



Te Reo
the Journal of the Linguistic
Society of New Zealand

Volume 69

Issue 1

Research Article

2026

Pages 110–141

October, 2026

The linguistic diversity of western Espiritu Santo Island (Vanuatu)

Tihomir Rangelov

Max Planck Institute for Evolutionary Anthropology

This paper is a peer-reviewed contribution from <https://www.nzlingsoc.org/journal/current-issue/>

©Te Reo – The Journal of the Linguistic Society of New Zealand

Editor: Victoria Chen

The linguistic diversity of western Espiritu Santo Island (Vanuatu)

TIHOMIR RANGELOV

Max Planck Institute for Evolutionary Anthropology

Abstract

Espiritu Santo, the largest island of Vanuatu, is known as a global hotspot of linguistic diversity, but this diversity remains severely underdocumented. This pilot study aims to determine the scale of linguistic diversity on Santo in greater detail based on newly collected lexical data from 17 varieties, ensuring dense geographical coverage of a selected area in the western half of the island. Adopting the methodology of the main previous study of the area, Tryon (1976), shared cognacy rates among the varieties are calculated. These results are compared to diachronic and typological patterns in the phonologies of the varieties, and to the results of a lexical variation analysis. Based on these findings, a few additional languages can be postulated in the studied area, alongside six previously identified languages, as well as greater dialectal variation. The language ideologies of speakers of these varieties are also discussed employing the Indigenous term *lanwis* in order to test how the etic results of the linguistic data analysis correspond to speakers' emic perceptions of *lanwis* boundaries. This study demonstrates that the extreme linguistic diversity of western Santo is greater and geographically more complex than previously assumed, and highlights the urgency of further documentation and analysis of Santo's linguistic varieties.

Keywords

Vanuatu, Espiritu Santo, linguistic diversity, *lanwis*, Oceanic languages

1 Introduction

Vanuatu has consistently ranked as the country with the most languages per capita in the world (Crowley, 2000, p. 50; Adelaar, 2009, p. 36; François et al., 2015, p. 8; Rangelov et al., 2025a, p. 1). The two largest islands of Vanuatu – Espiritu Santo and Malekula – are also among the regions with the highest geographical density of languages in the world (Rangelov, 2026a; cf. François et al., 2015, p. 8). An important reason we know these facts is that half a century ago, in 1976, Darrell Tryon published his seminal work *New Hebrides languages: An internal classification*,¹ and for the first time made it clear that the linguistic diversity of this archipelago was so substantial. Tryon's data included cognate-coded basic vocabulary from hundreds of villages throughout Vanuatu. Tryon performed a lexicostatistical analysis, investigated the

¹ New Hebrides is the colonial name of the group of islands that is now the Republic of Vanuatu, a name that was adopted at the time Independence was declared in 1980. In this article, the toponym Vanuatu is also used to refer to the archipelago during the time period before 1980.

phonological histories of the recorded varieties and arrived at a count of 105 distinct Indigenous² languages.³

Since Tryon (1976), considerable work has been carried out on the languages of Vanuatu. However, the amount of documentation, description and analysis varies, with Espiritu Santo – the largest island, also commonly called Santo – being by far the least well covered area (Rangelov et al., 2025b). Tryon (1976, p. 80) listed 29 languages for Santo, which form several overlapping dialect chains with intersecting isoglosses, later identified as *linkages* by Ross Clark (2009a, 2009b) following terminology introduced by Ross (1988). Part of the reason for the lack of attention to these languages is that the linguistically densest western half of Santo is relatively inaccessible, being dominated by a high mountain chain with large rivers and very poor infrastructure. In this paper, the focus is precisely on this area.

Tryon's (1976) count of 105 languages for Vanuatu has been revised up as documentation work proceeded. Adelaar (2009) reported 114 languages for Vanuatu. François et al. (2015) list 138 as a maximal number of languages for Vanuatu. Glottolog 5.3 (Hammarström et al., 2026) lists 118 languages based on a reconciliation of multiple works. This variation is due to whether extinct languages are counted, or whether, especially in a few cases, a dialect chain is counted as a single language or as separate languages. The addition of new languages is most often the result of new work that has been done in the past 50 years. For example, for Malekula, the second largest island of Vanuatu, Tryon (1976) listed 28 distinct languages, comparable to Santo's 29; however, Glottolog now lists 34 languages for Malekula. The increase of six languages is based on more recent linguistic data and analysis. Santo's language count remains at 29 in Glottolog today. However, multiple lines of evidence suggest that the linguistic diversity on Santo may be substantially higher:

- (1) Tryon (1976, p. 61) himself suggests lower confidence for the geographical coverage of Santo in his data collection at the time: "It is still possible that one or two additional languages may be discovered on the island of Espiritu Santo";
- (2) in a later work, Tryon (2010) postulates 35 languages for Santo (i.e. six more) based on reports by Santo-based *filwoka*⁴ and their perceptions about language boundaries and mutual intelligibility (see Section 2.3 on the *lanwis* concept); due to lack of data, the match between these perceptions and linguistic evidence has not been tested yet;
- (3) based on the addition of six new languages for Malekula, it could be expected that new language work could uncover even more new languages on Santo, which is larger and has even more rugged terrain, and where a much larger population lives inland and in other areas with poor accessibility, compared to Malekula, where most of the population lives on or near the coast. These conditions are assumed predictors of higher

² By "Indigenous (languages)" I refer to the Oceanic languages of Vanuatu, whose history predates contact between Indigenous populations and Western European colonisers. In other words, this set excludes Vanuatu's three official languages English, French and Bislama; Bislama might be considered an Indigenous language of Vanuatu by some measures, even if it emerged following, and as a consequence of, that contact.

³ Tryon uses an 80% shared cognacy threshold to distinguish a language from a dialect. This is discussed further in Sections 5 and 8.

⁴ The Vanuatu Cultural Centre, a government agency, runs the very successful *Filwoka* Programme, where community experts in Indigenous culture and languages, titled *filwoka* (a Bislama term, from English *fieldworker*), from throughout Vanuatu gather to exchange information and ideas, get training and promote their *kastom* and languages (Regenvanu, 2011; Tilon, 2011). Darrell Tryon served as facilitator of such workshops at the Vanuatu Cultural Centre (Rousseau, 2011) and the data in Tryon (2010) are based on interviews with such *filwoka*.

diversity and lower levels of endangerment, as well as higher chance of lower coverage in Tryon (1976);

- (4) Tryon (1976, pp. 87, 90–91) reports some dialectal variation for 12 of his 29 languages. Firstly, considering that Tryon (1976) collected data from 40 settlements across Santo, whereas the total number of settlements on the island is in the few hundreds, it can be expected that more substantial dialectal variation exists. Secondly, it is possible that some of the dialects established by Tryon may turn out to be different enough to be considered distinct languages.⁵

In this article, based on new data, I investigate in more detail the scale of linguistic diversity in western Santo. The aim is to determine diversity in terms of both the number of postulated distinct languages and variation within them.⁶ Following Tryon (1976), I start by analysing shared cognacy rates between the varieties in the sample, based on cognate-coded vocabulary data. This is done for a selected geographic area of western Santo.⁷ The added detail of this study is made possible because the new data are more comprehensive than Tryon's in two respects: (1) the geographic coverage is denser in that the new data were collected from 17 different locations in the selected area, compared to ten⁸ locations for the same area in Tryon (1976, p. 83), and (2) the data include basic vocabulary items, as well as some non-basic concepts. In this study, I also look at shared phonological history (based on Rangelov, 2026b) and lexical variation to validate the results of the shared cognacy analysis and establish more detail where possible. The non-basic vocabulary has allowed for a more confident analysis of the phonological history and lexical variation, also contributing to more confident cognate coding.

The 17 varieties included in the current sample are spoken on the western coast of Santo (12 geographically contiguous varieties) and on the eastern side of the Santo mountain chain (five varieties). All of them surely belong to the Espiritu Santo clade.⁹ The covered area is a subset of the most severely underdocumented area of Santo – its western half. In this sense,

⁵ Rangelov et al. (2025b: 37) list the following examples of Tryon (1976)/Glottolog “languages”, for which “there is a tradition to study different dialects separately, and sometimes arguments for considering the varieties to be separate languages”: Uripiv-Atchin-Wala-Rano (a.k.a Northeast Malekula, Glottocode: urip1239) (cf. Moore, 2019; Wessels, 2021); Malua-Tepërav (Glottocode: malu1245) with its dialects Espiegle's Bay and Malua Bay, of northern Malekula (cf. Wessels, 2013; Holmes, 2014); and the Sakao-Nkep language of northeastern Santo (known collectively as Wanohe in Glottolog, Glottocode: saka1289) (cf. Touati, 2014; Meyerhoff, 2017; Hog Harbour Community et al., 2021). In some cases, divergence in recent decades might be one reason why these varieties may be considered different languages now.

⁶ A single speaker was recorded per communalect, which does not allow for determining variation within communalects. However, documenting substantial dialectal variation (between communities) is possible with the current method (see Section 4 and the supplementary materials), and reports about perceived boundaries between varieties (see Sections 2.3, 3.1) do not suggest that any substantial variation has been overlooked within the geographical scope of the study.

⁷ The selected area (see Section 3.1) is not intended to encompass a specific subgroup of languages. Therefore, no attempt has been made to reconstruct any proto-forms or offer a classification hypothesis here.

⁸ This should be taken as an approximate number as the locations cannot be pinpointed exactly based on Tryon's maps. I have included all doculects that Tryon lists under languages that overlap with the current sample, see also Section 2.2.

⁹ As per Glottolog's (Hammarström et al., 2026) classification, Santo's languages form a discrete exclusive group: Austronesian > Malayo-Polynesian > Oceanic > North and Central Vanuatu > Espiritu Santo. Glottolog's classification is a reconciliation of Lynch (2020), Lynch & Crowley (2001), Tryon (2010) and Clark (2009a); see these works for alternative subgroupings. Ross et al. (2023, pp. 11, 586–587) also posit Espiritu Santo as a discrete exclusive clade with the following lineage: Oceanic > Southern Oceanic linkage > North Vanuatu linkage > Espiritu Santo linkage.

this is also a pilot study aiming to determine the urgency of continuing the linguistic surveying of Santo.

In Section 2, I elaborate on the background of this study and present an overview of previous work on Santo languages. The varieties included in the study and the data are presented in Section 3, followed by a description of the methods in Section 4. The following sections present the relevant findings of the shared cognacy analysis (Section 5), historical-comparative analysis, based on Rangelov (2026b) (Section 6) and the lexical variation analysis (Section 7). A discussion of the results is presented in Section 8 and some notes on future research conclude the paper in Section 9.

