Variation in the Fataluku voiced coronal (j)

James Grama,[1] Tyler M. Heston,[2] and Melody Ann Ross[1]
[1] Sociolinguistics Lab, University of Duisburg-Essen | [2] University of Kansas

This paper represents the first variationist investigation of the voiced coronal phone (j) in Fataluku, a Papuan language of Timor-Leste. Here, we implement the Boruta algorithm at the front end of our analysis pipeline to quantify predictor importance, then use classification trees and mixed-effects regression to disentangle observed effects. Analysis suggests that word position is highly predictive of (j) realization, with glides more likely word-medially and obstruents word-initially. Region is an important predictor word-medially; speakers in Tutuala show nearly categorical [j], indicating strong allophony. Outside of Tutuala, medial tokens vary according to gender and education; among speakers with limited formal education, men show higher rates of glides than women, but speakers with secondary education exhibit higher obstruent rates and no gender differences. Initial tokens, by contrast, are undergoing a change in progress towards affricate realizations. We interpret these findings in the context of locally-specific conceptions of place for Fataluku people in Timor-Leste, particularly that of Tutuala.

Keywords: Fataluku, Timor-Leste, East Timor, Lautém, variation in minority languages, Boruta, random forests, classification trees, geographical variation, sociophonetics, language documentation

1. Introduction

This paper explores the variable patterning in the realizations of (j), a voiced coronal phone with a large number of variants in Fataluku (ISO 639-3 ddg), a Papuan language of Timor-Leste.[1] Despite attested regional phonological variation (Hull, 2005; van Engelenhoven, 2009) and disagreement in the literature

1. We represent the variable under discussion here as (j), following conventions for East Fataluku in van Engelenhoven & Huber (2020) (see also Section 2.3).

https://doi.org/10.1075/aplv.00017.gra | Published online: 11 January 2024
Asia-Pacific Language Variation 9:2 (2023), pp. 125–155. ISSN 2215-1354 | E-ISSN 2215-1362

about the phonemic status of (j) (cf. Heston, 2015; van Engelenhoven & Huber, 2020), there has been little systematic investigation of this variable to date (but see Heston, 2019). Here, we present a variationist sociolinguistic account of (j), focusing on how linguistic and social factors bear on its variants.

Using a picture naming task that yielded nearly 1,000 realizations of (j), we corroborate the wide range of variants attested in Heston (2019), which include a voiced glide and obstruents that vary in manner and voicing (see Figures 4 and 6). We observe that position in the word is a strong predictor of (j) realization; glides are heavily favored word-medially, and obstruents word-initially. Within each of these domains, we observe principled variation, with region, age, gender, and education bearing on (j) variants. Medial instances of (j) exhibit regional stratification, with Tutuala showing functionally categorical proportions of the glide variant. Outside of Tutuala, women are more likely to produce obstruent variants than men, and men show an effect of formal education; those with less education produce higher rates of the glide than those with more formal education. Initial (j), by comparison, shows evidence of a change over time, with younger speakers exhibiting a shift away from fricative realizations towards affricates. Following the growing trend towards incorporating minority languages into the variationist paradigm (e.g., Meyerhoff, 2019), this study underscores how variationist methods can elucidate and strengthen language documentation. We further show that machine learning techniques (see Dickson & Durantin, 2019) when combined with statistical methods commonly used in sociolinguistics can help tease apart factors that constrain variation, even in instances where suitable phonetic data may be limited.

We first briefly discuss the recent history of Timor-Leste and Fataluku in Section 2, then detail data collection, elicitation methods, and coding in Section 3. Our analysis is discussed in Section 4, where we also show how various statistical methods help circumscribe the variation we observe. Finally, Section 5 generalizes our findings against the wider background of Fataluku and Timor-Leste.

2. Background

2.1 A brief history of Timor-Leste following European colonization

Timor-Leste is a country of intricately balanced linguistic relationships. The constitution names two co-official languages: the former colonial language, Portuguese, and a local language, Tetun Dili (Eng: *Tetun Dili* or *Tetun*; Port: *Tétum Praça* or *Tétum*). The constitution also gives the unique legal status of

'working language' to the former occupier language, Indonesian, and the symbolic international language, English (RDTL, 2002). In the western part of the country, the landscape is dominated by the Austronesian languages Mambae and Tetun Terik (a distinct variety from Tetun Dili, with less lexical influence from Portuguese; see discussion in Greksáková, 2018, pp. 409–410). In the eastern part of the country, the main languages are the Papuan languages Makasae and Fataluku. The boundaries between these languages, their dialects, and their relatives are porous and permeable. Apart from the few languages named here, Timor-Leste is also home to as many as 20 named varieties, and no single language comprises a majority mother tongue (Williams-van Klinken & Williams, 2015).

