicality hierarchy, which states that human and given referents tend to be pronominalised. The finding of this study also aligns with Genetti & Crain's (2003) *animate pronoun* constraint, which suggests that inanimate referents tend to have lexical references (see also Meakins, 2015). Schnell & Barth (2020) also find that human referents prefer free pronoun references over zero forms in the Vera'a language of northern Vanuatu. This study shows that, while true zero anaphora

is rare³⁶ in Lelepa, human referents also tend to have free pronoun references over subject proclitics. There is also a trend towards free pronoun references for human(-like) references compared with subject proclitics between the two non-lexical subject expressions in Lelepa, and the effect of speech-act participants (SAPs) may be helpful to account for this trend.

SAPs are found interacting with multiple influencing factors on free subject pronoun presence in Lelepa according to the results. Section 4.3.3 reveals that deictic speech-act participant referents are the most likely to have free subject pronoun references. In addition, as is briefly discussed in § 4.4.4 and § 4.4.5, human(-like) and direct-speech referents, both with higher proportions of speech-act participants, also tend to have free pronoun references. This study also supports Torres Cacoullos & Travis's (2018, pp. 88–89) observation that free pronouns are preferred in direct speech, especially referring to SAPs, with quantitative data from Lelepa. Furthermore, Schnell (2018b) hypothesizes that it is the deictic and shifting nature of SAP subject referents that accounts for their preference towards overt realisation in Vera'a. The higher likelihood in Study 2 of free subject pronoun presence for deictic speech-act participants, especially when there is a referent switch, supports this hypothesis and suggests its potential universality across Oceanic languages.

5.2.2 Functions

In addition to the disambiguation function mentioned in § 4.3.3 and § 5.2.1 above, the presence of a free subject pronoun in Lelepa can also be purely

³⁶ See 5.3.2 below for a proposal on treating all proclitics as pronouns in Lelepa.

pragmatic. This subsection discusses the pragmatic functions that account for the presence of free subject pronouns in difference instances with comparison to similar functions of free pronouns in other languages.

The most common pragmatic function of free subject pronouns is topic marking. The discussion here follows Frascarelli and Hinterhölzl's (2007, p. 89) tripartite division of topics: *shifting topic*, *contrastive topic*, and *familiar topic*. Lacrampe (2014, p. 214) mentions the use of free subject pronouns to establish a shifting topic when the referent is new or re-introduced in Lelepa, and this function is supported by evidence from corpus and quantitative data in this study. This study also provides examples of free pronouns used as familiar topics for topic continuity building on Lacrampe's (2014) observation (see also Sections 4.2.2, 4.4.2, and 5.2.1). Lacrampe (2014) also makes a distinction between contrastive topics and shifting or continued topics in Lelepa. According to Lacrampe (2014, pp. 120, 146, 191, 215–216), the referent of a free subject pronoun is marked as a contrastive topic³⁷ when the pronominal NP is left-dislocated or modified with topic marker *wei* or demonstrative *na*. Results of the corpus study nonetheless suggest the presence of free subject pronouns itself can encode contrast without left-dislocation or a topic particle (see also § 2.4 and § 4.2.2). Furthermore, a shifting topic can also be left-dislocated, as in (57) above in § 4.5.4.

³⁷ Existing studies also distinguish contrastive topic and contrastive focus in that if the referent is new then it is a contrastive focus (e.g. Arka & Sedeng, 2018). However, no instance of subject NP marking a new referent for contrast is found in the current corpus.

The use of optional free subject pronouns may also indicate floor shift in Lelepa. Examples (38) and (39) above in § 4.4.3 illustrate the use of 1sg free subject pronouns at the beginning of narratives where the 1sg referent is highly salient. The use of free subject pronouns as floor-shifting device in conversations is widely observed in pro-drop languages like Spanish (Davidson, 1996; Travis & Torres Cacoullos, 2012), Japanese (D.-Y. Lee & Yonezawa, 2008), Korean (N. Lee, 2021), and Javanese (Ewing, 2014). Therefore, it is possible that Lelepa speakers' use of free subject pronouns at the beginning of storytelling when they take over the role of the speaker yields a floor and introduces the new discourse topic, i.e. the narrative(s) they are making. Moreover, this floor-shifting function may also account for the higher likelihood of free subject pronoun presence in direct quoted speech, where narrators attempt to replicate a different conversation context. Since free subject pronouns are only used as floor-shifting devices in direct rather than indirect reported speech, their presence also distinguishes direct speech from indirect speech when the subject of the reported speech is first-person singular, especially in personal history narratives (see also § 4.2.5). However, since all the recordings this thesis draws on are monologues, a follow-up study of free subject pronoun variations in Lelepa conversations is necessary to examine whether free subject pronouns marks floor shifts only in direct quoted speech as a rhetorical device in storytelling to guide the listener's attention or also in everyday conversations.

Another pragmatic function of free subject pronouns in Lelepa maps onto *attention-getter*, which is similar to floor-shifting and commonly found in direct speech. For instance, in examples (16) in § 4.2.2 and (23) in § 4.2.5, the free subject pronouns *nag* and *kon(ou)* are present while the person and number values have already been expressed through the proclitics *ḡa=* and *a=*. The

presence of free pronouns as attention-getters in these cases puts an emphasis on the proposition of the clause instead of the referent. A similar function is observed in the use of 1sg pronoun topic *qá mà* in Longxi Qiang (see also Section 2.2.3), however, the 1sg pronoun *qá mà* as a pragmatic marker has been semantically bleached and is used for non-1sg referents as well (Zheng, 2020). The Lelepa examples suggest that free subject pronouns in Lelepa still have their referential property even as attention-getters, as they agree in person and number with the actual referents.

5.3 Subject proclitics

5.3.1 Distribution

Although subject proclitics are clearly the most chosen subject expression in Lelepa, results of Study 4 (see § 4.5) suggest that subject proclitics are not obligatory. Instead, they are deleted 16.52% of the time in the corpus, extending on Lacrampe's (2014) observation that proclitic deletion is possible. This subsection concentrates on the implications of subject proclitic distribution outlined in § 4.5 with reference to existing studies.

Firstly, subject proclitics are categorically absent in clause-chains in Lelepa, as noted in § 4.5.3 and § 4.5.5. Lelepa clause-chaining is similar to Nafsan (Thieberger, 2006, pp. 324–327) in that clause-chaining in both languages involves the omission of subject proclitics in non-initial clauses and clauses are prosodically connected in a clause-chain. In addition, examples (48) and (49) in § 4.5.3 exemplify head-tail linkage in Lelepa, which is also commonly attested in other Oceanic languages in Vanuatu (see also § 2.4). However, while Thieberger's (2006) examples only show subject proclitic deletion when tail-

head linkage takes place within a clause chain, the Lelepa example additionally suggests subject proclitic deletion also takes place in tail-head linkage within an independent clause in (49).

Section 4.5.4 notes subject proclitics are likely to be deleted where a coreferential lexical NP is present, especially when the referent (and their quantity) is highly accessible, possibly to also avoid unnecessary redundancy (see also § 5.4.2). However, plural proclitics are rarely deleted as number is not encoded in Lelepa nominal morphology, and proclitics can specify number and hence are not redundant. Conominated singular subject proclitics in Lelepa are also non-redundant when there is a need to specify the singularity of the referent. Hence, subject proclitics are retained when the number of the referent is inaccessible. Zapotec, an American language with no number marking on its nominals, exhibits a similar innovation in that proclitics are used to quantify the referent (Black, 2000). Subject proclitics are also partially redundant in clauses headed by a numeral which specifies the quantity already, and they tend to be either dropped or realised as a non-truth-conditional 3sg proclitic (see § 4.5.6 and example (58) in § 4.5.4). The use of a 3sg proclitic in this case also displays *disagreement* as there is a mismatch between the morphological and semantic number of the referent, similar to collective nouns in English like *committee* whose semantic number is plural but the verbal agreement can be 3sg (Sturt, 2022).