2 Background

2.1 Geography

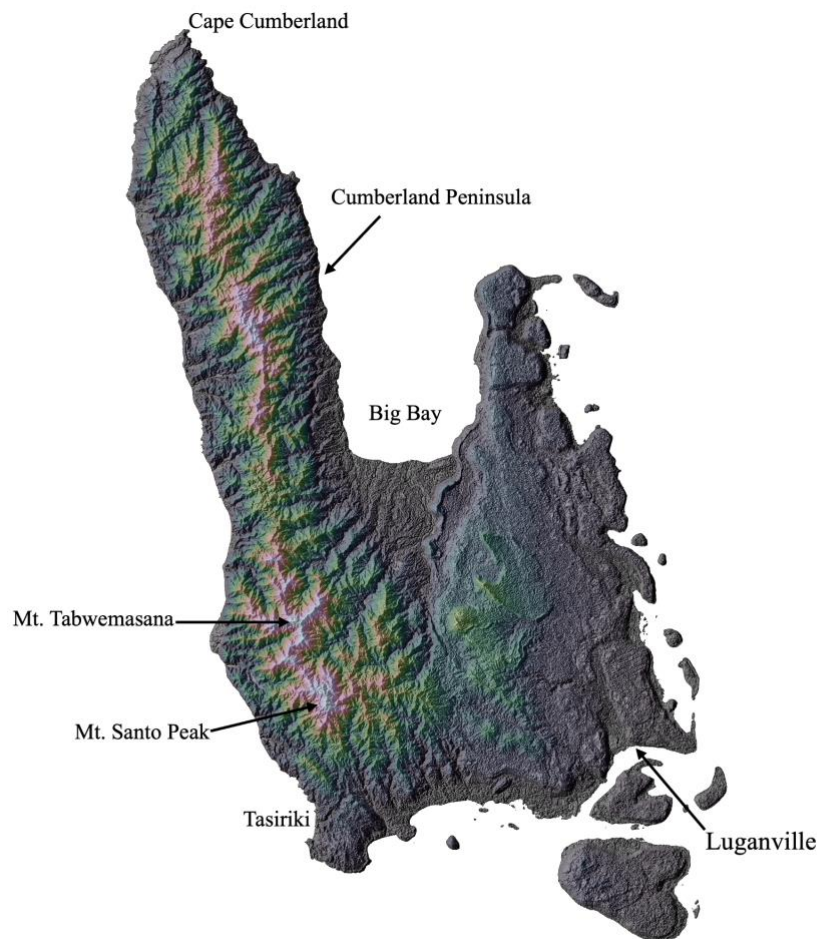
Santo is the largest island in Vanuatu with a land area of 3,955.5 sq. km. Together with its outlying islands it forms Vanuatu's Sanma province with a total population of 43,165, of which 17,719 people reside in Luganville, Vanuatu's second largest urban and commercial centre and cultural melting pot, and 25,446 reside in rural areas of Sanma province, according to the 2020 Census (VBOS, 2020). This makes rural Sanma province the area of Vanuatu with the lowest population density (*ibid.*).

Geographically, Santo is roughly divided in two areas, see Map 1. In the east, the island is dominated by lowlands and a plateau, as well as many offshore islands. Located in the southeast corner, the town of Luganville features an international airport and other infrastructure. There is a paved road extending along most of the east coast. The east coast and the offshore islands are where much of Sanma's tourism infrastructure is located. The entire eastern area, including the plateau has seen substantial numbers of both international and ni-Vanuatu immigrants for many decades, due to military activity by the US Navy during WWII, the establishment of many plantations and other agricultural businesses decades ago and their persistence, and, more recently a tourism and holiday home industry. There are a few poor-quality gravel roads leading into the plateau and extending into Big Bay. The historical developments and relative accessibility of eastern Santo have resulted in the restructuring of Indigenous social networks, the mixing of populations and the concentration of the population in Luganville, its surroundings and a few villages and plantations on the plateau and along the east coast. This has also reduced the vitality of many Indigenous languages in this area.

Along the south coast of Santo, a gravel road extends from Luganville to Tasiriki village in the southwestern corner of the island. The eastern half of this road is of relatively good quality and gives access to some peri-urban areas and larger agricultural businesses. The road quickly deteriorates when one goes further into the more remote areas of the western half of Santo.

The western half of Santo is dominated by the Santo Mountain Chain, starting from Cape Cumberland in the north, forming the entire Cumberland Peninsula and extending further to the southern coast. The mountain chain widens in the south, with the highest peak, Mt. Tabwemasana, at 1879 metres above sea level. Virtually, the entire ridge of the mountain chain is at above 1000 metres altitude. The mountain chain forms a sharp geographic and logistical boundary, as well as a climate boundary. Due to the predominantly southeasterly winds, the steeper west coast of Santo is in the rain shadow and is drier with shorter streams and rivers. The terrain and climate of the west coast mean that most human settlements are relatively far from each other; the potential for sustaining higher populations may be limited. The eastern slopes of the mountain chain see more precipitation which feeds longer and larger rivers

running east and north into Big Bay, the largest river being the Jordan river. The southern slopes are also dominated by the valleys of a few large rivers. Apart from the mentioned poor-quality roads into a few settlements on the Big Bay shore and along the south coast, the rest of the western half of Santo is only accessible by expensive, weather-dependent boat trips along the coast or on foot. This also means that there has been no industrial development and little immigration into the area, and traditional social networks are much better preserved than in the east. Overall, the vitality of most extant Indigenous languages in the west is greater than in the east, albeit all Indigenous languages of Santo perform poorly by some vitality measures, and future prospects are bleak (cf. Taman et al., 2025).



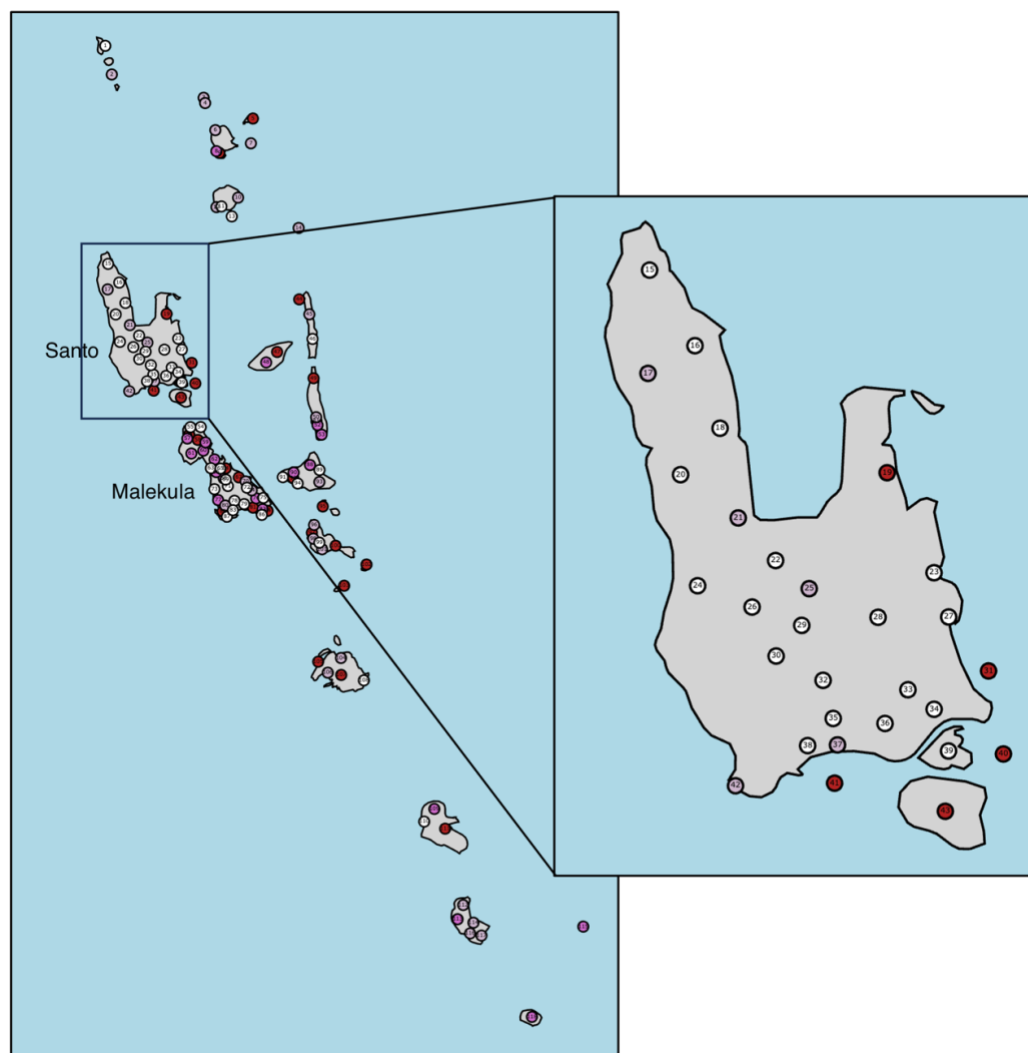
Map 1. A topographic map of Espiritu Santo Island and offshore islands (Sanma province). Map base: European Space Agency (2024); © DLR e.V. 2010–2014 and © Airbus Defence and Space GmbH 2014–2018 provided under COPERNICUS by the European Union and ESA.

2.2 *Santo's languages are un(der)documented*

Geographical and infrastructural factors are likely reasons why Santo is by far the least linguistically studied island (or island group)¹⁰ of Vanuatu, according to Rangelov et al. (2025b), who looked at how much work has been done on the Indigenous languages of Vanuatu, see Map 2. Most of the languages of Santo lack documentation, apart from the

¹⁰ Unless otherwise specified, when discussing the languages of Santo as a whole, all outlying large and small islands that appear on Map 1 are included.

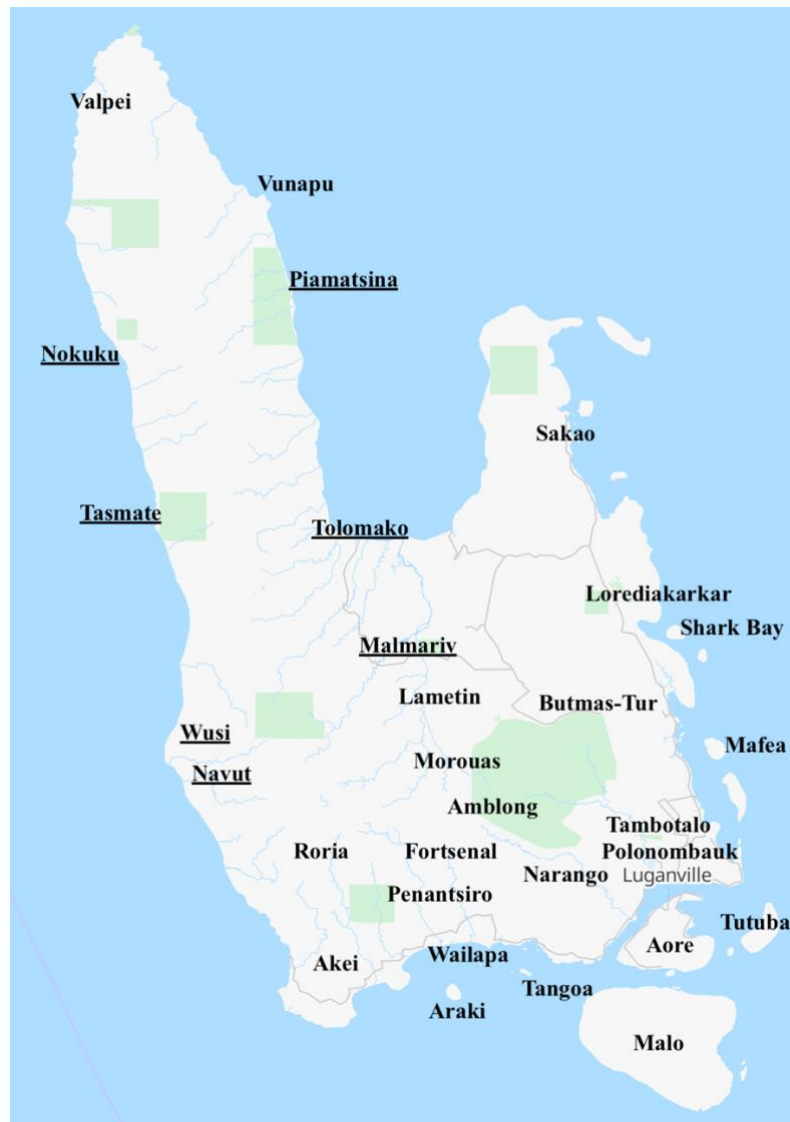
wordlists in Tryon (1976).¹¹ In terms of the amount of language work, Santo stands in stark contrast with Malekula, Vanuatu's second largest island (Map 2, cf. Section 1); while there are still notable gaps in linguistic research on Malekula's languages, its coverage is significantly higher and better distributed than Santo's due to the pioneering efforts of Jean-Michel Charpentier (1982) and Terry Crowley (1998, 2006a, 2006b, 2006c, 2006d), followed by the work of many other linguists and community members.