The languages and cultures of Timor-Leste have been shaped by millennia of contact, both with other indigenous groups in the region and with foreign colonizers. In the sixteenth and seventeenth centuries, both the Portuguese and the Dutch laid colonial claims to the island of Timor, vying to control the people of Timor and exploit their natural resources. The official boundaries between Dutch and Portuguese claims remained largely theoretical for much of their history, contested by European politicians until the early twentieth century (Fox, 2003). The western portions of the island claimed by the Dutch won independence in 1949, while decolonial processes prompted Portugal to recognize the eastern half of the island as the independent nation of Timor-Leste in 1975.

Weeks after Timor-Leste's independence, the Indonesian military invaded the newly sovereign state, disrupting local traditions. The conflict during occupation touched every life in the country over the ensuing two decades (Bovensiepen, 2014, p. 294). Timorese ways of interacting with other communities and cultural forbearers were broken as they were forcibly removed from their ancestral lands, restricted to new highly regulated 'resettlements', and severely punished for speaking languages unfamiliar to the Indonesians (*ibid*). In 1999, the United Nations intervened to re-establish Timor-Leste as an independent country, and following Timor-Leste's regained independence in 2001, communities attempted to return to their former homelands, rebuild, reconnect, and reclaim their lives and traditions.

This mix of social and linguistic pressures has created a unique sociolinguistic context: a robust history of local ethnolinguistic exchange; protracted subjugation by a rotating door of foreign powers; forced migration and isolation from traditional communities; and national sovereignty that marked a return to ancestral lands.

2.2 Fataluku

It is against this backdrop that the current study takes place. This paper is part of a larger project investigating phonological variation in Fataluku, a Papuan language with approximately 41,500 speakers in the easternmost portion of the island of Timor (see Figure 1; van Engelenhoven & Huber, 2020). Fataluku is closely related to Makasae and Makalero, both spoken in neighboring regions (Schapper, Huber & van Engelenhoven, 2014). In line with the baseline multilingualism that characterizes much of the population of Timor-Leste, many Fataluku speakers also speak Tetun Dili (the national *lingua franca*), Portuguese (the language of elite education until 1975), Indonesian (the language of public education from 1975–1999), and/or one of the other indigenous languages of Timor (Boon, da Conceição Savio, Kroon, & Kurvers, 2021). Though there are some L2 speakers, especially from neighboring cultural regions or people who marry into Fataluku households, most speakers culturally identify as Fataluku.

The Fataluku-speaking population has experienced significant recent social upheaval. In addition to the aforementioned social changes to Timor-Leste, many Fataluku are leaving traditional subsistence practices and moving to larger, urban centers in search of salaried employment. These social and economic changes are accompanied by increasing shifts towards Tetun Dili, though community attitudes towards Fataluku language and culture remain strong (da Conceição Savio, Kurvers, van Engelenhoven, & Kroon, 2012; Boon et al., 2021).

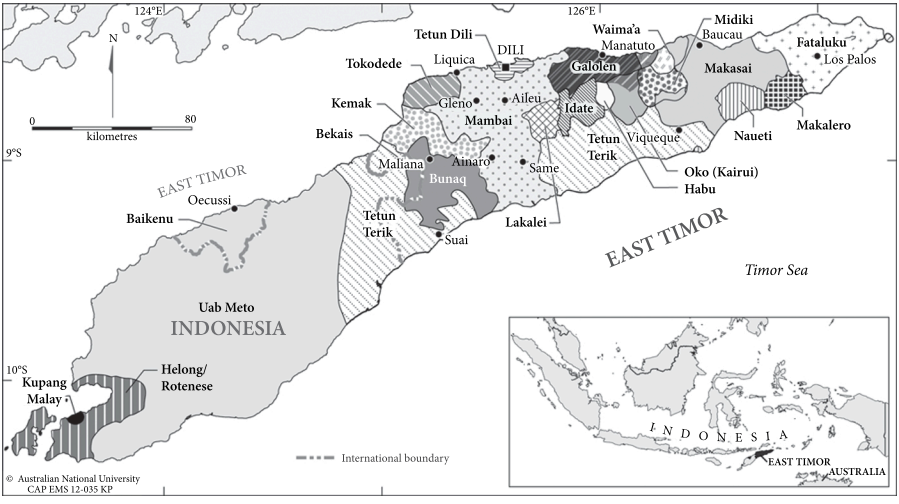


Figure 1. The languages of Timor-Leste (CartoGIS, College of Asia and the Pacific, ANU)

The Fataluku-speaking region is roughly coterminous with the easternmost district of Lautém, whose capital and largest village, Lospalos, lies at its center. Approximately 220 km from Dili (the nation's capital and largest city), Lospalos is primarily accessible via the main road along the northern coast by turning south at the village of Lautem.[2] From Lospalos, there is a passable road east to Tutuala, a smaller road west to Cacavei, and a more difficult road south to Lore I.