Furthermore, § 4.5.5 mentions the tendency for the 1sg subject proclitic *a=* to be deleted before the irrealis marker *ga*, while the common pattern for other proclitics is for *ga* to lose its final vowel. However, it is possible that the *ga* form in this case is a result of metathesis and may be analysed as *g=a*

‘IRR=1sg.NOM’, similar to Nafsan (Thieberger, 2006), whose irrealis marker *k*³⁸ is suffixed to proclitics with an onset and prefixed to proclitics with zero onset³⁹. As evidence from historical comparative accounts shows, the reconstructed proto-Efatic *q is reflected in Nafsan as *k* but in Nguna and Lelepa as *g*⁴⁰ (R. Clark, 1985, p. 11). It is therefore reasonable to assume that the Nafsan irrealis particle *k* and the Lelepa irrealis particle *g(a)* are cognates, and the realisation of 1sg.IRR as *ga* may be a sign of language change in progress, given the striking similarities in language change between Nafsan and Lelepa (cf. Billington et al., 2020; Lacrampe, 2012). In addition, the metathesis of irrealis marker attaching to zero-onset subject proclitics is also observed in other Oceanic languages like Wogeo, an Oceanic language of northeast New Guinea (Exter, 2012)⁴¹. This hypothesis is, however, not well supported, since no 3sg.IRR proclitic is realised as **g*=*e* in the corpus. A follow-up investigation of more recent Lelepa narratives is necessary before assuming a diachronic change.

Finally, the more fusional proclitics are also less likely to be deleted in Lelepa. § 4.5.5 notes that 2sg subject proclitics are rarely deleted, possibly because they are portmanteau morphemes encoding not only the grammatical relation, person, and number of the referent, but also the mood of the clause, while other subject proclitics, no matter whether plural or not, rely on the independent irrealis particle *ga* to express mood. Recalling Neeleman & Szendrői’s (2005),

³⁸ Except for the 2sg proclitic whose irrealis form is non-analysable like in Lelepa.

³⁹ For example, *k-a* ‘IRR-1sg’ versus *ta-k* ‘1du.incl-IRR’ (Krajinović, 2017; Thieberger, 2006).

⁴⁰ The <*g*> refers to the velar nasal /ŋ/ rather than plosive.

⁴¹ For example, *g-o* ‘IRR-1sg’ versus *to-g* ‘1du-IRR’.

proposal on the correlation between fusional case morphology and non-pro-drop, results of this study suggest fusionality may also have a similar effect on subject proclitic variation in Lelepa, as 2sg proclitics with fusional modality marking are less likely to be deleted than other proclitics with agglutinating modality marking.

Apart from deletion, Section 4.3.3 also mentions the exclusive/anaphoric use of first-person plural inclusive proclitic *tu=*. One tentative explanation for the exclusive use of inclusive proclitics relates the metaphorical use of inclusive pronouns in Bislama, the lingua franca in Vanuatu which is also spoken by most Lelepa speakers. According to Meyerhoff (1998), Bislama speakers may use inclusive pronouns in exclusive cases to signify an inclusion in the mental domain, for example, a shared identity, rather than the real world⁴². In addition, exclusive uses of inclusive proclitics observed in the current corpus are all from procedural narratives and *naluokia* (see also § 3.2 for definitions). It is possible that Lelepa speakers use inclusive proclitics metaphorically because they regard the listener(s) as a member of their community and would like to include them in their routines and instructions. Alternatively, it is also possible that the speakers refer to first-person plural exclusive referents using the inclusive proclitic form in order to resolve the ambiguity between 1pl.excl and 3pl subject proclitics. Evidence for this hypothesis comes from one recording⁴³ where the clausal subjects constantly switch between 1pl.excl and 3pl referents with the same proclitic *ur=*. The speaker uses free pronouns to distinguish these two

⁴² It is possible that this usage in Bislama is influenced by pre-existing patterns in Vanuatu's indigenous languages, that have just not been investigated yet, seeing that Bislama syntax is influenced by Oceanic languages (cf. Camden, 1979)

⁴³ This recording is *SL1-20080424PeaceCeremony*.

referents at first, and then starts to use the 1pl.incl proclitic *tu=* instead to denote the first-person plural referent and maintains it for the rest of the recording, with no change in the referential entity. Consequently, no first-person plural free pronoun has been used in the narrative since the switch in proclitic form, hinting that the use of inclusive proclitics may bear a comparable function to free pronouns in this case. This switch in subject expressions is manifested in examples (31) – (35) in § 4.3.3 above.

5.3.2 Status

As to whether the argument indexes in a language are agreement markers or pronouns, Corbett (2003, pp. 188–189) comments that only agreement markers can occur more than once in a clause. This criterion is nonetheless unsatisfactory for Lelepa subject proclitics since the definition of a basic *clause* in Lelepa (Lacrampe, 2014, p. 237) is that it cannot have more than one subject proclitic. While Lacrampe (2014, p. 244) points out the possibility of analysing conominated subject proclitics as anaphoric agreement markers, I propose that subject proclitics are in fact consistently (bound) pronouns. This subsection presents my proposal based on the implication of the results.

As illustrated in § 4.5.4, the cooccurrence of a subject proclitic and a subject NP is constrained in Lelepa. When the coreferential subject NP is headed by a pronoun, either personal or demonstrative, linear reanalysis takes place and the proclitic is then hosted by the pronoun, forming a single referential expression. In addition, pronouns ending with vowels also go through different degrees of final vowel reduction in this process, for example, *kon=a* ‘1sg=1sg’. When it comes to proclitics coreferential with lexical NPs, they are commonly deleted except for cases where there is a need to specify the number of the referent

(e.g. example (64) in § 4.5.6). While Mithun (2003) mentions that a common pattern of grammatical change is for free pronouns to evolve into bound forms, and then into redundant agreement markers, the conominated proclitic variation in Lelepa in fact avoids redundant subject expressions (see also § 5.4.2), hence, subject proclitics in Lelepa should be consistently treated as bound pronouns. Moreover, since Haspelmath's (2013) definition indicates that only nominals within a narrow clause⁴⁴ count as conominals, subject proclitic variation may in turn be used to identify left-dislocated subject NPs, which are prosodically independent and unreduced with coreferential proclitics on the verb complex, if the proclitics are treated as pronouns and hence unlikely to allow conominals.

The presence of expletive proclitics in Lelepa also supports this bound pronoun hypothesis. Commonly found in existential clauses during preliminary observation of the corpus data (see also § 3.5.2), expletive proclitics in Lelepa do not agree with an anaphoric referent in person and number. Therefore, they are non-referential, and the presence of an expletive subject proclitic is only a placeholder to satisfy the structural requirement for a Lelepa clause to have a subject. Furthermore, the referentiality of a proclitic does not have a very significant effect on its presence or deletion, according to the results of the random forest (see also § 4.5.1), hence, expletive proclitics are not more likely to be deleted referential proclitics, although instances are still observed. A potential counterargument to this proposal may come from Italian, a pro-drop language with rich agreement morphology and null expletive subjects. In Standard Italian, verbs inflect for number existential clauses, whereas in Northern Italian varieties the verbs in existential clauses are always singular

⁴⁴ This is equivalent to a *basic clause* in Lelepa (Lacrampe, 2014, p. 244)

(Cardinaletti, 1997; Pescarini, 2022). Formalist analyses tend to treat the null expletive subject as silent arguments (Cardinaletti, 1997). From this perspective, verbs in existential clauses with null subject agree with the associate in Standard Italian, while they agree with the covert expletive in Northern varieties. This proposal suggests while the expletive can be null, the agreement marker should be categorically present. However, existential subject proclitics in Lelepa, like other proclitics, are optionally deleted. Hence, this thesis adopts the pronoun analysis based on the findings of the quantitative studies. Future studies from a formalist account may help elucidate the status of Lelepa existential proclitics.

Finally, although the omission of free subject pronouns in Oceanic languages is commonly regarded as pro-drop (e.g. Polinsky & Potsdam, to appear), based on this proposal, Lelepa is not a pro-drop language. Subject pronouns are mostly present in Lelepa, in the form of either a proclitic, a free pronoun, or a coalesced coreferential pronoun. However, null subjects are also observed when the referent is highly salient. Cross-linguistic evidence also suggests that pronouns are still able to be omitted in non-pro-drop languages like English, and the constraints for pronoun omission in non-pro-drop languages are found to be very similar to those in pro-drop languages like Spanish (e.g. Torres Cacoullos & Travis, 2019 *inter alia*).