Map 2. The level of grammatical description of the Indigenous languages of Vanuatu and Santo (the inset). Darker colours mean more comprehensive description; white means no description. Santo clearly stands out as the least well covered area. This map first appeared in the supplementary material of Rangelov et al. (2025b) and is used with permission. That article also reports on other language work beyond grammatical descriptions – lexical description, linguistic research, corpora, and Bible translations. The results for those other measures show similar patterns, i.e. Santo is the least well covered area of Vanuatu in terms of those types of works as well.

¹¹ Most of the language work done on Santo is on a few languages along the more accessible east and south coasts (see Rangelov et al., 2025b for details). For the rest of the island, there are short grammar sketches and other short descriptive works for a handful of languages (Codrington, 1885; Gordon, 1889; Ray, 1926; Ludvigson, 1989; Chung, 2005; Clark, 2006; Meyerhoff, 2009). One western Santo language is included in Clark's (2009a) historical analysis of North-Central Vanuatu languages. Clark (2009b) proposes a classification of Santo's languages based on Tryon's (1976) data. There are translations of parts of the Bible for three languages.

Tryon (1976, p. 83) collected 41 wordlists from 40 locations on Santo. His analysis identified 29 distinct languages (Map 3), 24 of which comprise his West Santo clade, a primary subgroup of the Espiritu Santo family, which consists of overlapping dialect continua or linkages.



Map 3. The 29 distinct languages identified by Tryon (1976, p. 8) for Santo. The languages that appear in the approximate area relevant to the present study were underlined by me. Map base: © OpenStreetMaps contributors.

The newly collected data, on which the present study is based, are from varieties spoken in a relatively compact area of Santo, covering roughly a quarter of the island's land area (see Section 3.1 for details on the varieties and their geographic distribution). In approximately the same area, on Map 3, the following seven languages identified by Tryon (1976) appear: Nokuku, Tasmate, Wusi, Tie (alt. name Navut), Tiale (alt. name Malmariv), Tolomako and Piamatsina.¹² The seven languages correspond to eleven of Tryon's wordlists (doculects) (see Tryon, 1976, pp. 83, 90, note that there are two wordlists for location Wusi). Figure 1 presents a matrix of Tryon's shared cognacy rates among these eleven doculects.

¹² In Tryon (2010), no new languages were added to this area based on reports by the *filwoka*, although some languages may appear under a different name in this later work.

	Nokuku	Tasmate	Kerepua	Wusi I	Wusi II	Piamatsina	Tolomako	Malmariv	Nonona	Navut
Tasmate	71.5% 151									
Kerepua	50.3% 169	64.9% 151								
Wusi I	50.3% 167	64.9% 151	88.4% 181							
Wusi II	52.2% 157	63.0% 146	87.7% 162	93.3% 163						
Piamatsina	67.9% 168	74.8% 151	53.5% 185	54.4% 182	55.6% 162					
Tolomako	57.6% 165	67.8% 149	59.4% 180	57.7% 215	61.0% 159	67.6% 182				
Malmariv	52.3% 153	61.0% 136	63.3% 166	63.8% 213	68.2% 151	58.1% 167	63.1% 198			
Nonona	50.9% 165	66.0% 147	80.1% 181	80.1% 231	85.6% 160	54.9% 182	59.9% 217	67.9% 218		
Navut	49.7% 155	62.0% 137	62.0% 166	65.9% 214	68.0% 153	56.3% 167	57.5% 200	77.6% 214	71.8% 220	
Matae	52.4% 164	64.8% 145	67.4% 178	68.4% 225	71.1% 159	56.4% 179	58.7% 213	75.6% 217	73.5% 234	81.8% 220

Figure 1: Cognacy rates for the relevant doculects from Tryon (1976, pp. 116–117). In each cell, the number on top is the shared cognacy percentage, below is the number of concepts, based on which the calculation was made (i.e. the number of concepts for which there was data for both compared varieties). Cells where shared cognacy is above 80% are in black. This was Tryon’s threshold for distinguishing language from dialect. Rates above 70% are in grey. This highlighting is mine (cf. Figure 2 in Section 5).

2.3 Language, dialect and lanwis

The main purpose of this study is to establish a level of linguistic diversity, both between languages and within languages. The issue of distinguishing a language from a dialect is a well-known conundrum in linguistics, especially in the case of dialect chains or linkages. Tryon (1976) used an 80% threshold of shared cognacy to distinguish language from dialect, which is based on heuristic observations regarding mutual intelligibility in the Vanuatu context (i.e. in what cases speakers of different variants can understand each other and/or whether they resort to the *lingua franca* Bislama instead). The criterion of mutual intelligibility is especially problematic in areas of considerable small-scale multilingualism, such as Vanuatu and Santo in particular. People acquire competence in neighbouring varieties for various reasons, including cooperation and trade between communities, exogamy, and other forms of contact. Competence in multiple, especially neighbouring, language varieties is highly regarded (Lindstrom, 1994, pp. 75–76). Some varieties may be more likely to be acquired by non-L1 speakers after gaining in prestige, as is the case with the Nokuku, Tasmate, and Tolomako varieties in the sample, which became more prestigious because they were chosen to be used in the newly established religious and formal educational institutions (cf. Section 3.1). In the case of Vanuatu, and Santo, the fact that some varieties gain prestige for such reasons contributes to a transition to a hierarchical model from an older, egalitarian model of small-

scale multilingualism (cf. François, 2012; Lüpke, 2016; Pakendorf et al., 2021), which further complicates the application of the mutual intelligibility criterion.¹³

As is well-known, a boundary between languages may also be drawn on the basis of language attitudes and ideologies. In Vanuatu, the Bislama term *lanwis* (from English *language*)¹⁴ is used broadly to distinguish Indigenous Oceanic languages from the set of colonial languages Bislama, English and French. However, *lanwis* (and its translational counterparts in the Indigenous languages) can also be used to mark identity and belonging to a place (Lindstrom, 1994; Ridge, 2025). For the varieties included in this study, local attitudes and ideologies regarding boundaries between different *lanwis* are reported in Section 3.1 (cf. also Taman et al., 2025 for the varieties along the western coast). For example, the speakers of three pairs of varieties in the sample (those spoken in Kerepua+Wusi, Elia+Panas, and Linduri+Winsao) report that the two varieties in their respective pair are part of the same respective *lanwis* that are different from other *lanwis* in the sample.¹⁵ In Section 8, I discuss how these attitudes and ideologies map to the findings of the linguistic analysis.

3 Data

This study is based on new data from 17 Santo varieties collected by the author in 2022, 2023, and 2025. The collected data include lexical data and sociolinguistic interviews. For most varieties, a single speaker was recorded, usually a male community leader (chief or church leader). For the varieties Penouru-Hurahi and Sauriki-Kitano, data were elicited from a group of rememberers, since these two varieties are dormant. Details about the varieties can be found in Section 3.1. Details about the methods used for collecting the lexical data, its annotation and organisation are discussed in Section 3.2 and in the supplementary material.

3.1 *The varieties in this sample*

The 17 varieties appear on Map 4. The labels have the following format [variety index]_[geographic location (village/hamlet)]_[variety endonym]. In the rest of the article the varieties may be referred to only by their indexes for brevity.

¹³ Another problem with the mutual intelligibility criterion is that substantial phonological and grammatical divergence may reduce mutual intelligibility between varieties that otherwise have high shared cognacy. In Section 6, we see that there are important phonological differences between some of the varieties in this study, but they are perhaps not very substantial and, in some cases, are likely compensated for by small-scale multilingualism. Regarding grammatical differences between the varieties, this remains a question for future documentation and analysis.

¹⁴ In this paper, the form *lanwis* is used for both singular and plural referents.

¹⁵ Places not included in this sample where some of the present *lanwis* are spoken, are also mentioned, but further discussion of this is beyond the scope of this article.



Map 4. A map of Santo and larger neighbouring islands with the varieties included in this study. In grey are relevant settlements and other locations mentioned in this article. Variety 17 is placed at its historical location (see below for details). The indexes start from the northwesternmost variety and follow the west coast south to then continue in a rough south-to-north direction on the eastern side of the mountains. This map first appeared in Rangelov (2026b) and is used with permission.

Table 1 presents details about the 17 varieties that are relevant to the present study, including reported perceptions of *lanwis* boundaries and mutual intelligibility with/competence

in other languages. Mutual intelligibility/competence may be due to both genealogical and areal effects. These data were provided during variety and speaker background interviews with the recorded speakers, and, in the case of a few details, during informal conversation with other speakers. The details in the table include data collected by the author as well as previously published data as indicated. Other details about the varieties can be found in Taman et al. (2025) and Rangelov (2026b).

Table 1. Sociolinguistic and other details about the 17 varieties that are relevant to this article

Variety	Speakers' attitudes and ideologies regarding <i>lanwis</i> boundaries	Reported competence of and mutual intelligibility with other varieties	Notes
01_Nokuku_Tanokuk	The same <i>lanwis</i> is the L1 in a few settlements to the north, as well as Penouru, Nampeko and Ravlepa to the south	Speakers report some competence in Tavanlav (a <i>lanwis</i> spoken further north), and in varieties 02,03,04, and, to a lesser extent, 05.	It corresponds to the Nokuku or Nogugu language (Glottocode: noku1237, ISO 639-3 code: nkk). Nokuku wordlist in Tryon (1976); language discussed in Clark (2009a, 2009b). <i>Lingua franca</i> of the northwestern coast of Santo. Speakers of 02 have shifted to 01.
02_Penouru_Lela	Speakers consider it its own <i>lanwis</i> .	Speakers report some competence in Tavanlav, and in 03,04,05. Mutual intelligibility with an unidentified variety on the eastern side of the mountains (Big Bay coast) due to historical connections before migration from the hills.	Dormant, speakers switched to 01 after migrating to the coast from the hills.
03_Petawata_Hurahi	Speakers consider it its own <i>lanwis</i> .	All speakers are also fluent in 01. Some competence in Tavanlav, as well as 04,05 and less so in 07,08.	
04_Wunavae_Takani	Speakers consider it its own <i>lanwis</i> .	Speakers report some competence in varieties 01,03,05,07,08 on the west coast. They also report that there are communities on the Big Bay coast that speak varieties very similar to theirs since they originated from	

Variety	Speakers' attitudes and ideologies regarding <i>lanwis</i> boundaries	Reported competence of and mutual intelligibility with other varieties	Notes
		the same hill communities, but migrated to different coasts.	
05_Tasmate_Oa	The same <i>lanwis</i> is also spoken in Vasalea and Sulesai to the north of Tasmate.	Speakers report some competence in 01,04,06,07,08 and less so in 10,11. They also report that their language is historically closely related to a "Tolomako <i>lanwis</i> " spoken in Big Bay.	Corresponds to Tasmate/Meri/'Oa language (Glottocode: tasm1246, ISO 639-3 code: tmt). Wordlist from Tasmate in Tryon (1976), also discussed in Clark (2009b). Some older residents of neighbouring west coast settlements from Wunavae to around Wusi may have had exposure to this variety in their childhood, as the only primary school in the area until the 1980s was in Tasmate, where students boarded.
06_Wunaruu_Joa	Speakers report that the same <i>lanwis</i> is spoken in the Big Bay village Malau (<i>lanwis</i> endonym Mojoa) and an inland Big Bay village of unknown name (<i>lanwis</i> endonym Nojoa).	Speakers report some competence in the west coast varieties 07,08,09 to the south and 05 and even 01 to the north	Alternative endonym: Japwele Moved to the current location from Big Bay in 1998.
07_Elia_Jo	Speakers consider it a separate <i>lanwis</i> from 08 (which is spoken in the southern half of Elia, Kerepua, Wusi) and other west coast varieties. They consider it the same <i>lanwis</i> as variety 13.	Speakers report fluency in 05,08 and some competence in 04,06,10,11.	
08_Kerepua_Tuijo	Speakers consider it the same <i>lanwis</i> as 09 (with slight differences), as well as the variety of the southern half of Elia. Reportedly, the same <i>lanwis</i> is also spoken by close relatives in	Speakers report mutual intelligibility with or competence of 05,06,07 and 10,11,14.	Tryon (1976) has a wordlist from Kerepua, and two Wusi lists which are listed under the language Kula/Wusi (Glottocode: wusi1237, ISO 639-3 code: wsi). This language is discussed in Clark (2009b).