What research exists on variation in Fataluku largely identifies five geographically-defined dialects (Hull, 2001; van Engelenhoven, 2009; van Engelenhoven & Huber, 2020; but see McWilliams, 2007, who reports as many as seven, and Bovensiepen, 2014, who reports fourteen). Most researchers agree on the existence of Central (spoken in and around Lospalos), North (Lautem), South (Lore I), and East dialects (Tutuala) but disagree as to whether to include an additional Northwest dialect centered on the village of Baiduro (Hull, 2001; van Engelenhoven, 2009) or a Western dialect centered on Cacavei (Valentim, 2002; see summary in English in van Engelenhoven, 2009, p. 334). The principal evidence for these groupings comes from sound correspondences reported by van Engelenhoven (2009, pp. 334–335; see also van Engelenhoven & Huber, 2020, pp. 365–369). Voiced coronal (j) forms one of the three phonological characteristics that underpin these groupings, with [z] occurring in the North (and Northwest) and Central dialects, and [j] occurring in the dialects of the South and East. Despite the widespread acknowledgment of dialect boundaries, there have been no attempts to date to combine acoustic and statistical analysis to empirically describe variation in Fataluku.

Figure 2 shows a map of the district of Lautém, with the areas represented in the current study labeled. Superimposed on this map are denotations of the dialect regions suggested by past research, with the North and Northwest dialect grouped together under the umbrella of 'North'. Of note, the village Raça serves as putative transition areas, lying geographically between the Central and North dialect zones.

2. Note that the municipality – Lautém – differs from the northern town – Lautem. The municipality originally drew its name from the town which once served as its capital. After 1946, the capital was shifted to Lospalos.

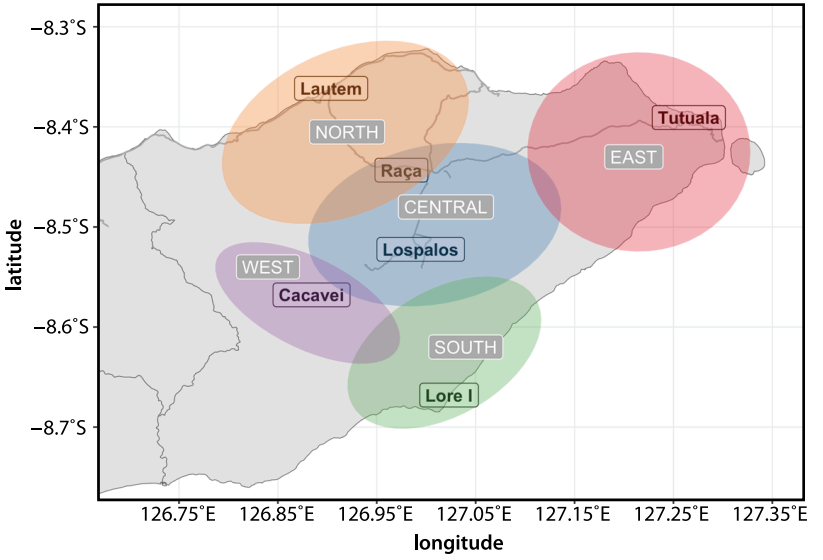


Figure 2. Lautém district with towns in present study and putative dialect regions

2.3 (j) in Fataluku

Fataluku has a (C)V(V)(C) syllable template, with a moderately small phoneme inventory of five vowels (/i, e, a, o, u/) and around fourteen consonants (shown in Table 1). The voiced coronal (j), with its chief variants [j] and [z], is a prominent example of a phoneme that is reported to vary cross-dialectally, variation that has led to differing claims about the phonemic status of the two sounds. Hull (2005) and Heston (2015), both of whom focus on the Central variety of Fataluku spoken in and around Lospalos, analyze /z/ and /j/ as separate phonemes, the latter finding evidence of a few near-minimal pairs, such as [aza] 'rain' and [paja] 'liquid' (Heston, 2015, p. 78). Other work recognizes no such contrast. Campagnolo (1973), who conducted fieldwork in Lore I (representing the South dialect) observes allophony between [z] and [j], identifying that [z] occurs word-initially and after /i/, while [j] occurs after other vowels. Working in the East in Tutuala, van Engelenhoven and Huber (2020, pp. 351–355) similarly find complementary distribution, with [z] root-initially and [j] root-medially. Attempts at creating an orthography for Fataluku have also identified variation. Valentim's (2002) monolingual Fataluku dictionary uses an orthography that uses both [z] and [j] (which he writes as <i>), but he often cross-references between variants (e.g., both *paia* and *paza* 'necklace' are listed as headwords, with the definition given under *paia*). The Fataluku Language Project in 2004 and the Fataluku Lan-