5.4 Referential choice

5.4.1 Referential choice and information structure

This section aims to provide a unified approach to referential choice for subjects in Lelepa. Subject expressions involved in this section are genuine null subjects

(as in clause-chains), subject proclitics, free personal pronouns, and lexical NPs, recalling that the size of data does not allow for taking the distribution of demonstrative pronouns in Lelepa into consideration. This subsection summarises referential choice based on information structure.

Multiple studies on the effect of information structure have acknowledged the cross-linguistic trend that the more *predictable* or *accessible* referents tend to be less marked (e.g. Ariel, 1991; Givón, 1983b; Gundel et al., 1993). In addition, according to the topicality hierarchy theory, human, animate, singular, given or speech-act participant referents are more salient and more accessible to pronominalisation (Dunbar, 1983; Duranti, 1979; Hawkinson & Hyman, 1974 *inter alia*). However, this study only finds that the effect of humanness and discourse anaphora in Lelepa accords with the theory, while the preferred subject expressions for referents with different person and number are idiosyncratic and cannot be hierarchically mapped (see also § 4.3.3, § 4.5.6, and § 5.3.1). Furthermore, the findings additionally demonstrate that humanness conditions referential choice at two different stages: human referents prefer the less-marked non-lexical forms at first, and then tend to have overt free pronoun realisations, suggesting a hierarchical analysis like the accessibility hierarchy is also not satisfactory in predicting the subject expressions based on humanness. Table 5-1 presents the referential choice of subjects in Lelepa based on the accessibility (Ariel, 1991, p. 449), or predictability (Givón, 1983c, p. 359), of referents. In the table, there are four options for anaphoric referents and three for deictic referents, as they are not found with a sole lexical NP reference in the corpus. Null subject refers to the absence of an overt subject expression for the most continuous referents in clause-chains or tail-head linkage. Subject proclitic indicates a sole proclitic with

no coreferential NP. Expletive subjects are excluded from this table because they only have proclitic or, less commonly, zero references.

Table 5-1: Preferred subject expressions in Lelepa based on accessibility

Anaphoric	Null subject	Subject proclitic	Free pronoun	Lexical NP
Deictic	Null subject	Subject proclitic	Free pronoun	
<– Unmarked			Marked –>	
Ambiguity	No	Contextual	Morphological + contextual	
Givenness	Highly continuous	Continuous	Re-introduced	New
<– Accessible			Inaccessible –>	

Essentially, a referent can be referenced with subject expressions on the same level or to its right in the hierarchy where applicable. Although cross-referencing normally would not lead to any misunderstanding in this case, the results suggest overt subject NPs are usually avoided as they would still lead to unnecessary redundancy.

5.4.2 Referential choice and redundancy

Section 5.4.1 above summarises how Lelepa speakers' choice of differential subject expressions in the corpus associates with information structure. This subsection extends the discussion to how Lelepa speakers' referential choice relates to redundancy. Previous chapters and sections in this thesis suggest redundant subject expressions are only present for semantic or pragmatic purposes and are disallowed elsewhere in Lelepa. The analysis presented below will draw on both discourse and semantic redundancy (see also § 2.2 for the distinction).

When it comes to semantic redundancy, Leufkens (2020, pp. 87, 95) notes a partially redundant lexical item may be used to complement the information provided by a grammatical marker. This proposal sheds light on measuring the degrees of semantic redundancy of Lelepa subject expressions based on the additional information they provide, in order to account for the final realisation of subject reference. According to Sections 4.3.3 and 4.5.6, the semantic functions of redundancy in Lelepa subject expressions are observed only for the partially redundant references which provide additional information. Coreferential subject expressions providing no additional information are regarded as totally redundant, hence only present with pragmatic functions. Based on the usage frequency and distribution of differential subject references unveiled in this study, a prototypical redundancy hierarchy of subject expressions is proposed below in Table 5-2:

Table 5-2: Prototypical semantic redundancy hierarchy of subject expressions in Lelepa

Free pronoun	Subject proclitic	Lexical NP
<– Redundant		Informative –>

With subject proclitics being the default in Lelepa, the presence of a coreferential free subject pronoun is prototypically redundant as the information it provides overlaps with the proclitic and it lacks a dual distinction in number. In Vera'a, overt number agreement on the verb also relates to null subject pronoun so that the realisation is not redundant (Schnell & Barth, 2020). However, free pronouns are also informative when cross-referencing proclitics with ambiguous forms, i.e. *ur*= '1pl.excl/3pl' and *ar*= '1du.excl/3du'. As illustrated in Table 5-3, the free pronoun specifies the person and the proclitic specifies the number of

the referent. Therefore, both subject expressions are present to resolve ambiguity despite being partially redundant on the information ‘*the referent is non-second-person and non-singular*’. In addition, this hierarchy accounts for not only free pronoun deletion but also reduction, since free pronouns cross-referencing proclitics are prosodically unaccented, commonly undergo final-vowel reduction, and form prosodic words with the proclitics (see also § 4.5.4). As suggested by Nichols (2018 *inter alia*) in the *complementarity hypothesis*, free pronouns and argument indexes tend to be in complementary distribution. Findings of this thesis suggest another possible realisation of the cooccurrence between free subject pronouns and subject proclitics from Lelepa data, i.e. prosodic coalescence.

Table 5-3: Person and number values of 1du.excl, 1pl.excl, 3du, and 3pl free pronouns and subject proclitics

		Person			
		1	3	1	3
Number	du	kenem	naara	ar=	
	pl(.excl)			ur=	

When a coreferential lexical NP is present within the basic clause, the proclitic tends to be considered superfluous as the referent has been specified by the NP already, and hence the proclitic is commonly found deleted in the corpus (see Sections 4.5.4 and 5.3.1). An exception is that when there is a need to quantify the referent, proclitics are present despite redundantly expressing the referent’s person and grammatical relation. However, as mentioned in § 4.5.4, left-dislocated subject NPs always (n = 18) leave a coreferential proclitic trace in the basic clause, therefore, it is reasonable to consider the basic clause as

the smallest unit of semantic redundancy detection in Lelepa as cross-references of left-dislocated subjects are not avoided.

However, the presence of the semantically informative expressions is also conditioned by discourse redundancy. While Section 5.4.1 suggests the optimal choice for unambiguous references, Section 5.2.1 summarises that Lelepa speakers primarily use subject NPs when there is a referent switch. In addition, in the few instances of subject NPs present for continued referents, the semantically less informative free pronouns are preferred to lexical NPs. Moreover, coreferential expressions are not more prevalent in cases of morphological ambiguity, such as *ur*= '1pl.excl/3pl', than elsewhere, unless there is also a referent switch and the referent has then become contextually ambiguous. In addition, the 3sg proclitic is also found deleted despite additionally encoding the number for the lexical NP when the number of the referent is accessible (see also Sections 4.5.4, 4.5.6, and 5.3.1). Therefore, discourse redundancy is avoided even if the expression is semantically partial- or non-redundant.

To sum up, both discourse and semantic redundancy conditions referential choice in Lelepa. Subjects are not (cross-)referenced with either discourse- or semantic-redundant expressions in most cases. However, free subject pronouns with pragmatic functions can still be unconditionally present, as discussed in 5.2.2 above. Based on the findings, three constraints are proposed in a – c below. A cross-reference in Lelepa should not violate constraints a or b, unless it abides by c, based on the corpus data:

- a. DISCOURSE REDUNDANCY: Avoid discourse redundancy for cross-references.

- b. **SEMANTIC REDUNDANCY:** Avoid total⁴⁵ semantic redundancy for cross-references within a basic clause.
- c. **PRAGMATIC FREE PRONOUNS:** A redundant free pronoun may be present if pragmatically licenced, whereas a redundant lexical NP is not allowed in this case.

Finally, the redundancy resolution based on constraints a – c for cross-references between a subject NP and a proclitic within basic clauses in Lelepa can be formulated as in Table 5-4 below:

Table 5-4: Redundancy resolution for cross-references between a subject NP and a proclitic in Lelepa

Input	Pronoun + proclitic			
Scenario	Switch referents with ambiguous proclitic forms	Continuous referents with ambiguous proclitic forms	Elsewhere	
Redundant expression	No	Pronoun	Pronoun	Pronoun
Discourse redundancy	No	Yes	Yes	Yes
Semantic redundancy	No	No	Yes	Yes
Discourse function	Yes/No ⁴⁶	No	Topic-marking, floor-shifting, attention-getting	No
Constraint violated	-	a, c	a, b	a, b, c
Output	Reduced pronoun + proclitic	Proclitic	Reduced pronoun + proclitic	Proclitic

⁴⁵ However, partial redundancy is allowed.