Variety	Speakers' attitudes and ideologies regarding <i>lanwis</i> boundaries	Reported competence of and mutual intelligibility with other varieties	Notes
	the villages Narata and Maljia on the eastern slopes of the mountains.		
09_Wusi_Kula	Speakers consider it the same <i>lanwis</i> as 08 (with slight differences).	Speakers report mutual intelligibility with or competence of 05,06,07 and 10,11,14.	Tryon (1976) has two wordlists from Wusi, which together with his Kerepua variety, is listed under the language Kula/Wusi (Glottocode: wusi1237, ISO 639-3 code: wsi). This language is discussed in Clark (2009b).
10_Linduri_Tie	Speakers report that the same <i>lanwis</i> Tie is spoken in the villages Winsao (variety 14), as well as Lawuswus and Wuniu on the eastern slopes of the mountains.	Speakers report mutual intelligibility with or some competence of 07,08,09 and 11,12, but consider their variety to be a separate <i>lanwis</i> from these.	A language under the name Tie (also known as Navut, ?Bura, ?Sinia) is known in previous literature with the Glottocode navu1237 and ISO 396-3 code nsw. Tryon's (1976) list from Navut is likely closely related to 10.
11_Sauriki_Tiale	Speakers report that the same <i>lanwis</i> , Tiale, is spoken by all villages in the Sauriki river valley (except the highest village Narumaji, where reportedly a <i>lanwis</i> called Akei is spoken), the Pareo river valley to the south, and some villages in the Kerewai river valley further south, as well as some villages on the eastern slopes of the mountains.	Speakers report that their <i>lanwis</i> has relatively high mutual intelligibility with variety 10, but it is not the same <i>lanwis</i> , and they have some competence of 05,07,08,09.	A language known as Tiale (Glottocode tial1239 and ISO 639-3 code mnl) is included in Clark (2009b). Glottolog lists Malmariv as an alternative linguonym of this language. Tryon (1976) has a wordlist labelled Malmariv, which comes from a location at low altitude on the eastern side of the mountains, around the river Jordan (cf. Map 3). Untangling this language's dialectal distribution and its relationship to neighbouring languages is pending future data collection.
12_Sauriki_Kitano	Speakers consider it a separate <i>lanwis</i> from 11.		Dormant, speakers switched to 11.
13_Panas_Jo	Speakers consider it the same <i>lanwis</i> as 07, different from all other <i>lanwis</i> .	Speakers report some competence in 05,10,11,14.	Speakers relocated from Elia in 2015 due to the frequent droughts on the west coast.
14_Winsao_Tie	Speakers report that the same <i>lanwis</i> is spoken by the people of Linduri on the west	Speakers report varying degrees of competence in 11,16,17 and other	See note on variety 10.

Variety	Speakers' attitudes and ideologies regarding <i>lanwis</i> boundaries	Reported competence of and mutual intelligibility with other varieties	Notes
	coast (variety 10), as well as other communities on the eastern slopes of the mountains, as well as communities on the southern slopes of the mountains.	varieties in their vicinity.	
15_Petel_Na	Speakers report that the same <i>lanwis</i> is spoken in a few villages and hamlets close to Petel, including Pelwus, Pialulup, and Vas.	Speakers report some competence in 16, and the <i>lanwis</i> Pesena to the north, as well as 17 to the south, and some comprehension of 04 on the west coast due to historical ties.	
16_Piamatsina_Pwelmaw	Speakers report that the same <i>lanwis</i> is spoken in the nearby villages Pialalau and Piavu, as well as that it is the same <i>lanwis</i> as variety 03, due to historical connections before 20th century migrations.	Speakers report some competence in the Pesena <i>lanwis</i> to the north and to Tolomako varieties to the south. The latter is due to Tolomako serving as a <i>lingua franca</i> in large parts of coastal Big Bay.	Tryon (1976) has a wordlist from Piamatsina village, based on which Piamatsina has been listed as a separate language (Glottocode: piam1242, ISO 639-3 code: ptr). This language is discussed in Clark (2009b).
17_Portolry_Tolomako	Speakers report that they use the same <i>lanwis</i> as some communities along the Big Bay coast (cf. Clark, 2006; Meyerhoff, 2009).	Speakers report some competence in Sakao (the other language spoken in Port Olry).	The ancestors of this variety migrated around 100 years ago from Tolomako village in the southwestern corner of Big Bay to their current location (Monnier, 1994, p. II), see Map 1. Tryon (1976) has a wordlist from Tolomako village. A language known as Tolomako, corresponding to these varieties (Glottocode: tolo1255; ISO 639-3: tlm) is very likely also the variety described by Codrington (1885) and Ray (1926).

3.2 The lexical data

For the lexicostatistical analysis, a subset of the lexical data was used. The sample was based on two principles: (1) including basic vocabulary, which is a core principle in lexicostatistics, also followed by Tryon (1976), and (2) aiming for comparability with Tryon's (1976) sample in order to be able to compare the results of this study with his. Based on these, all concepts on the basic vocabulary list and those items from the other lists that also appear on Tryon's (1976)

list were included. The included concepts are indicated by a ‘yes’ in the ‘included_in_lexicostatistics’ column of the data spreadsheet. This resulted in a list of 250 concepts. While this list is shorter than Tryon’s maximum list of 293 concepts, gaps in Tryon’s data mean that in reality his lexicostatistical analysis is rarely based on a list of more than 200 concepts, at least for the varieties relevant to this study (cf. Figure 1 in Section 2.2).

A few statistics about the dataset are as follows (more details can be found in the R-notebook in the supplementary material):

- 8615 datapoints in total for over 700 concepts;
- 250 concepts included in lexicostatistical analysis, 4200 data points across them;
- The variety with the fewest concepts represented in the lexicostatistical analysis is 17_Portolry_Tolomako with 244 data points;
- For 225 concepts there is data for all varieties; of the remaining 25 concepts data is lacking for one or two varieties in most cases.

Regarding the subset of lexical data used for the phonological analysis, see Section 6, Rangelov (2026b) and the supplementary material. Regarding the data used for the lexical variation analysis, see Section 7.

4 Methods

To investigate the scale of linguistic diversity in the selected area based on the data set described above, I adopt both an etic and an emic perspective through the language-dialect-*lanwis* framework outlined in Section 2.3. The lexical, phonological and phonetic data are analysed to arrive at a formal categorisation of the varieties in the sample driven by language structure and an attempt at achieving maximum scientific objectivity. I take into account the emic perspective by mapping these results to the *lanwis* boundaries reported by the speakers.

The shared cognacy analysis, the historical and typological phonological analysis, and the lexical variation analysis are all based on the cognate-coded lexical data. Some details about the methods of each analysis appear in the respective sections. Following Tryon (1976, pp. 75–76), in the lexicostatistical analysis only homosemantic cognates were allowed. In the handful of doubtful cases, I erred on the conservative side and gave different cognate codes. The methods related to data annotation, organisation, and cognate coding are described in detail in the supplementary material.

5 Shared cognacy

Figure 2 displays the shared cognacy rates for each possible pair of varieties, which were calculated using R code (R Core Team 2026), including the tidyverse package (Wickham et al., 2019) in R-Studio (Posit Team 2026), see the R notebook in the supplementary material.

	01_Nokuku_Tanokuk	02_Penouru_Lela	03_Petawata_Hurahi	04_Wunavae_Takani	05_Tasmate_Oa	06_Wunaru_Joa	07_Elia_Jo	08_Kerepua_Tuijo	09_Wusi_Kula	10_Linduri_Tie	11_Sauriki_Tiale	12_Sauriki_Kitano	13_Panas_Jo	14_Winsao_Tie	15_Petel_Na	16_Piamatsina_Pwelmau	17_Portolry_Tolomako
02_Penouru_Lela	77.2% 246																
03_Petawata_Hurahi	70.6% 248	78.5% 246															
04_Wunavae_Takani	57.5% 247	64.1% 245	65.2% 247														
05_Tasmate_Oa	65.7% 248	69.5% 246	69.4% 248	73.7% 247													
06_Wunaru_Joa	46.2% 249	52.2% 247	51.4% 249	64.5% 248	61.4% 249												
07_Elia_Jo	45.2% 248	47.2% 246	46.4% 248	58.3% 247	55.6% 248	71.1% 249											
08_Kerepua_Tuijo	43.7% 245	46.9% 243	45.9% 246	56.1% 244	55.1% 245	70.7% 246	83.3% 245										
09_Wusi_Kula	42.2% 244	46.7% 242	45.1% 244	56.0% 243	52.7% 245	68.6% 245	79.5% 244	85.2% 243									
10_Linduri_Tie	40.5% 247	43.7% 245	43.1% 248	50.4% 246	50.6% 247	56.9% 248	55.9% 247	61.2% 245	61.7% 243								
11_Sauriki_Tiale	41.6% 243	44.8% 241	45.1% 244	50.8% 244	50.6% 243	55.3% 244	56.4% 243	61.4% 241	60.3% 239	81.1% 243							
12_Sauriki_Kitano	40.8% 245	44.4% 243	44.7% 246	50.4% 244	49.8% 245	53.7% 246	53.5% 245	57.4% 244	59.1% 242	76.3% 245	80.5% 241						
13_Panas_Jo	44.7% 244	47.1% 242	45.5% 244	55.9% 245	54.5% 244	69.4% 245	86.1% 244	81.4% 242	78.0% 241	59.3% 243	57.3% 241	55.4% 242					
14_Winsao_Tie	40.8% 245	42.6% 244	43.1% 246	49.6% 246	48.2% 245	55.7% 246	55.9% 245	59.7% 243	61.0% 241	82.9% 245	80.7% 243	75.3% 243	58.0% 243				
15_Petel_Na	52.2% 245	56.4% 243	56.7% 245	71.1% 246	62.6% 246	59.3% 246	52.7% 245	49.6% 242	48.8% 242	44.7% 244	45.7% 243	45.9% 242	48.6% 243	45.1% 244			
16_Piamatsina_Pwelmau	62.1% 248	73.6% 246	74.2% 248	64.0% 247	64.7% 249	51.0% 249	48.0% 248	44.9% 245	44.9% 245	43.3% 247	43.2% 243	42.4% 245	46.3% 244	43.7% 245	64.2% 246		
17_Portolry_Tolomako	49.4% 243	53.1% 241	54.5% 244	66.5% 242	58.4% 243	63.5% 244	55.6% 243	53.7% 242	51.2% 240	49.0% 243	48.8% 240	46.9% 241	52.3% 239	47.3% 241	74.3% 241	59.3% 243	

Figure 2. Shared cognacy rates for the 17 varieties. Following Tryon's (1976) presentation (cf. Figure 1), in each cell, the number on top is the percentage of shared cognacy, below is the number of concepts on which the calculation was made (i.e. the number of concepts for which there is data for both varieties). Cells with shared cognacy above 80% are in black, those over 75% are in dark grey and above 70% are in light grey.