guage Council in 2012 both proposed <j> for this phone, but allow for variation in its representation (see discussion in van Engelenhoven & Huber, 2020, p. 370).[3]

Table 1. The consonant phonemes of Fataluku; the status of (j) is debated

	Bilabial	Labiodental	Coronal	Velar	Glottal
Stops	p		t	k	ʔ
Affricates			t͡s		
Fricatives		f v	s (j)		h
Nasals	m		n		
Taps/Trills			r		
Laterals			l		

Despite these claims, most work characterizes (j) as monolithic within dialect region, with little suggestion of variation beyond the realizations [j] and [z]. By contrast, Heston (2019) finds a much wider range of articulations in Lospalos and Tutuala than previously reported, including voiced and (partially) devoiced alveolar and postalveolar fricatives and affricates. Although he finds little consistent patterning among the obstruent variants, he finds complementary distribution between obstruents and glides in Tutuala, with obstruents occurring initially and glides medially (see also van Engelenhoven & Huber, 2020, p. 355). Heston (2019) attributes the variation in medial position in speakers from Lospalos to speaker- and lexeme-level effects. While this variation is not modeled statistically, Heston suggests that speaker region and that of the speaker's parents may have motivated differences in (j) realizations, especially if the speaker was from Lospalos (Heston, 2019, pp. 86–87). Lospalos's heterogeneity, he argues, may thus be the result of contact between dialects with a phonological contrast in medial position and those without. However, he hedges this claim by suggesting that principled

3. This issue has implications for both documentation and education. A document from the 2004 FLP proposes that sounds should be represented as they are pronounced, and that this may vary by region (leading to different regional orthographies). In the original Tetun, it reads: "Regra 1: Hakerek ne'ebé rona. Regra ne'e arti katak makdalen iha rejiaun seluk hakerek tuir sira-nia sistéma rasik." (van Engelenhoven, n.d., p. 1) (Eng: Rule 1: Write how it sounds/what you hear. This rule means that writing in each region is according to their own system. [all translations by the third author]). A later document from federal early-grades educational programs in Fataluku states: "Hakerek letra j iha fatin ne'ebé ko'alia j, no hakerek letra i iha fatin ne'ebé ko'alia i." (Langford, 2014, p. 4) (Eng: Write the letter j in places where j is said, and write the letter i in places where i is said.); however, later literacy materials chose the <j> variant (Langford, personal communication, 2023).

variationist analysis with wider geographic coverage is needed to more adequately characterize these patterns.

3. Methods

3.1 Materials

The data for this study were collected by the second author over one month as part of a larger project on regional variation in Fataluku. A list of 99 words was created to illustrate all phonemic contrasts in phonetically-controlled environments, with a particular focus on phones reported to vary across dialects. The wordlist was limited to nouns to facilitate elicitation by picture, and was based on roughly 4,100 combined entries from Nácher's (2012) Fataluku-Portuguese dictionary and Heston's (2015) Fataluku-English wordlist (both of which argue for a phonemic contrast between /z/ and /j/).[4] Since the variable under study is infrequent in the Fataluku lexicon, it was possible to include the majority of lexical items reported to contain (j). The list (see Table A in Appendix) contained eight examples word-initially and eleven examples word-medially (three of which are reported to have a medial glide and eight a medial obstruent); these materials thus include (j) in all licit positions.

3.2 Participants

Fieldwork was conducted in Timor-Leste in 2018, during which 33 speakers were recruited by word of mouth. Because of the importance of region in previous reports, a primary focus was placed on collecting data from as many of the major regions as possible. The distribution of speakers is presented in Table 2, which includes macro-regional groups (e.g., North) following precedent in the literature (see Section 2.2). Older speakers had comparatively more difficulty completing the task and thus are underrepresented in the current sample. Importantly, a larger share of the data comes from young men. The lack of women is largely due to disproportionate household responsibilities assigned to young women, which made one-on-one recordings (especially those appropriate for phonetic analysis) an undue burden on their time and a potentially insensitive contravention of local notions of propriety. As a result, our ability to disentangle the effects of gender and other social predictors is limited.