⁴⁶ This indicates the referential expression would be realised as follows whether or not it has a pragmatic function, henceforth.

Input	Lexical NP + proclitic			
Scenario	Unpredictable referents with unspecified quantity	Predictable referents		Elsewhere
Redundant expression	No	Lexical NP	Lexical NP	Proclitic
Discourse redundancy	No	Yes	Yes	Yes
Semantic redundancy	No	No	No	No
Discourse function	Yes/No	Topic-marking	No	No
Constraint violated	-	a, c	a, c	a, c
Output	Lexical NP + proclitic	Reduced pronoun + proclitic	Proclitic	Lexical NP

5.5 Conclusion

What this thesis has presented is nonetheless merely the tip of the iceberg and awaits future research in various areas. This current corpus is suitable for more fine-grained analyses based on either a specific subset of data, such as human referents, or additional variables, for example, the effect of anaphoric distance⁴⁷ or ergativity⁴⁸, or the correlation between redundancy and speech rate⁴⁹, which may depict a more diverse pattern of referential choice beyond that presented in this thesis. Future studies will also benefit from a larger amount of data. Studies drawing on longitudinal data by additionally including either older materials, more recent speech data, or both may help reveal the variation in

⁴⁷ cf. Givón (1983b).

⁴⁸ cf. Du Bois (1987).

⁴⁹ cf. Lieberman (1963).

usage frequency of the first-person plural exclusive pronoun *kenem* and third-person plural pronoun *naara* before and after the *ur*= ‘1pl.excl/3pl’ syncretism, or allow an examination of whether the hypothetical *ga/g* ‘IRR’ metathesis has gradually taken place in zero-onset proclitics. An investigation of subject expressions in Lelepa conversations would also be preferable to test the findings of this thesis, based on monologues. Finally, cross-linguistic typological and theoretical analysis will effectively extend the implications of the findings to a larger scope. Similar studies on other languages with or without verbal argument indexes can help demonstrate whether the proposal of redundancy resolution in this thesis is universal across languages or language-specific. It is also worth investigating subject proclitics in Lelepa existential clauses from a formalist perspective and offering a final suggestion on whether they are expletive pronouns or agreement markers.

On the other hand, this thesis constitutes the first systematic investigation of differential subject expressions in Lelepa from a frequentist account. The findings support Lacrampe’s (2014) observation that subject NPs, especially free pronouns are uncommon in Lelepa, and that subject proclitics are the most prevalent subject expression. Referential choice in Lelepa revealed in this study also displays a strong conformity to other languages with predominant argument indexes and optional subject NPs. As for free subject pronouns, the thorough quantitative analyses additionally show that, apart from simply being optional, free subject pronouns are commonly used for switch, human(-like), direct-speech, and deictic SAP referents for pragmatic purposes like topic-marking, floor-shifting, and attention-getting. Free subject pronouns are also commonly present to resolve ambiguity arising from homonymic proclitics. When it comes to subject proclitics, the results indicate that they are very likely

optional when totally redundant. The distribution of subject proclitics also suggests they should be considered as syntactic arguments rather than grammatical agreement markers, based on which a proposal that Lelepa is not a pro-drop language is made. Overall, this thesis expands upon Lacrampe's (2014) description and finds that referential choice for subjects in Lelepa exhibit a clear pattern that speakers tend to overtly express the subjects where ambiguous or unpredictable, while unnecessary redundancy is consistently avoided. In addition, the undersampling and cross-validation methods used in Study 1 demonstrate practical approaches to performing quantitative analyses of small and/or unbalanced datasets, a common challenge in quantifying under-documented languages or peripheral structures, and may help identify typologically uncommon patterns and variations. The findings extend our understandings of differential subject expressions and pro-drop cross-linguistically, by contributing new data from Lelepa, an understudied Oceanic language. These findings also contribute to the typology of referential choice, especially in languages with obligatory person indexes which are commonly found in Oceania, by innovatively drawing a connection between multiple subject expressions and the theory of redundancy.

References

- Amaral, P., & Schwenter, S. A. (2005). Contrast and the (non-) occurrence of subject pronouns. In D. Eddington (Ed.), *Selected Proceedings of the 7th Hispanic Linguistics Symposium* (pp. 116–127). Cascadia Proceedings Project.
- Ariel, M. (1991). The function of accessibility in a theory of grammar. *Journal of Pragmatics*, 16(5), 443–463. [https://doi.org/10.1016/0378-2166\(91\)90136-L](https://doi.org/10.1016/0378-2166(91)90136-L)
- Ariel, M. (2001). Accessibility theory: An overview. *Text Representation: Linguistic and Psycholinguistic Aspects*, 8, 29–87.
- Arka, I. W. (2003). *Balinese morphosyntax: A lexical-functional approach*. Pacific Linguistics.
- Arka, I. W., & Sedeng, I. N. (2018). Information structure in Sembiran Balinese. In S. Riesberg, A. Shiohara, & A. Utsumi (Eds.), *Perspectives on information structure in Austronesian languages* (pp. 139–175). Language Science Press.
- Arnold, J. E., & Zerkle, S. A. (2019). Why do people produce pronouns? Pragmatic selection vs. rational models. *Language, Cognition and Neuroscience*, 34(9), 1152–1175. <https://doi.org/10.1080/23273798.2019.1636103>
- Aznar, J., & Gala, N. (2020). *The Nisvai Corpus of Oral Narrative Practices from Malekula (Vanuatu) and its Associated Language Resources*. 2649–2656.
- Barbosa, P. P. (2019). pro as a Minimal nP: Toward a Unified Approach to Pro-Drop. *Linguistic Inquiry*, 50(3), 487–526. https://doi.org/10.1162/ling_a_00312

- Bates, D., Maechler, M., Bolker, B., & Walker, S. (2023). *lme4: Linear Mixed-Effects Models using “Eigen” and S4* (1.1-34) [Computer software].
<https://cran.r-project.org/web/packages/lme4/index.html>
- Beaver, D. I. (2004). The optimization of discourse anaphora. *Linguistics and Philosophy*, 27, 3–56.
- Berman, S., & Hestvik, A. (1997). Split antecedents, noncoreference, and DRT. *Atomism and Binding*, 1–29.
- Biberauer, T. (2010). Semi null-subject languages, expletives and expletive pro reconsidered. *Parametric Variation: Null Subjects in Minimalist Theory*, 153–199.
- Billington, R., Stoakes, H., & Thieberger, N. A. (2021). The Pacific Expansion: Optimizing phonetic transcription of archival corpora. *Proc. Interspeech 2021*, 4029–4033.
- Billington, R., Thieberger, N., & Fletcher, J. (2020). Phonetic evidence for phonotactic change in Nafsan (South Efate). *Ital. J. Linguist. (Special Issue: The Dynamics of Vowel Reduction and Loss: Phonetic Mechanisms and Phonological Outcomes)*, 125–150.
- Black, C. A. (2000). *Number marking innovations in Zapotec*. The Indigenous Voice of Oaxaca conference, University of California, Los Angeles.
- Bleotu, A. C. (2012). *Why does IT always rain on me? On weather verbs*. 59–81.
- Bloomfield, L. (1984). *Language*. University of Chicago Press.
- Blust, R. A. (2003). *Thao dictionary*. Institute of Linguistics (Preparatory Office), Academic Sinica.
- Bresnan, J. (2001). The emergence of the unmarked pronoun. In G. Legendre, J. Grimshaw, & S. Vikner (Eds.), *Optimality-theoretic syntax* (pp. 113–142). MIT Press.