First, I compare these results with Tryon's (cf. Figure 1 in Section 2.2 above), where possible. In Table 2 below, Tryon's shared cognacy results are matched with the new results for pairs of varieties, which are very likely the same in the two samples based on the locations where the data were collected (cf. Tryon, 1976, p. 83 for locations). Tryon's shared cognacy rates are consistently higher, the average difference being 8.1 percentage points. There are at least two possible factors that can explain this difference:

- (1) The new results are based on more consistent data. In the current analysis, shared cognacy was calculated based on at least 239 concepts, compared to 150-200 concepts for most pairs in Tryon's calculations (and as low as 136). Tryon (1976, p. 163) tests the effect of wordlist size (for 100, 200, and 300 items) between ten pairs of selected varieties from throughout Vanuatu and finds that a larger wordlist size consistently yields lower shared cognacy rates with an average difference of 7 pct. points between the 100-item and 300-item lists and an average difference of 1.5 pct. points between the 200-item and 300-item lists. This suggests that, compared to Tryon's shared cognacy rates, the new results might be lower by a few pct points due to wordlist size effects.
- (2) It is possible that, at least in some cases, varieties have diverged somewhat in the past 50 years. The smallest difference in Table 2 is between the Kerepua and Wusi varieties, whose speakers report awareness of speaking the same *lanwis* and also live in relatively close proximity, which promotes contact. It is possible that contact and language attitudes and ideologies may have constrained divergence between these two varieties through the decades, while they also jointly diverged from other varieties further away, where we see much higher difference scores. Besides, some closely related varieties that are no longer geographically close after migrations (cf. Section 3.1), may have

diverged from each other since. These hypotheses, however, remains to be tested based on more detailed data and modelling.

Table 2. Comparison of the new shared cognacy results to Tryon's (1976) for comparable pairs of varieties

Pair	Tryon's (1976) shared cognacy rate	Shared cognacy rate (this study)	Difference (pct. pts.)
Nokuku~Tasmate (01~05)	71.5%	65.7%	5.8
Nokuku~Kerepua (01~08)	50.3%	43.7%	6.6
Nokuku~Wusi I (01~09)	50.3%	42.2%	8.1
Nokuku~Wusi II (01~09)	52.2%	42.2%	10
Nokuku~Piamatsina (01~16)	67.9%	62.1%	5.8
Tasmate~Kerepua (05~08)	64.9%	55.1%	9.8
Tasmate~Wusi I (05~09)	64.9%	52.7%	12.2
Tasmate~Wusi II (05~09)	63%	52.7%	10.3
Tasmate~Piamatsina (05~16)	74.8%	64.7%	10.1
Kerepua~Wusi I (08~09)	88.4%	85.2%	3.2
Kerepua~Wusi II (08~09)	87.7%	85.2%	2.5
Kerepua~Piamatsina (08~16)	53.5%	44.9%	8.6
Wusi I~Piamatsina (09~16)	54.4%	44.9%	9.5
Wusi II~Piamatsina (09~16)	55.6%	44.9%	10.7
Average difference:			8.1

Figure 3 offers a summary of the relationships between the varieties based on the new shared cognacy results. In order to identify how the varieties cluster in a way that is both meaningful and clear, I have decided on a round 75% cut-off point, around which to organise the analysis. This is five percentage points lower than Tryon's 80% threshold, which is also in line with the discussion above.¹⁶ In Section 8, this threshold is discussed further.

¹⁶ To demonstrate sensitivity to the cut-off point, in comparison with Figure 3, at a 73% threshold, stronger links between varieties 04+05 and 15+17 would have to be posited, and variety 16 would be more strongly related to 02 and 03; at 77 %, Cluster 3 would have to be split in two; at 79%, varieties 01, 02, and 03 would stand on their own, and Cluster 2 would be less coherent.

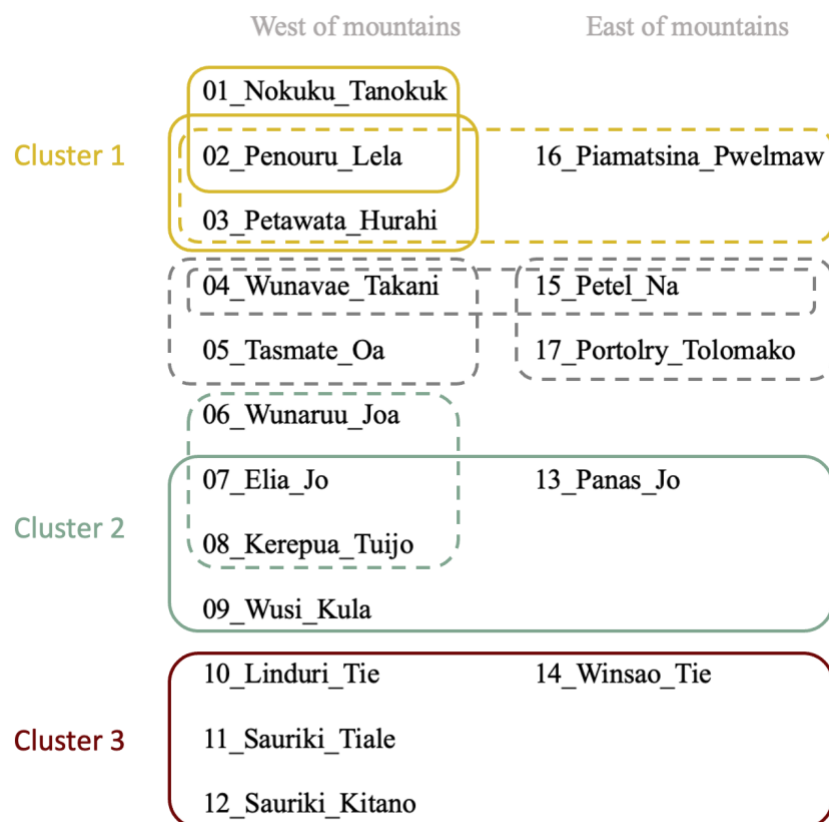


Figure 3. Relationships between the varieties based on the lexicostatistical analysis. Solid lines group varieties that share 75% or more cognacy, dashed lines indicate 70-74.9% shared cognacy. Differential colouring is used to denote the identified clusters.

As can be seen in Figure 3, three clusters of varieties clearly emerge. I number these from north to south and discuss them in this order below.

Cluster 1 in the northwest is centred around 01_Nokuku_Tanokuk, 02_Penouru_Lela and 03_Petawata_Hurahi. This cluster is somewhat less cohesive than the other two. The pairs 01+02 and 02+03 both share 77-78% cognacy within themselves. However, the 70.6% shared cognacy between varieties 01 and 03 is relatively low. Besides, varieties 02 and 03 have been influenced by 01, which is the local *lingua franca* and has even replaced variety 02 (cf. Section 3.1). The speakers of varieties 02 and 03 have migrated in remembered times to their present location and their varieties likely converged with 01 due to contact in which 01 has been the dominant variety (cf. Section 3.1). On the other side of the mountains, 16_Piamatsina_Pwelmau shares 73-74% cognacy with 02 and 03, but much less, 61.7%, with 01, and can be considered as a marginal member of this cluster.

In Cluster 2, 07_Elia_Jo, 08_Kerepua_Tuijo, 09_Wusi_Kula and 13_Panas_Jo all share at least 78% cognacy. 06_Wunaruu_Joa may also be considered a marginal member of this cluster, sharing around 71% cognacy with 07 and 08. The varieties in this cluster are geographically contiguous, except 13_Panas_Jo, which is separated from the rest by the mountains, but whose speakers are known to have migrated away from Elia recently (cf. Section 3.1). On the other hand, the current coastal villages are known to have been formed through migrations from the hills around the 1970s, which suggests that these varieties were close enough to each other already at that earlier stage.

The clusters that were identified in §5 are largely confirmed by the phonological-typological patterns in Table 3. The varieties of Cluster 3 have identical consonant inventories, which also stand out for being significantly smaller than those of most other varieties. In Cluster 2, two internal subgroups (08+09 and 07+13) emerge based on the **s > h* innovation and a partial **m > n* change respectively (see below). Cluster 1 is less cohesive in terms of synchronic phonological typology.

Table 4. Matrix of key sound changes. Code 0 means retention of proto-sound. Notes: (1) in the labiovelarisation column, no distinction is made between languages that always have labiovelarised labials where expected and those that have exceptions; (2) only evidence of linguolabial shift is taken into account, not its extent; (3) in variety 16, the [ndr] reflex of *d is considered a retention of the PNCV state¹⁷, hence coded 0; (4) in the pair 15+17 grey shading is used to indicate differences between pair members.

sound changes	1.*d>[ɟ] 2.*d>k 3.*d>ʔ/Ø	1.*q>ʔ 2.*q>k	1.*k>x/y/h 2.*k>ʔ/Ø	1.*g>n 2.*g>x/h/y	loss of labio- velarisation	linguo- labial shift	1.*s>h
01 Nokuku Tanokuk	3	1	0	1	0	0	0
02 Penouru Lela	3	1	1	1	0	0	0
03 Petawata Hurahi	3	1	1	0	0	0	0
04 Wunavae Takani	2	2	2	1	0	0	0
05 Tasmate Oa	2	2	2	1	0	0	0
06 Wunaru Joa	1	2	1	1	0	0	0
07 Elia Jo	1	2	2	1	0	1	0
08 Kerepua Tuijo	1	2	2	1	0	0	1
09 Wusi Kula	1	2	2	1	0	0	1
10 Linduri Tie	2	2	2	1	1	0	0
11 Sauriki Tiale	2	2	2	1	1	0	0
12 Sauriki Kitano	2	2	2	1	1	0	0
13 Panas Jo	1	2	2	1	0	1	0
14 Winsao Tie	2	2	2	1	1	0	0
15 Petel Na	1	1	1	2	1	0	0
16 Piamatsina Pwelmau	0	2	1	0	0	0	0
17 Portolry Tolomako	1	2	1	2	1	1	0

The patterns observed in Table 4 largely match the three clusters identified in Section 5. The following discussion aims to compare the patterns that emerge from the two analyses, building also on results outlined in Rangelov (2026b).

- Cluster 1 is again less cohesive than the other clusters. We see some shared innovations with exceptions. Variety 16 has retained PNCV *d as [ndr] and has [k] as a reflex of *q (unlike varieties 01,02,03), 01 has retained *k and 03,16 have retained *g. Furthermore, as mentioned, 01 is the only variety that has developed a sixth vowel. These results suggest that Cluster 1 constitutes a linkage. Here, variety 16 also stands out as the most conservative variety in the sample.
- Cluster 2 is more coherent; however, two pairs of varieties clearly emerge here based on two innovations: the **s > h* innovation in 08+09 and the **m > n* shift (which surely included an intermediate linguolabial stage, cf. Rangelov, 2026b; Clark, 2009b; Lynch, 2020) in the pair 07+13. Both of these changes are somewhat unusual in the geographic vicinity, their speakers report awareness of them and that they act as shibboleths,

¹⁷ PNCV *d reflects both POc *d and *dr, the latter being a prenasalised coronal trill with a plosive-like release, such as [ndr]. Clark (2009a) does not list *dr as a PNCV phoneme, but notes that the distinction between *d and *dr persisted in some NCV languages (Clark, 2009a, p. 11).

distinguishing the two pairs from each other, and from other varieties, as separate *lanwis* (Rangelov, 2026b). Variety 06, which was identified as a satellite variety to Cluster 2 by shared cognacy, has not participated in these two innovations, but shares three other innovations with the cluster.

- Cluster 3 is coherent in terms of phonological history, with its varieties sharing all innovations; this cluster also emerges as the most innovative with five shared innovations.
- Of the four varieties that are not part of the three clusters, 04+05 on the west coast form a tight pair. The east coast pair 15+17 is somewhat looser with the linguolabial shift being restricted to 17, and different reflexes of *q. Lastly, the relatively close relationship between varieties 04 and 15 attested in shared cognacy analysis is not supported by shared innovations at all.