- Bril, I. (2004). Coordination strategies and inclusory constructions in New Caledonian and other Oceanic languages. *Typological Studies in Language*, 58, 499–536.
- Bybee, J. (2006). From usage to grammar: The mind's response to repetition. *Language*, 711–733.
- Calhoun, S. (2015). The interaction of prosody and syntax in Samoan focus marking. *Lingua*, 165, 205–229.
<https://doi.org/10.1016/j.lingua.2014.11.007>
- Camden, W. (1979). Parallels in structure and lexicon and syntax between New Hebrides Bislama and the South Santo language spoken at Tangoa. In P. Muhlhausler, E. Lattey, E. Woolford, W. Camden, M. Steffenson, M. Clyne, W. Peet Jr., U. Mosel, I. Smith, J. Gebhard, D. Nguyen, G. R. Dreyfuss, D. Oka, & {de Reigo (Eds.), *Papers in Pidgin and Creole Linguistics No. 2* (pp. 51–118). Pacific Linguistics.
<https://openresearch-repository.anu.edu.au/handle/1885/144547>
- Cameron, R. (1992). *Pronominal and null subject variation in Spanish: Constraints, dialects, and functional compensation* [Ph.D. thesis]. University of Pennsylvania.
- Cameron, R. (1995). The scope and limits of switch reference as a constraint on pronominal subject expression. *Hispanic Linguistics*, 6, 1–28.
- Cardinaletti, A. (1997). Agreement and control in expletive constructions. *Linguistic Inquiry*, 521–533.
- Carnie, A. (2021). *Syntax: A generative introduction* (4th ed.). Wiley-Blackwell.
<http://gen.lib.rus.ec/book/index.php?md5=203A1C3F74905EB25755336147E9EBFF>
- Carvalho, A. M., & Child, M. (2011). Subject pronoun expression in a variety of Spanish in contact with Portuguese. In J. Michnowicz & R. Dodsworth

- (Eds.), *Selected proceedings of the 5th Workshop on Spanish Sociolinguistics* (pp. 14–25). Cascadilla Proceedings Project.
- Chafe, W. L. (1994). *Discourse, consciousness, and time: The flow and displacement of conscious experience in speaking and writing*. University of Chicago Press.
- Chen, V., & McDonnell, B. (2019). Western Austronesian voice. *Annual Review of Linguistics*, 5(1), 173–195. <https://doi.org/10.1146/annurev-linguistics-011718-011731>
- Clark, H. H., & Begun, J. S. (1971). The semantics of sentence subjects. *Language and Speech*, 14(1), 34–46.
- Clark, R. (1973). Transitivity and case in Eastern Oceanic languages. *Oceanic Linguistics*, 12(1/2), 559–605.
- Clark, R. (1985). The Efate dialects. *Te Reo*, 28, 3–35.
- Comrie, B., Haspelmath, M., & Bickel, B. (2015). *Leipzig glossing rules*. Conventions for Interlinear Morpheme-by-Morpheme Glosses. Leipzig: Max Planck Institute for Evolutionary Anthropology. <https://www.eva.mpg.de/lingua/resources/glossing-rules.php>
- Corbett, G. G. (2003). Agreement: The range of the phenomenon and the principles of the Surrey Database of Agreement. *Transactions of the Philological Society*, 101(2), 155–202.
- Corston-Oliver, S. H. (2003). Core arguments and the inversion of the nominal hierarchy in Roviana. In J. W. D. Du Bois, L. E. Kumpf, & W. J. Ashby (Eds.), *Preferred argument structure: Grammar as architecture for function* (pp. 273–300). John Benjamins.
- Coupé, C. (2018). Modeling linguistic variables with regression models: Addressing non-Gaussian distributions, non-independent observations,

- and non-linear predictors with random effects and generalized additive models for location, scale, and shape. *Frontiers in Psychology*, 9, 513.
- Crowley, T. (1998). *An eromangan (Sye) grammar*. University of Hawaii Press.
- Davey, K. J. (2021). *A quantitative study of directional constructions in Matukar Panau* [Honours thesis]. The Australian National University.
- Davidson, B. (1996). 'Pragmatic weight' and Spanish subject pronouns: The pragmatic and discourse uses of 'tú' and 'yo' in spoken Madrid Spanish. *Journal of Pragmatics*, 26(4), 543–565.
- Debeer, D., Hothorn, T., & Strobl, C. (2021). *permimp: Conditional Permutation Importance* (1.0-2) [Computer software]. <https://cran.r-project.org/web/packages/permimp/>
- Donohue, M., & Sawaki, Y. (2007). Papuan Malay pronominals: Forms and functions. *Oceanic Linguistics*, 46(1), 253–276.
- Du Bois, J. W. (1987). The discourse basis of ergativity. *Language*, 805–855.
- Duhamel, M.-F. (2019). The possessive classifiers in Raga, Vanuatu: An investigation of their use and function in natural speech. *Te Reo*, 62(1), 26–48.
- Dunbar, R. W. (1983). The Topicality Hierarchy in German Syntax. *Monatshefte*, 141–146.
- Duranti, A. (1979). Object clitic pronouns in Bantu and the topicality hierarchy. *Studies in African Linguistics*, 10(1), 31–45.
- Early, R. (1994). *A grammar of Lewo, Vanuatu* [Ph.D.]. The Australian National University.
- ELAN (Version 6.6). (2023). [Computer software]. Nijmegen: Max Planck Institute for Psycholinguistics, The Language Archive. <https://archive.mpi.nl/tla/elan>

- Elff, M. (2022). *mclogit: Multinomial Logit Models, with or without Random Effects or Overdispersion* (0.9.6) [Computer software]. <https://cran.r-project.org/web/packages/mclogit/index.html>
- Evans, B. (2008). Subject agreement in Marovo: Diachronic explanations of synchronic conditions. *Oceanic Linguistics*, 383–408.
- Ewing, M. C. (2014). Motivations for first and second person subject expression and ellipsis in Javanese conversation. *Journal of Pragmatics*, 63, 48–62.
- Exter, M. (2012). ‘Realis’ and ‘irrealis’ in Wogeo: A valid category? In N. Evans & M. Klamer (Eds.), *Melanesian languages on the edge of Asia: Challenges for the 21st century* (pp. 174–190). University of Hawai’i Press.
- Fernandez, A., Garcia, S., Herrera, F., & Chawla, N. V. (2018). SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary. *Journal of Artificial Intelligence Research*, 61, 863–905. <https://doi.org/10.1613/jair.1.11192>
- Flego, S., & Forrest, J. (2021). Leveraging the temporal dynamics of anticipatory vowel-to-vowel coarticulation in linguistic prediction: A statistical modeling approach. *Journal of Phonetics*, 88(1), 1–19. <https://doi.org/10.1016/j.wocn.2021.101093>
- Flores-Ferrán, N. (2004). Spanish subject personal pronoun use in New York City Puerto Ricans: Can we rest the case of English contact? *Language Variation and Change*, 16(1), 49–73.
- Fox, J., Weisberg, S., Price, B., Friendly, M., & Hong, J. (2022). *effects: Effect Displays for Linear, Generalized Linear, and Other Models* (4.2-2) [Computer software]. <https://cran.r-project.org/web/packages/effects/index.html>

- François, A., Franjeh, M., Lacrampe, S., & Schnell, S. (2015). The exceptional linguistic density of Vanuatu. In A. François, M. Franjeh, S. Lacrampe, & S. Schnell (Eds.), *The languages of Vanuatu: Unity and diversity* (pp. 1–21). Asia-Pacific Linguistics.
- Frascarelli, M., & Hinterhölzl, R. (2007). Types of topics in German and Italian. In K. Schwabe & S. Winkler (Eds.), *On Information Structure, Meaning and Form: Generalizations across languages* (pp. 87–116). John Benjamins Publishing Company. <https://doi.org/10.1075/la.100.07fra>
- Genetti, C., & Crain, L. D. (2003). Beyond preferred argument structure: Sentences, pronouns and given referents in Nepali. In W. J. Ashby, L. E. Kumpf, & J. W. Du Bois (Eds.), *Preferred argument structure: Grammar as architecture for function* (pp. 197–223). John Benjamins.
- Givón, T. (1983a). *Topic continuity in discourse: A quantitative cross-language study*. John Benjamins. <https://benjamins.com/catalog/tsl.3>
- Givón, T. (1983b). Topic continuity in discourse: An introduction. In T. Givón (Ed.), *Topic continuity in discourse: A quantitative cross-language study* (pp. 3–41). John Benjamins.
- Givón, T. (1983c). Topic continuity in spoken English. In T. Givón (Ed.), *Topic continuity in discourse: A quantitative cross-language study* (pp. 343–363). John Benjamins.
- Goffman, E. (1974). *Frame analysis: An essay on the organization of experience*. Harper.
- Goffman, E. (1981). *Forms of talk*. University of Pennsylvania Press.
- Greenfield, P. M., & Smith, J. H. (1976). *The structure of communication in early language development*. Academic Press.
- Gudmestad, A., & Geeslin, K. L. (2010). Exploring the roles of redundancy and ambiguity in variable subject expression: A comparison of native and

- non-native speakers. In C. Borgonovo, M. Español-Echevarría, & P. Prévost (Eds.), *Selected proceedings of the 12th Hispanic linguistics symposium* (pp. 270–283). Cascadilla Proceedings Project.
- Gundel, J. K. (1996). Relevance theory meets the givenness hierarchy. An account of inferrables. *Pragmatics and beyond New Series*, 141–154.
- Gundel, J. K. (2010). Reference and Accessibility from a Givenness Hierarchy Perspective. *International Review of Pragmatics*, 2(2), 148–168.
<https://doi.org/10.1163/187731010X528322>
- Gundel, J. K., Hedberg, N., & Zacharski, R. (1993). Cognitive status and the form of referring expressions in discourse. *Language*, 274–307.
- Haig, G., & Schnell, S. (2014). *Annotations using GRAID (Grammatical Relations and Animacy in Discourse). Manual Version 7.0.*
https://www.academia.edu/10328418/Haig_Geoff_and_Stefan_Schnell_2014_Annotations_using_GRAID_Grammatical_Relations_and_Animacy_in_Discourse_Manual_Version_7_0
- Hartig, F. (2022). *DHARMa: Residual Diagnostics for Hierarchical (Multi-Level / Mixed) Regression Models* (0.4.6) [Computer software]. <https://cran.r-project.org/web/packages/DHARMa/index.html>
- Hartmann, R. R. K., & Stork, F. C. (1972). *Dictionary of Language and Linguistics*. Applied Science Publishers.
- Haspelmath, M. (2011). The gradual coalescence into ‘words’ in grammaticalization. In B. Heine & H. Narrog (Eds.), *The Oxford Handbook of Grammaticalization*. Oxford University Press.
<https://doi.org/10.1093/oxfordhb/9780199586783.013.0027>
- Haspelmath, M. (2013). Argument indexing: A conceptual framework for the syntactic status of bound person forms. In *Argument indexing: A conceptual framework for the syntactic status of bound person forms*

- (pp. 197–226). De Gruyter Mouton.
<https://doi.org/10.1515/9783110331127.197>
- Hastie, T. J., & Pregibon, D. (1992). Generalized linear models. In *Statistical models in S* (pp. 195–247). Routledge.
- Hawkinson, A. K., & Hyman, L. M. (1974). Hierarchies of natural topic in Shona. *Studies in African Linguistics*, 5(2), 147–170.
- Hayashi, R. (1987). *A study of floor management of English and Japanese conversation*. University of Illinois at Urbana-Champaign.
- Hayashi, R. (1991). Floor structure of English and Japanese conversation. *Journal of Pragmatics*, 16(1), 1–30.
- Heltai, P. (2005). Explication, redundancy, ellipsis and translation. *New Trends in Translation Studies. In Honour of Kinga Klaudy. Budapest: BPH*, 46–49.
- Hockett, C. F. (1958). *A course in modern linguistics*.
- Horning, A. S. (1979). On defining redundancy in language: Case notes. *Journal of Reading*, 22(4), 312–320.
- Hothorn, T., Hornik, K., Strobl, C., & Zeileis, A. (2023). *party: A Laboratory for Recursive Partytioning* (1.3-13) [Computer software]. <https://cran.r-project.org/web/packages/party/index.html>
- Huang, C. (2011). Referential choice in Mandarin child language: A discourse-pragmatic perspective. *Journal of Pragmatics*, 43(7), 2057–2080.
- Huang, L. M. (1995). The syntactic structure of Wulai and Mayrinax Atayal: A comparison. *Bulletin of National Taiwan Normal University*, 40, 261–294.
- Huang, X., & Hsieh, F. (2006). The pragmatics of case marking in Saisiyat. *Oceanic Linguistics*, 45(1), 91–109.

- Hyslop, C. (2001). *The Lolovoli dialect of the North-east Ambae language, Vanuatu*. Pacific Linguistics, Research School of Pacific and Asian Studies, The
- Iemmolo, G. (2010). Topicality and differential object marking Evidence from Romance and beyond. *Studies in Language*, 34, 239–272.
<https://doi.org/10.1075/sl.34.2.01iem>
- Itagaki, N., & Prideaux, G. D. (1985). Nominal properties as determinants of subject selection. *Lingua*, 66(2–3), 135–149.
- Ivanova-Sullivan, T. (2014). Deictic pronouns in oral production. In *Theoretical and experimental aspects of syntax-discourse interface in heritage grammars* (pp. 111–124). Brill.
https://doi.org/10.1163/9789004246171_008
- Jacobson, S. M. (2016). *Cognizer Verbs and the First-Person Pronoun Yo in Contact: A Corpus Study*.
- Kibrik, A. A., Khudyakova, M. V., Dobrov, G. B., Linnik, A., & Zalmanov, D. A. (2016). Referential choice: Predictability and its limits. *Frontiers in Psychology*, 7, 1429.
- Kidima, L. (1987). Object agreement and topicality hierarchies in Kiyaka. *Studies in African Linguistics*, 18(2), 175–210.
- Klamer, M. A. F. (1998). *Kambera* (Vol. 18). Walter de Gruyter.
- Krahmer, E., & Van Deemter, K. (2012). Computational generation of referring expressions: A survey. *Computational Linguistics*, 38(1), 173–218.
- Krajinović, A. (2017, December 7). *Some issues in TAM from an Oceanic perspective* [Conference presentation]. ALS2017, Sydney, Australia.
- Krejci, B. (2014). *What is Raining? English Weather" it" Revisited*. 5, 16–1.
- Krifka, M. (2008). Basic Notions of Information Structure. *Acta Linguistica Hungarica*, 55(3–4), 243–276.

- Krifka, M., & Musan, R. (2012). Information structure: Overview and linguistic issues. In M. Krifka & R. Musan (Eds.), *Information structure: Overview and linguistic issues* (pp. 1–44). De Gruyter Mouton.
<https://doi.org/10.1515/9783110261608.1>
- Kuhn, M. (2023). *caret: Classification and Regression Training* (6.0-94) [Computer software]. <https://cran.r-project.org/web/packages/caret/>
- Lacrampe, S. (2009). *Possession in Lelepa, a language of Central Vanuatu* [M.A. Thesis]. University of the South Pacific.
- Lacrampe, S. (2012). Simplifying a system: A story of language change in Lelepa, Vanuatu. In M. Ponsonnet, L. Dao, & M. Bowler (Eds.), *Proceedings of the 42nd Australian Linguistic Society Conference—2011* (pp. 223–244). Australian Linguistic Society.
<https://openresearch-repository.anu.edu.au/handle/1885/9402>
- Lacrampe, S. (2014). *Lelepa: Topics in the grammar of a Vanuatu language* [Ph.D. thesis, The Australian National University].
<https://openresearch-repository.anu.edu.au/handle/1885/12893>
- Lacrampe, S. (2017). *Description and documentation of Lelepa, an endangered language of Central Vanuatu*. Endangered Languages Archive. <http://hdl.handle.net/2196/00-0000-0000-000F-BAF7-9>
- LaFond, L. L. (2001). *The Pro-drop Parameter in Second Language Acquisition Revisited: A Developmental Account*.
<https://doi.org/10.7282/T34B304Q>
- Lastra, Y., & Butragueño, P. M. (2015). Subject pronoun expression in oral Mexican Spanish. In A. M. Carvalho, R. Orozco, & N. L. Shin (Eds.), *Subject pronoun expression in Spanish: A cross-dialectal perspective* (pp. 39–57). Georgetown University Press.