It needs to be pointed out that most of the shared innovations discussed in the above paragraphs do not define the clusters exclusively, but rather span across clusters and/or pairs. This supports Clark's (2009b) argument that the languages of Santo form a linkage. Data from more varieties, including undocumented varieties that may fall within the clusters discussed here, would be necessary to further explore this issue and propose classification hypotheses.

7 Lexical variation

This section examines patterns of lexical variation among the varieties, focusing on concepts for which the data is divided between more than one cognate set, regardless of whether the variation results from lexical innovation, retention and/or borrowing. The approach follows that of Clark (2009b), who analysed lexical variation across Santo using concepts whose reflexes fall into two or three cognate sets. From Clark's original set of 38 concepts, the 14 concepts that are relevant to the present study are included, i.e. those for which there is data and for which variation is attested within the sample. Twenty-three other concepts were identified in the dataset that show a clear lexical split between two main cognate sets.

The purpose of this endeavour is primarily to perform a qualitative test of the lexicostatistical analysis in Section 5 (inasmuch as the lack or presence of cognacy is already reflected in that analysis) and to attempt to uncover more detail.

The full dataset can be found in the supplementary material. Tables 5–7 display a subset of examples that demonstrate most of the observed patterns, discussed below. I start by looking for support for the clusters identified by the lexicostatistical analysis, followed by a discussion of other patterns.

Cluster 1 (varieties 01-03,16) shares innovations/retentions in a few concepts, cf. Table 5. In a few cases varieties 04 or 05 patterns with Cluster 1. Variety 16 is not part of the same cognate set in FISH, where Cluster 1 appears to have retained reflexes of PNCV *ika unlike variety 16 and most other varieties. The concept SIX represents the organisation of the cardinal numeral systems of the respective varieties, i.e. Cluster 1 + variety 05 employ the original Austronesian 10-based system, while the rest of the varieties have a system that uses 5 as an anchor, which is widespread but unevenly distributed among Vanuatu languages (cf. Barlow, 2023). When looking at the entire dataset, the finding that Cluster 1 hangs less tightly together is reflected also in the lexical variation data: there are four concepts where variety 01 does not pattern with the rest of the varieties from the cluster; there is also a mismatch between varieties 02 and 03 in two concepts.

Table 5. Examples of lexical variation where Cluster 1 is distinguished from the rest of the varieties

concept	LAUGH *mana *uru *nisa	NIGHT *dodor *bogi	INTESTIN ES *tinae *tali	TAIL *labwe *v(w)[u,i]di -	BLOOD *makavi *daRa	SIX *ono *5 + *rave	MAN *lamane *taudu	FISH 1. *ika 2. *mazi
01_Nokuku_Tanokuk	man	ʔoʔor	tine-ʔ	lap ^w en	mekoβ	ono	leman	ik
02_Penouru_Lela	man	oʔor	tine-ʔ	lap ^w ena	mahaβ	ono	laman	iha / ma:tʃe
03_Petawata_Hurahi	mana	n ^a ʔo:n ^a ʔo	tine-ʔu	lap ^w ena	mahaβi	uono	lamane	iha
04_Wunavae_Takani	uru	poni	tali-m	βucina	kai-m	mo-lima-raβe	lamane	ma:tʃi
05_Tasmate_Oa	mana	pon	tali-na	βocina	maʔaβi	ono	taʔua	iʔa
06_Wunaru Joa	uru	pon	tali	wutʃina	tʃae	lima-raβe	tautʃu	ma:tʃi
07_Elia_Jo	uru	poni	tajtal	wutʃina	tʃaʔa	lima-rave	tautʃu	na:tʃi
08_Kerepua_Tuijo	uru	poni	tali-m	wutʃina	tʃae-m	lima-raβe	tautʃu	matʃi
09_Wusi_Kula	uru	poni	tali-na	wutʃina	tʃae	lima-raβe	tautʃu	ma:tʃi
10_Linduri_Tie	nisa	po:ni	tali-ga	βicina	kae-ka	lima-raβe	tauku	ma:tʃi
11_Sauriki_Tiale	konisa	pon	tali-ka	βicina	kae-ka	lima-raβe	tauku	ma:tʃi
12_Sauriki_Kitano	konisa	pon	tali-ka	βicina	kae-m	mo-lima-raβe	tauku	ma:tʃi
13_Panas_Jo	uru	pon	tali-m	witʃina	tʃai-m	lima-raβe	tautʃo	na:tʃi
14_Winsao_Tie	nisa	taponi	takli-ʔga	βicina	kae-ʔga	ta-lima-raβe	tauku	ema:tʃe
15_Petel_Na	uru	pohi	tajtali-m	wutʃina	tʃae-m	ma-raβe	tatʃua	matʃi
16_Piamatsina_Pwelmau	mana	n ^a ro:n ^a ro	tine-ku	lap ^w ena	mahaβi-ku	o:no	lamane	matʃi
17_Portolry_Tolomako	uru	poyi	tali-m	βutʃina	tʃaj	la-rave	tatʃua	natʃi
included in Clark (2009b)?			yes					yes
comment	Cluster 3 Cluster 1 (+05)	Cluster 1	Cluster 1	Cluster 1	Cluster 1 (+05)	Cluster 1 (+05)	Cluster 1 (+04)	Cluster 1 (-16+05)

Cluster 2 (varieties 06-09,13) shares innovations/retentions in the concepts WATER and YAM and clearly stands out from its varieties' immediate neighbours in HOUSE and FIRE (Table 6). Variety 06, which is an outlier within this cluster only patterns with the rest of the cluster in one of these concepts. Looking at the entire dataset, there are only three concepts in which there is variation within the core varieties (07-09,13) of this cluster (FLOWER, BUSH NUT, BELLY). However, even in these cases the two identified subgroups within this cluster (08+09 and 07+13 respectively) always match.

Cluster 3 (varieties 10-12,14) is represented by its own cognate set in the concepts STEAL, OWL, FLOWER, EARTH, LAUGH and it shares a cognate set with variety 01 for SUN (Table 6). When looking at the entire dataset there is very little variation within the cluster: variety 12 stands out for idiosyncratic innovations, retentions or borrowings in 3-4 concepts; for the concept CANOE, there are Indigenous terms for two varieties of this cluster, with variety 10 retaining a reflex of *waqa 'canoe' while in variety 14 *ove 'canoe tree (Gyrocarpus)' changed to CANOE by synecdoche (as canoes are made from the wood of this tree), an innovation shared with many other varieties in western Santo (cf. Clark, 2009a, 2009b).

Table 6. Examples of lexical variation where Cluster 2 and Cluster 3 respectively are distinguished from the rest of the varieties

[illegible]

Finally, at least a third of the concepts in the dataset demonstrate a somewhat clear north-south split between the varieties (Table 7). If we compare this with the results of the lexicostatistical analysis (Figure 2 in Section 5), this division seems to be organised around a shared cognacy rate threshold of approx. 50%, i.e. the varieties within each of these two groups tend to have shared cognacy of over 50%, while those that belong to different groups exhibit less than 50%. This result is likely due to both higher-level subgrouping of Santo languages (cf. Clark, 2009b) and borrowing between geographically contiguous varieties before and after the documented recent migrations.

Table 7. Examples of lexical variation whose distribution follows a geographic north~south pattern

[illegible]

8 Discussion

The aim of this study was to advance our knowledge about the linguistic diversity of Vanuatu's largest island, Espiritu Santo, which is already known as a global hotspot in this regard, while also being severely underdocumented. The study is based on data from 17 varieties in a selected area of the western part of Santo. Data coverage was larger in breadth and depth compared to the previous survey published in Tryon (1976). Following Tryon, a lexicostatistical analysis based on cognate-coded basic vocabulary was performed. The results were matched with patterns in the diachronic and synchronic phonologies of the 17 varieties (based on Rangelov, 2026b) and, following Clark (2009b), on patterns of lexical variation.

The shared cognacy analysis revealed three main clusters of varieties, to which all but four varieties belong; between some of these four varieties there were more tenuous connections. It is worth reiterating here that no subgrouping hypothesis is being offered here, as we lack data about the current distribution of other varieties which may belong to the same clusters or be part of larger linkages. Earlier data can be used to complement the current data set in refining earlier subgrouping hypotheses, but this is beyond the scope of this study.

All three clusters straddle the mountain chain, which is unexpected, considering that in previous listings of the languages of Santo (Tryon, 1976, 2010; Clark, 2010b), the mountain chain, especially outside of its northern and southern extremes, constitutes a clear linguistic boundary. The main reason for the current geographic separation of communities speaking very similar varieties is a relatively recent series of migrations. In fact, some of Tryon's data were collected before many of the migrations from hill settlements towards the coasts. For example, the 1970s saw people from the same, or apparently very closely related, ethnolinguistic communities living in the hills, splitting and migrating to opposite sides of the mountains (Taman et al., 2025), as in the case of the people of Winsao and the rest of the speakers of the varieties in Cluster 3. A similar scenario was also reported by the people of Piamatsina and Petawata, as well as Petel and Wunavae, respectively. In the case of Panas, we see a subsequent migration over the mountains in the 2010s, driven by droughts on the west coast. It is likely that more such migrations will happen in the near future due to the high risks associated with climate change and natural disasters (Aleksandrova et al., 2021; Taman et al., 2025).

The three clusters are also corroborated to a large extent, albeit with some exceptions, by diachronic and typological phonological evidence. Lexical innovations patterned in the same way, although a north~south separation emerged that blurs some boundaries, especially of Cluster 1 in the north. The north~south pattern is likely evidence of horizontal transfer that may otherwise not always be detectable in the data.

The evidence clearly shows that the southernmost Cluster 3 appears to be the most cohesive, while Cluster 1 in the north is the loosest. In Cluster 2, a clear split between the pairs 07+13 and 08+09 emerged based on two salient phonological innovations.

So far, the relationships between the varieties were discussed around the three clusters. These clusters are likely part of higher-order subgroups/linkages that involve varieties outside of the current sample. In the following paragraphs I discuss relationships within the clusters and take as a starting point the language-dialect-*lanwis* framework outlined in Section 2.3. The shared cognacy results demonstrate that the linguistic diversity in the surveyed area is substantially higher than previously known or even suspected. If we were to apply Tryon's 80% shared cognacy threshold for identifying different languages, then according to the results in Figure 2, seven extra languages would have to be proposed for the area. Namely, following Glottolog (Hammarström et al., 2026), the previously known (labelled) languages spoken in the area are seven:

- Nokuku (= 01_Nokuku_Tanokuk)
- Tasmate (= 05_Tasmate_Oa)
- Kula/Wusi (= 08_Kerepua_Tuijo and 09_Wusi_Kula)
- Tie (a.k.a Navut) (= 10_Linduri_Tie and 14_Winsao_Tie)
- Tiale (a.k.a Malmariv) (= 11_Sauriki_Tiale)
- Piamatsina (= 16_Piamatsina_Pwelmau)
- Tolomako (= 17_Portolry_Tolomako)

Based on an 80% threshold, these are confirmed as separate languages, except that the varieties corresponding to the Tie and Tiale languages would have to be merged into a single language.¹⁸ Then, the following seven “new” languages can be postulated:

- 02_Penouru_Lela (dormant)
- 03_Petawata_Hurahi
- 04_Wunavae_Takani
- 06_Wunaruu_Joa
- 07_Elia_Jo+13_Panas_Jo – both share over 80% cognacy with 08, but not with 09 within Cluster 2
- 12_Sauriki_Kitano (dormant) – shares 80.5% cognacy with 11, but not with the other two varieties in Cluster 3
- 15_Petel_Na

If we were to adjust the shared cognacy threshold down to 75%, a more conservative estimate¹⁹, then 12_Sauriki_Kitano would fall within Tie/Tiale (Cluster 3), and 07_Elia_Jo+13_Panas_Jo would be counted under the Kula/Wusi language within Cluster 2. Cluster 1 then remains with blurrier boundaries between languages in a dialect continuum/linkage situation as the pairs 01+02 and 02+03 both share over 75% cognacy within themselves but the pair 01+03 does not do so. Here, it would be more reasonable to draw a boundary between varieties 01 and 02, as the phonological history of 01 suggests a separate development (retention of *k, changes in the vowel inventory). In fact, the shared cognacy between 01 and 02/03 may be somewhat inflated because of undetected horizontal transfer where 01 is the source variety, being more prestigious (it is the language that the rememberers of 02 have switched to, and the speakers of 03 use daily). In this case, we would then have to posit “only” four additional languages for the selected area, as in Figure 4.