- Lee, D.-Y., & Yonezawa, Y. (2008). The role of the overt expression of first and second person subject in Japanese. *Journal of Pragmatics*, 40(4), 733–767. <https://doi.org/10.1016/j.pragma.2007.06.004>
- Lee, N. (2021). Overt Subjects Signaling Floor Shifts in Korean Discourse. *Language Research (어학연구)*, 57(3), 329–354.
- Leufkens, S. (2020). A functionalist typology of redundancy. *Revista Da ABRALIN*, 19(3), 79–103.
- Levin, B., & Krejci, B. (2017). *Talking about the weather: A case study of precipitation verbs*. Linguistics Colloquium. University at Buffalo.
- Levshina, N. (2016). When variables align: A Bayesian multinomial mixed-effects model of English permissive constructions. *Cognitive Linguistics*, 27(2), 235–268.
- Levshina, N. (2022). Comparing Bayesian and Frequentist Models of Language Variation: The Case of Help + (to-)Infinitive. In J. Schlüter & O. Schützler (Eds.), *Data and Methods in Corpus Linguistics: Comparative Approaches* (pp. 224–258). Cambridge University Press; Cambridge Core. <https://doi.org/10.1017/9781108589314.009>
- Li, P. J. (2000). Some Aspects of Paze Syntax. *Oceanic Linguistics Special Publications*, 29, 89–108.
- Lieberman, P. (1963). Some Effects of Semantic and Grammatical Context on the Production and Perception of Speech. *Language and Speech*, 6(3), 172–187. <https://doi.org/10.1177/002383096300600306>
- Longacre, R. E. (1999). Grammatical units. *Concise Encyclopedia of Grammatical Categories*, 172–177.
- Longacre, R. E. (2007). Sentences as combinations of clauses. In T. Shopen (Ed.), *Language Typology and Syntactic Description: Volume 2:*

Complex Constructions (2nd ed., Vol. 2, pp. 372–420). Cambridge University Press; Cambridge Core.

<https://doi.org/10.1017/CBO9780511619434.007>

Lynch, J. (2016). Chapter 6. Oceanic Languages: Grammatical Overview. In *Chapter 6. Oceanic Languages: Grammatical Overview* (pp. 100–165). University of Hawaii Press. <https://doi.org/10.1515/9780824842581-010>

Lynch, J., & Crowley, T. (2001). *Languages of Vanuatu: A New Survey and Bibliography*. Pacific linguistics, Research School of Pacific and Asian Studies, Australian National University.

Lynch, J., Crowley, T., & Ross, M. (2002). *The Oceanic languages*. Curzon.

Macdonald, D. E. (2017). *Oceanic languages: A comparative investigation of pre-clausal constructions: Vol. Doctor of Philosophy (PhD)* [Doctoral, University of Waikato]. <https://hdl.handle.net/10289/11363>

MacKenzie, L. (2020). Comparing constraints on contraction using Bayesian regression modeling. *Frontiers in Artificial Intelligence*, 3, 1–18.

Massam, D., Peter, V., & Wang, L. (2012). Radical pro-drop in Niuean. In Paula Caxaj (Ed.), *Proceedings of the 2012 Annual Conference of the Canadian Linguistic Association* (pp. 1–10). Canadian Linguistic Association.

McCawley, J. D. (1999). Participant Roles, Frames, and Speech Acts. *Linguistics and Philosophy*, 22(6), 595–619.

Meakins, F. (2015). Not obligatory: Bound pronoun variation in Gurindji and Bilinarra. *Asia-Pacific Language Variation*, 1(2), 128–161. <https://doi.org/10.1075/aplv.1.2.02mea>

- Meyerhoff, M. (1998). Accommodating your data: The use and misuse of accommodation theory in sociolinguistics. *Language & Communication*, 18(3), 205–225.
- Miller, G. (1945). *Havannah Harbour, West Efate*.
- Mithun, M. (1990). Third-person reference and the function of pronouns in Central Pomo natural speech. *International Journal of American Linguistics*, 56(3), 361–376.
- Mithun, M. (2003). Pronouns and agreement: The information status of pronominal affixes. *Transactions of the Philological Society*, 101(2), 235–278. <https://doi.org/10.1111/1467-968X.00119>
- Nalborczyk, L., Batailler, C., Løevenbruck, H., Vilain, A., & Bürkner, P.-C. (2019). An introduction to Bayesian multilevel models using brms: A case study of gender effects on vowel variability in standard Indonesian. *Journal of Speech, Language, and Hearing Research*, 62(5), 1225–1242.
- Neeleman, A., & Szendrői, K. (2005). Pro drop and pronouns. In J. Alderete, C. Han, & A. Kochetov (Eds.), *Proceedings of the 24th West Coast Conference on Formal Linguistics* (Vol. 24, pp. 299–307). Cascadia Proceedings Project.
- Neeleman, A., & Szendrői, K. (2007). Radical pro drop and the morphology of pronouns. *Linguistic Inquiry*, 38(4), 671–714. <https://doi.org/10.1162/ling.2007.38.4.671>
- Nicenboim, B., & Vasishth, S. (2016). Statistical methods for linguistic research: Foundational Ideas—Part II. *Language and Linguistics Compass*, 10(11), 591–613.
- Nichols, J. (2018). Agreement with overt and null arguments in Ingush. *Linguistics*, 56(4), 845–863.

- Ordóñez, F., & Treviño, E. (1999). Left dislocated subjects and the pro-drop parameter: A case study of Spanish. *Lingua*, 107(1), 39–68.
[https://doi.org/10.1016/S0024-3841\(98\)00020-5](https://doi.org/10.1016/S0024-3841(98)00020-5)
- Orozco, R. (2015). Pronominal variation in Colombian costeño Spanish. In A. M. Carvalho, R. Orozco, & Naomi Lapidus Shin (Eds.), *Subject pronoun expression in Spanish: A cross-dialectal perspective* (pp. 17–37). Georgetown University Press Washington, DC.
- Palmer, B. (2009). Clause order and information structure in Cheke Holo. *Oceanic Linguistics*, 48(1), 213–249.
- Pan, C.-J. (2012). *A grammar of Lha'alua, an Austronesian language of Taiwan* [Ph.D. thesis, James Cook University].
<https://researchonline.jcu.edu.au/32386/>
- Paulo, J. R., Pires, G., & Nunes, U. J. (2021). Cross-subject zero calibration driver's drowsiness detection: Exploring spatiotemporal image encoding of EEG signals for convolutional neural network classification. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 29, 905–915.
- Pearce, E. (2012). Dual and plural in languages of Vanuatu. *Linguistic Typology*, 16(2), 191–232. <https://doi.org/10.1515/lity-2012-0007>
- Pearson, J., Stevenson, R., & Poesio, M. (2001). The effects of animacy, thematic role, and surface position on the focusing of entities in discourse. In M. Poesio (Ed.), *Proceedings of the First Workshop on Cognitively Plausible Models of Semantic Processing*.
- Pescarini, D. (2022). Expletive Subject Clitics in Northern Italo-Romance. *Languages*, 7(4), 265.
- Pesetsky, D. (1987). Wh-in-situ: Movement and unselective binding. *The Representation of (in) Definiteness*, 98, 98–129.

- Poedjosoedarmo, S. (1968). Javanese Speech Levels. *Indonesia*, 6, 54–81.
<https://doi.org/10.2307/3350711>
- Polinsky, M., & Potsdam, E. (to appear). Austronesian syntax. In B. Palmer (Ed.), *The languages and linguistics of Island Southeast Asia and the Pacific: A comprehensive guide*. Mouton de Gruyter.
- Posio, P. (2011). Spanish subject pronoun usage and verb semantics revisited: First and second person singular subject pronouns and focusing of attention in spoken Peninsular Spanish. *Journal of Pragmatics*, 43(3), 777–798.
- Posio, P. (2013). The expression of first-person-singular subjects in spoken Peninsular Spanish and European Portuguese: Semantic roles and formulaic sequences. *Folia Linguistica*, 47(1), 253–292.
- R Core Team. (2023). *R: A language and environment for statistical computing* (4.3.1) [Windows]. R Foundation for Statistical Computing.
<https://www.r-project.org/>
- Ross, M. (1988). *Proto Oceanic and the Austronesian languages of western Melanesia*. Dept. of Linguistics, Research School of Pacific Studies, the Australian National University.
- Rubino, C. R. G. (1997). *A Reference Grammar of Ilocano* (p. xxii 522) [Phd]. UMI.
- Ruda, M. (2015). Rich agreement and dropping patterns: Pro-drop, agreement drop, no drop. In Y. Oseki, M. Esipova, & S. Harves (Eds.), *Proceedings of the 24th Meeting of Formal Approaches to Slavic Linguistics* (pp. 321–341). Michigan Slavic Publications.
- Schachter, P., & Otnes, F. T. (1983). *Tagalog reference grammar*. Univ of California Press.