¹⁸ The Glottolog languages Tie and Tiale likely encompass varieties beyond the ones in the current sample (cf. Section 3.1), which possibly are also part of a dialect continuum or linkage, so based on these results it cannot be argued that the two Glottolog languages need to be merged. This demonstrates a limitation of this paper’s sampling method. More data from other varieties can shed light on this issue.

¹⁹ In Section 5 it was pointed out that the current data yield systematically lower shared cognacy rates compared to Tryon’s, and the possible reasons for this were discussed (larger dataset, language divergence during the past decades). A precise quantification of all the effects has not been attempted here and 75% was chosen as a rounded estimate, see also Section 5.

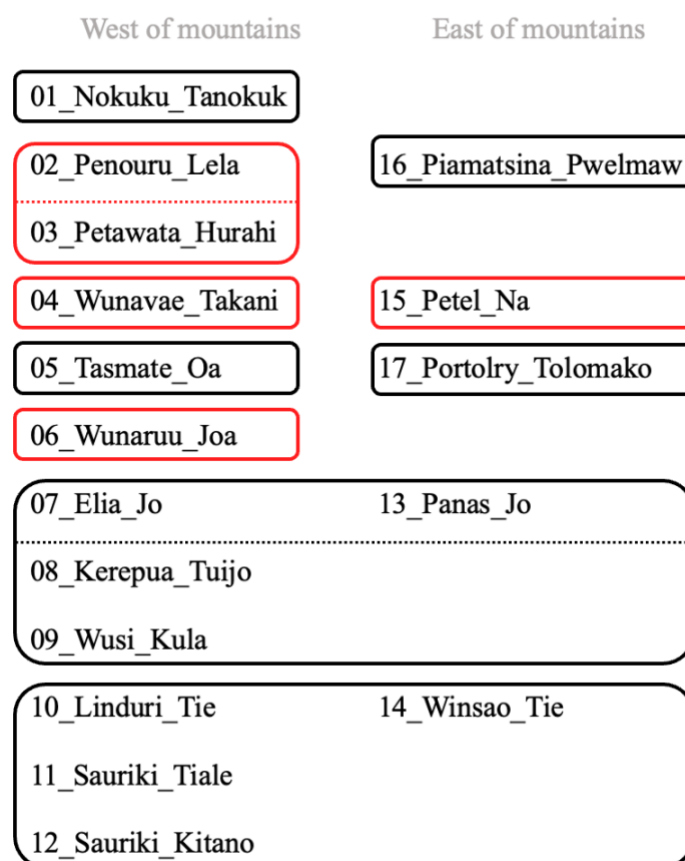


Figure 4. The six (Tie and Tiale are considered as one here) known languages in the studied area are encircled in black. The four new languages postulated based on a 75% shared cognacy threshold are encircled in red: 02+03, 04, 06, 15. Dotted lines mark dialect/*lanwis* boundaries.

In terms of dialectal variation, the historical and typological-phonological analysis adds important detail to what can be inferred from the shared cognacy results. Varieties 02 and 03 could be considered different dialects of the same language, with 03 partially retaining the PNCV velar nasal *g, while 02 having completely merged this with *n. This may be due to a closer historical relationship (genealogical and/or areal) between 03 and 16, which also retains *g. These three varieties can be argued to form a linkage (at a higher level, together with 01 and possibly other undocumented varieties).

Then, within the core of Cluster 2, two dialects emerge – 08+09 and 07+13 – based on the shared phonological innovations and lexical variation.

Within Cluster 3, the conditioned prenasalised plosive reflexes of *b and *q in the geographic outlier variety 14 stand out (Rangelov, 2026b). Their distribution is not considered substantial enough to postulate dialectal variation within this language. It is unknown whether or to what extent these reflexes are descendants of some dialectal variation that existed already before the 20th-century migrations and/or more recent contact effects with other varieties that retain a prenasalisation distinction.

Next, I integrate emic perceptions of lectal distinctions with the etic comparative linguistic data. In Section 3.1, speakers' attitudes and ideologies about *lanwis* boundaries and perceived mutual intelligibility were reported. Mutual intelligibility is problematic because it may reflect

effects of small-scale egalitarian multilingualism rather than historical relations, so here the discussion is mostly limited to perceived *lanwis* boundaries. The results outlined above are corroborated by these perceptions to a large extent, but not completely.

Starting with the matches, the Glottolog languages Nokuku (01), Tasmate (05), Kula/Wusi (08,09), Piamatsina (16) and Tolomako (17) are considered separate *lanwis* by their speakers and by speakers of other *lanwis*; the dialects within the core of Cluster 2 and within the posited 02+03 language are also perceived as separate *lanwis*, as are the languages 04, 06, and 15. A more tenuous perceived historical relationship has also been corroborated: the speakers of variety 15 reported historical ties with the speakers of 04 on the west coast (although they did not consider the two varieties to be the same *lanwis*); this is confirmed by a 71.1% shared cognacy but not by shared phonological innovations.

Turning to the mismatches, the main disagreement between the results and speakers' perceptions is in the most cohesive Cluster 3. Namely, the speakers of varieties 10/14 (Tie) and 11 (Tiale) all agree that theirs are separate *lanwis*. Similarly, 12 (Kitano) was reported as a separate *lanwis* by its rememberers and other speakers of 11, while the data indicates that all these can be considered as part of the same language (as per the criteria adopted here). Specifically, the shared cognacy rates between the two Tie varieties and the Tiale variety are over 80%. As pointed out above (Footnote 18) it is possible that these varieties are part of a larger cluster (dialect chain or linkage) that extends further south and east. Its northern boundary appears to lie somewhere around a Linduri-Winsao line, as there is a clear separation from Cluster 2 (barely over 60% shared cognacy between the neighbouring villages Wusi and Linduri; 58% between the neighbouring villages Winsao and Panas) and other varieties (e.g. less than 50% shared cognacy between Winsao and the nearby historical location of 17_Portolry_Tolomako, where reportedly other dialects of the Tolomako language are spoken, cf. Clark, 2006).

Another mismatch involves varieties 03 and 16. The speakers of 16 reported that these two varieties were the same *lanwis*, but the speakers of 03 did not report 16 as being the same *lanwis* as theirs. The shared cognacy of 74.2% does not lend very strong support either. However, we see possible signal of a close historical relationship between 03 and 16 in the retention (partial in 03) of the PNCV velar nasal *g, and some signal in the lexicon (Rangelov, 2026b, Section 5.1.3). It is possible that variety 02 may also have had a close historical relationship with the 03+16 pair, as evidenced by the shared cognacy rates. The rememberers of 02 report close historical ties with speakers of lects on the Big Bay coast, but did not name 16 specifically, while the speakers of 16 may not have reported similar perceptions with 02 due to its status as a dormant variety.

Further south, the reports of the speakers of 05 that their variety is closely related to (however, not the same *lanwis* as) the "Tolomako *lanwis*" spoken in Big Bay, is out of line with the shared cognacy rates between 05 and 17 (58.4%). Indeed, they surely meant a different Tolomako variety than 17, which is spoken today at Port Olry, but such major disparity between 17 and other varieties that are normally labelled with the Tolomako linguonym is not well supported by existing reports (Clark, 2006; Meyerhoff, 2009). Similarly, the speakers of variety 06 reported that they speak the same *lanwis* as people in Malau village in Big Bay, which is normally included under the Tolomako language area (Tryon 1976, 2010; Lynch & Crowley, 2001, p. 46; François et al., 2015, p. 3), but the observed shared cognacy between 06 and 17 of 63.5% does not support this. The latter two cases remain open questions that may be answered by future surveying of the linguistic diversity in the Big Bay region, which may appear to be greater and more complex than we are currently aware of.

Summarising the extent to which emic perceptions confirm the result of the linguistic analysis, firstly, we see that the term *lanwis*, which reflects emic perceptions, appears to apply differences down to the dialect level from an etic perspective. Secondly, the analysis demonstrates some false positives (i.e. cases where a perceived *lanwis* boundary does not correspond to a language or dialect boundary), but, importantly, no unambiguous false negatives (i.e. cases where the data show a language/dialect boundary, but this boundary is not perceived by the speakers).

9 Conclusion

The results of this pilot study clearly show that the linguistic diversity in the selected area of Santo is much greater than previously known. Four additional languages and greater dialectal variation were postulated, compared to what was previously known.

One limitation of this study is that the sampled varieties offer less coherent geographical coverage on the eastern side of the mountains compared to the west coast. Another limitation is that it has not been possible to form any substantial subgrouping hypotheses based on the sample. Both of these limitations can be addressed by extending the linguistic survey of Santo to cover the entire island, and especially its western and southern parts that are its most underdocumented areas. This would be a logistically challenging endeavour, considering the number of settlements on this island, but the result that speakers' reports about *lanwis* boundaries largely corroborate the result of lexical data analysis, and, crucially, that there were no unambiguous false negatives, means that emic perceptions can be a meaningful proxy for variety sampling for future data collection and analysis. Tryon's (2010, p. 284) *lanwis* boundaries based on reports by *filwoka* from Santo did not add any new *lanwis* to the area covered by this survey; one reason for this may be that Tryon likely did not have access to *filwoka* representing all regions of Santo. Most of Tryon's (*ibid.*) six additional *lanwis* are in the higher altitude areas of the Big Bay and South Santo regions, but in this study, we have also seen evidence of a dense mosaic of *lanwis* distribution in other areas too.

Uncovering this mosaic can open many avenues for future research in various fields of linguistics and adjacent disciplines. Considering that all of Santo's Indigenous varieties are endangered to various degrees, including some having recently become extinct or dormant, the documentation and analysis of language diversity on the entire island is an urgent matter.

Supplementary material

The supplementary material can be found at the following link:

<https://doi.org/10.5281/zenodo.23383896>

The following files are included:

- An ODS file (open in LibreOffice, Microsoft Excel or similar) called 'WSanto-cognatecoded.ods', containing all lexical data and metadata,
- An ODS file called 'WSanto-lexvariation.ods' containing the lexical variation dataset,
- A folder called 'R' with the R code for calculating cognacy rates and statistics about the dataset used for the lexicostatistical analysis, including:
 - An R notebook ('WSanto-data-process.nb.html') containing the code, pseudocode and comments,
 - The input file ('WSanto-cognatecoded.csv'), which is a CSV export of the ODS data file above,

- The two files output by the code: ‘[cognacy-rates.csv](#)’ and ‘[cognate-code-matrix.csv](#)’.
- A PDF file ‘[SantoLingDiversity-MethodsEtc.pdf](#)’ with a detailed description of the data and methods related to data collection, annotation, organisation and cognate coding.