- Schneider, C. (2009). Information structure in Abma. *Oceanic Linguistics*, 1–35.
- Schnell, S. (2018a). Attention focus and information packaging in Vera'a demonstratives. In S. Riesberg, A. Shiohara, & A. Utsumi (Eds.), *Perspectives on information structure in Austronesian languages* (pp. 81–113). Language Science Press.
<https://doi.org/10.5281/zenodo.1402571>
- Schnell, S. (2018b). Whence subject-verb agreement? Investigating the role of topicality, accessibility, and frequency in Vera'a texts. *Linguistics*, 56(4), 735–780.
- Schnell, S., & Barth, D. (2020). Expression of anaphoric subjects in Vera'a: Functional and structural factors in the choice between pronoun and zero. *Language Variation and Change*, 32(3), 267–291.
- Schütz, A. J. (1969). *Nguna grammar*. University of Hawaii Press.
- Schwenter, S. A. (2006). Null objects across South America. In T. L. Face & C. A. Klee (Eds.), *Selected proceedings of the 8th Hispanic Linguistics Symposium* (Issue 1986, pp. 23–36). Cascadia Proceedings Project.
- Schwenter, S. A. (2014). Two kinds of differential object marking in Portuguese and Spanish. In P. Amaral & A. M. Carvalho (Eds.), *Portuguese-Spanish interfaces: Diachrony, synchrony, and contact* (pp. 238–260). John Benjamins.
- Shin, N. L. (2014). Grammatical complexification in Spanish in New York: 3sg pronoun expression and verbal ambiguity. *Language Variation and Change*, 26(3), 303–330.
- Shin, N. L., & Cairns, H. S. (2009). Subject pronouns in child Spanish and continuity of reference. In J. Collentine, M. García, B. Lafford, & F. M.

- Marín (Eds.), *Selected proceedings of the 11th Hispanic linguistics symposium* (pp. 155–164). Cascadilla Proceedings Project.
- Siewierska, A. (1999). *From anaphoric pronoun to grammatical agreement marker: Why objects don't make it*. 33(1–2), 225–252.
<https://doi.org/10.1515/flin.1999.33.1-2.225>
- Simard, C., & Wegener, C. (2017). Fronted NPs in a verb-initial language – clause-internal or external? Prosodic cues to the rescue! *Glossa: A Journal of General Linguistics*, 2(1), Article 1.
<https://doi.org/10.5334/gjgl.211>
- Sneddon, J. N., Adelaar, K. A., Djenar, D. N., & Ewing, M. (2012). *Indonesian: A comprehensive grammar*. Routledge.
- Sneddon, & Neil, J. (2010). *Indonesian Reference Grammar*. Allen & Unwin.
<http://gen.lib.rus.ec/book/index.php?md5=20b0a997891a6f8305f7c1bc3fc28bd7>
- Spportiche, D., Koopman, H., & Stabler, E. (2013). *An introduction to syntactic analysis and theory*. Wiley-Blackwell.
- Stalnaker, R. (2002). Common ground. *Linguistics and Philosophy*, 25(5/6), 701–721.
- Sturt, P. (2022). Syntactic and semantic mismatches in English number agreement. *Glossa Psycholinguistics*, 1(1).
<https://doi.org/10.5070/G6011159>
- Sun, C. (2023). *An Investigation of Syllable Structure in Lelepa* [Master's thesis, The Australian National University]. <https://openresearch-repository.anu.edu.au/handle/1885/299626>
- Suwarno, B., & Yanwar, A. (2020). *Speech Level in Some Indonesian Regional Languages and Implication for Conservation*.
<https://doi.org/10.2991/assehr.k.200130.066>

- Teng, S. (2005). Grammatical relations in Puyuma. In null I Wayan Arka and Malcolm Ross (Ed.), *The many faces of Austronesian voice systems: Some new empirical studies* (pp. 137–151). Pacific Linguistics.
- Teng, S. (2008). *A Reference Grammar of Puyuma: An Austronesian Language of Taiwan*. <https://www.semanticscholar.org/paper/A-Reference-Grammar-of-Puyuma%3A-An-Austronesian-of-Teng/aafc1e707777e05620f01e93ea85992e04fa5da4>
- Thieberger, N. (2006). *A grammar of South Efate: An Oceanic language of Vanuatu*. University of Hawaii Press.
- Torres Cacoullos, R., & Travis, C. E. (2018). Subject Pronoun Expression: Reconsidering the Constraints. In *Bilingualism in the Community: Code-switching and Grammars in Contact* (pp. 74–110). Cambridge University Press. <https://doi.org/10.1017/9781108235259.005>
- Torres Cacoullos, R., & Travis, C. E. (2019). Variationist typology: Shared probabilistic constraints across (non-) null subject languages. *Linguistics*, 57(3), 653–692.
- Travis, C. E., & Torres Cacoullos, R. (2012). What do subject pronouns do in discourse? Cognitive, mechanical and constructional factors in variation. *Cognitive Linguistics*, 23(4), 711–748.
- Trillo, J. R. (1997). Your attention, please: Pragmatic mechanisms to obtain the addressee's attention in English and Spanish conversations. *Journal of Pragmatics*, 28(2), 205–221.
- Tryon, D. T. (1976). *New Hebrides languages: An internal classification*. Dept. of Linguistics, Research School of Pacific Studies, The Australian
- Vasishth, S., Chen, Z., Li, Q., & Guo, G. (2013). Processing Chinese relative clauses: Evidence for the subject-relative advantage. *PloS One*, 8(10), e77006.

- Vollmer, M. (2023). *Comparing zero and referential choice in eight languages with a focus on Mandarin Chinese*. <https://doi.org/10.1075/sl.21072.vol>
- Walczak, B., & Massart, D. L. (2000). Calibration in wavelet domain. In B. Walczak (Ed.), *Data Handling in Science and Technology* (Vol. 22, pp. 323–349). Elsevier. [https://doi.org/10.1016/S0922-3487\(00\)80040-4](https://doi.org/10.1016/S0922-3487(00)80040-4)
- Wegener, C. (2012). *A Grammar of Savosavo*. De Gruyter Mouton. <https://doi.org/10.1515/9783110289657>
- Wickham, H. (2023). *tidyverse: Easily Install and Load the “Tidyverse”* (2.0.0) [Computer software]. <https://cran.r-project.org/web/packages/tidyverse/index.html>
- Wickham, H., Chang, W., Henry, L., Pedersen, T. L., Takahashi, K., Wilke, C., Woo, K., Yutani, H., & Dunnington, D. (2023). *ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics* (3.4.3) [Computer software]. <https://cran.r-project.org/web/packages/ggplot2/index.html>
- Wit, E., & Gillette, M. (1999). *What is linguistic redundancy*. University of Chicago. <http://www.math.rug.nl/~ernst/linguistics/redundancy3.pdf>
- Yamada, F. S. (2006). The pronoun system in Galeya: Arguments against a clitic analysis. *Oceanic Linguistics*, 45(2), 447–490.
- Yeh, M. M. (2003). A syntactic and semantic study of Saisiyat verbs. *PhD Diss., National Taiwan Normal University*.
- Zeitoun, E. (1997). The Pronominal System of Mantauran (Rukai). *Oceanic Linguistics*, 36(2), 312–346. <https://doi.org/10.2307/3622988>
- Zeitoun, E. (2000). A reference grammar of Tsou. *Taipei: Yuan-Liou Publishing Co.[In Chinese.]*.
- Zeitoun, E. (2005). Tsou. *The Austronesian Languages of Asia and Madagascar*, 259–290.
- Zeitoun, E. (2007). *A grammar of Mantauran (Rukai)*.

Zheng, W. (2020). The first person singular pronoun topic as attention-getter in interaction: A study of qá mà in Longxi Qiang. *Journal of Pragmatics*, 156, 1–15. <https://doi.org/10.1016/j.pragma.2019.12.002>