Acknowledgements

Data collection was funded by the Department of Linguistic and Cultural Evolution at the Max Planck Institute for Evolutionary Anthropology in Leipzig, Germany. Fieldwork and research were carried out under an ethics approval by the Max Planck Society and a research agreement with the Vanuatu Cultural Centre. I would like to thank the Vanuatu Cultural Centre, the Tabwemasana Council of Chiefs, the Santo Sunset Environment Network, the Area Administrators of West Coast and Northwest Santo areas, and all members of the communities that I visited, and especially Allan Taman and family, Sabrina Taman, Chief Ben Frank Palla and family, Chief Noel Sari, Deacon James Linrow, George Frank, Elder Vaiaken Aulu, Tani Marklin, Joshua Lewalev, Stephen Tavue, Chief Richard Poi, Charly Lapiho, Chief Taitas Rojo, Stefan Pipite, Chief Song Vanua, Pune Ropwel, Peka James, and Chief Puha Lapan. I also want to thank Russell Gray, Mary Walworth, Miriam Meyerhoff, Ross Clark, Elizabeth Pascal, Ben King, Isaac Stead, Russel Barlow, Jakob Halfmann, Natalie Korobzow, and Hans Schmidt for their encouragement, company, and/or comments on an earlier draft of this paper. I am grateful to the editor and two anonymous reviewers for their constructive comments and suggestions, which helped improve the manuscript. All remaining errors are my own.

References

- Adelaar, A. K. (2009). Language documentation in the West Austronesian world and Vanuatu: An overview. In M. Florey (Ed.), *Endangered Languages of Austronesia* (pp. 12–42). Oxford University Press.
<https://doi.org/10.1093/acprof:oso/9780199544547.003.0002>
- Aleksandrova, M., Balasko, S., Kaltenborn, M., Malerba, D., Mucke, P., Neuschäfer, O., Radtke, K., Prütz, R., Strupat, C., Weller, D., & Wiebe, N. (2021). *World Risk Report 2021*. Bündnis Entwicklung Hilft.
<https://reliefweb.int/sites/reliefweb.int/files/resources/2021-world-risk-report.pdf>
- Barlow, R. (2023). Papuan-Austronesian contact and the spread of numeral systems in Melanesia. *Diachronica*, 40(3), 287–340. <https://doi.org/10.1075/dia.22005.bar>.
- Charpentier, Jean-Michel. (1982). *Atlas linguistique du Sud-Malakula / Linguistic atlas of South Malakula (Vanuatu)*. SELAF.
- Chung, Y. S. H. (2005). *A descriptive grammar of Merei (Vanuatu)*. Pacific Linguistics.
- Clark, R. (2006). *Notes on Big Bay language (Tolomako)* [Unpublished manuscript].
- Clark, R. (2009a). **Leo Tuai: A comparative lexical study of North and Central Vanuatu languages*. Pacific Linguistics.
- Clark, R. (2009b, June 22–26). *Espiritu Santo as a “linguistic area”* [Conference presentation]. 11th International Conference on Austronesian Linguistics (11ICAL), Aussois, France.
- Codrington, R. (1885). *Espiritu Santo, Marina*. Clarendon Press.
- Crowley, T. (1998). A salvage sketch of Nati (Southwest Malakula, Vanuatu). In D. T. Tryon (Ed.), *Papers in Austronesian Linguistics* (No. 5), pp. 101–148. Pacific Linguistics.
- Crowley, T. (2000). The Language situation in Vanuatu. *Current Issues in Language Planning*, 1(1), 47–132. <https://doi.org/10.1080/14664200008668005>

- Crowley, T. (2006a). *Naman: A vanishing language of Malakula (Vanuatu)*. Pacific Linguistics.
- Crowley, T. (2006b). *Nese: A diminishing speech variety of northwest Malakula (Vanuatu)*. Pacific Linguistics.
- Crowley, T. (2006c). *Tape: A declining language of Malekula (Vanuatu)*. Pacific Linguistics.
- Crowley, T. (2006d). *The Avava language of central Malakula (Vanuatu)*. Pacific Linguistics. <http://hdl.handle.net/1885/146274>
- European Space Agency. (2024). *Copernicus Global Digital Elevation Model*. OpenTopography. <https://doi.org/10.5069/G9028PQB>
- François, A. (2012). The dynamics of linguistic diversity: Egalitarian multilingualism and power imbalance among northern Vanuatu languages. *International Journal of the Sociology of Language*, 214, 85–110. <https://doi.org/10.1515/ijsl-2012-0022>
- François, A., Franjeh, M., Lacrampe, S., & Schnell, S. (2015). The exceptional linguistic diversity of Vanuatu. In A. François, S. Lacrampe, M. Franjeh, & S. Schnell (Eds.), *The languages of Vanuatu: Unity and diversity*, pp. 1–21. Asia-Pacific Linguistics.
- Gordon, J. D. (1870). *A phonetic primer for acquiring the art of reading by syllables in one of the dialects of Espiritu Santo, New Hebrides*. J. L. Sherriff.
- Hammarström, H., Forkel, R., Haspelmath, M., & Bank S. (2026). *Glottolog 5.3*. Max Planck Institute for Evolutionary Anthropology. <https://doi.org/10.5281/zenodo.18840935>
- Hog Harbour Community, Prenter, I., & Meyerhoff, M. (2021). *First dictionary of Nkep*. The community of Hog Harbour.
- Holmes, R. (2014). *Espiegle's Bay: Grammatical topics* [Unpublished master's thesis]. University of Waikato.
- Lindstrom, L. (1994). Traditional cultural policy in Melanesia (Kastom polisi long kastom). In L. Lindstrom & G. M. White (Eds.), *Culture-kastom-tradition: Developing cultural policy in Melanesia*, pp. 67–81. University of the South Pacific.
- Ludvigson, T. (1989). *Vara Kiai: A Kiai wordlist*. Department of Anthropology, University of Auckland.
- Lüpke, F. (2016). Uncovering small-scale multilingualism. *Critical Multilingualism Studies*, 4(2), 35–74.
- Lynch, L., & Crowley, T. (2001). *Languages of Vanuatu: A new survey and bibliography*. Pacific Linguistics.
- Lynch, J. (2020). The Bilabial-to-linguolabial shift in Southern Oceanic: A subgrouping diagnostic? *Oceanic Linguistics*, 58(2), 292–323.
- Meyerhoff, M. (2009). *Aspects of the language spoken at Matantas, Big Bay (Vanuatu)* [Unpublished manuscript].
- Meyerhoff, M. (2017). Writing a linguistic symphony: Analyzing variation while doing language documentation. *Canadian Journal of Linguistics/Revue canadienne de linguistique*, 62(4), 525–549. <https://doi.org/10.1017/cnj.2017.28>
- Monnier, P. (1994). *Les lettres du Père Alphonse Ardouin (1880 – 1940)*. Maison Mariste.
- Moore, A. (2019). *A grammar sketch of Uripiv* [Master's thesis, University of Waikato]. Research Commons. <https://hdl.handle.net/10289/12292>
- Pakendorf, B., Dobrushina, N., & Khanina, O. (2021). A typology of small-scale multilingualism. *International Journal of Bilingualism*, 25(4), 835–859.
- Posit Team. (2026). *RStudio: Integrated development environment for R*. Version 2026.04.0
- R Core Team. (2026). *R: A language and environment for statistical computing*. Version 4.4.3.
- Rangelov, T. (2026a). *Language density hotspots*. Zenodo. <https://doi.org/10.5281/zenodo.18647227>

- Rangelov, T. (2026b). The phonological history of 17 varieties in western Espiritu Santo (Vanuatu). *Language and Linguistics in Melanesia*, 44, 52–104.
- Rangelov, T., Ridge E., & Takau, L. (2025a). Introduction to the three Special Issues on Vanuatu languages. *Te Reo*, 68(2), 1–6.
- Rangelov, T., Ridge E., & Takau, L. (2025b). Linguistics in Vanuatu 45 years after Independence. *Te Reo*, 68(2), 31–78.
- Ray, S. H. (1926). *A comparative study of the Melanesian island languages*. Cambridge University Press.
- Regenvanu, R. (2011). Welkam toktok. In J. Taylor & N. Thieberger (Eds.), *Working together in Vanuatu. Research histories, collaborations, projects and reflections*. 125–132. ANU E Press.
- Ridge, E. (2025). What does lanwis mean? Parallels between Indigenous languages of Vanuatu for young people in Mele Maat. *Te Reo*, 68(2), 133–161.
- Ross, M. D. (1988). *Proto Oceanic and the Austronesian languages of Western Melanesia*. Pacific Linguistics.
- Ross, M., Pawley, A., & Osmond, M. (1998). *The lexicon of Proto Oceanic: The culture and environment of ancestral Oceanic society. Volume 1: Material culture*. Australian National University.
- Ross, M., Pawley, A., & Osmond, M. (2007). *The lexicon of Proto Oceanic: The culture and environment of ancestral Oceanic society. Volume 2: The physical environment*. Australian National University.
- Ross, M., Pawley, A., & Osmond, M. (2008). *The lexicon of Proto Oceanic: The culture and environment of ancestral Oceanic society. Volume 3: Plants*. Australian National University.
- Ross, M., Pawley, A., & Osmond, M. (2011). *The lexicon of Proto Oceanic: The culture and environment of ancestral Oceanic society. Volume 4: Animals*. Australian National University.
- Ross, M., Pawley, A., & Osmond, M. (2016). *The lexicon of Proto Oceanic: The culture and environment of ancestral Oceanic society. Volume 5: People: body and mind*. Australian National University.
- Ross, M., Pawley, A., & Osmond, M. (2023). *The lexicon of Proto Oceanic: The culture and environment of ancestral Oceanic society. Volume 6: People: society*. Australian National University.
- Rousseau, B. (2011). Shifting others: Kastom and politics at the Vanuatu Cultural Centre. In J. Taylor & N. Thieberger (Eds.), *Working together in Vanuatu. Research histories, collaborations, projects and reflections*, pp. 225–237. ANU E Press.
- Taman, A., Taman, S., & Rangelov, T. (2025). Language ecosystems in Western Santo. *Te Reo*, 68(2), 186–245. <https://doi.org/10.5281/zenodo.18723278>
- Tilon, E. (2011). Wok olsem wan filwoka. In J. Taylor, & N. Thieberger (Eds.), *Working together in Vanuatu. Research histories, collaborations, projects and reflections*, p. 223. ANU E Press.
- Touati, B. (2014). *Description du sakao, langue du nord-est Santo (Vanuatu)* [Doctoral dissertation, Université Paris-Sorbonne]. <https://theses.fr/2014PA040244>
- Tryon, D. (1976). *New Hebrides languages: An internal classification*. Pacific Linguistics.
- Tryon, D. (2010). The languages of Espiritu Santo, Vanuatu. In J. Bowden, N. P. Himmelmann, & M. Ross (Eds.), *A journey through Austronesian and Papuan linguistic and cultural space: Papers in honour of Andrew K. Pawley*, pp. 283–290. Australian National University.
- Vanuatu Bureau of Statistics. (2020). *2020 Vanuatu National Population and Housing Census*. <https://vbos.gov.vu/index.php/en/census-and-surveys/census-report>

- Wessels, K. (2013). *Malua Bay: A description of the Malua Bay language (Malekula, Vanuatu)* [Unpublished master's thesis]. University of Waikato.
- Wessels, K. (2021, February 10–12). *The status of Atchin language: North-East Malekula or not?* [Conference presentation]. 2020 Language and Society Conference of the Linguistic Society of New Zealand, Hamilton, New Zealand.
- Wickham H., Averick M., Bryan J., Chang W., McGowan L. D., François R., Golemund G., Hayes A., Henry L., Hester J., Kuhn M., Pedersen T. L., Miller E., Bache S. M., Müller K., Ooms J., Robinson D., Seidel D. P., Spinu V., Takahashi K., Vaughan D., Wilke C., Woo K., & Yutani H. (2019). Welcome to the Tidyverse. *Journal of Open Source Software*, 4(43), 1686. www.doi.org/10.21105/joss.01686