4. In two cases, nominal phrases were elicited, as well as one example of an elicited reflexive pronoun.

Table 2. Breakdown of participants for the current study

		Region						
		Central/West		North		South/East		
Age	Gender	Lospalos	Cacavei	Lautem	Raça	Lore I	Tutuala	*Total*
Old	F			1	1	2	3	7
	M	1	2	2	1	1	2	9
Young	F	1	2					3
	M	5	1	1	4	1	2	14
Total		**7**	**5**	**4**	**6**	**4**	**7**	**33**

All participants were self-reported native speakers of Fataluku, except for one high-proficiency L2 speaker who had lived in Lospalos for decades; this speaker's productions of (j) did not differ substantially from comparable speakers in the sample. At least four speakers from each field site were included that represented the aforementioned consensus dialect regions: a Central/West grouping (including Lospalos and Cacavei), North (Lautem and Raça), South (Lore I), and East (Tutuala) regions. Hull's (2001) Northwest region is not represented, due to the difficulty associated with accessing this area during fieldwork.

3.3 Elicitation procedure

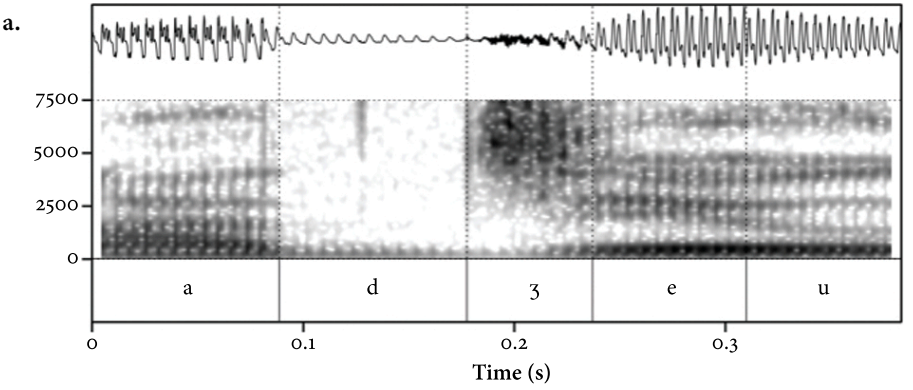
Each elicitation took place over two sessions, which helped prevent participant fatigue and accommodate time constraints (i.e., between school, work, and family responsibilities). Since there exists no widely used writing system for Fataluku (cf. van Engelenhoven & Huber, 2020, p. 370), target words were elicited by picture. Laminated index cards were created with a line drawing of the target word on one side and a representation of the word in a working orthography on the other, to help ensure that participants produced the intended word. Participants were asked to review the deck of pictures in pseudorandom order between one and three times, producing each target word in the frame /ana _ toto/ 'I am looking at _'. Each speaker also completed a sociolinguistic background questionnaire that included questions about family history, education level, and work status.

Elicitation procedures were conducted in Fataluku by the second author, aided by a young male speaker of Makasae, fluent in English, Tetun, and Fataluku. Recordings were made using a Zoom H6 digital recorder with either the built-in microphones or an external headset microphone (Shure SM35 or WH30), sampled at 44.1k/16bit. Most recording sessions were conducted outside on a covered patio (a culturally appropriate place to meet), either at a speaker's home or

place of work. This procedure yielded a total of 1,298 potential instances of (j); 349 tokens were excluded because they did not contain the target lexemes, yielding 949 tokens for analysis.[5] All primary data have been archived (Heston, 2018).

3.4 Data coding

The second author manually identified the on- and offset of each elicited token together with its frame in Praat (Boersma & Weenink, 2022). A Praat script was written to extract each token into its own short sound file, labeled with its gloss, speaker code, and position in the original file. All realizations of (j) were then auditorily coded by the second author, using additional information from the waveform and spectrogram. Tokens were coded for place (alveolar, postalveolar, palatal), manner of articulation (fricative, affricate, glide), and voicing (voiced, [partially] devoiced). Figure 3 illustrates the three most common variables: the voiced affricate [dʒ], voiced alveolar fricative [z], and voiced palatal glide [j].



5. Of note, all instances of the word /lojasu/ 'boat' were removed ($n = 35$), as it was invariably produced without the target segment. Realizations such as [leu̯.asi], [lo.asu], and [lo̯es] were common, potentially due to the token's complex etymological provenance, likely from some combination of Fataluku *loi* or Tetun *roo* 'boat' and Tetun *asu* or Portuguese *aço* 'iron', (cf. Williams-van Klinken & Williams, 2015; van Engelenhoven & Huber, 2020).

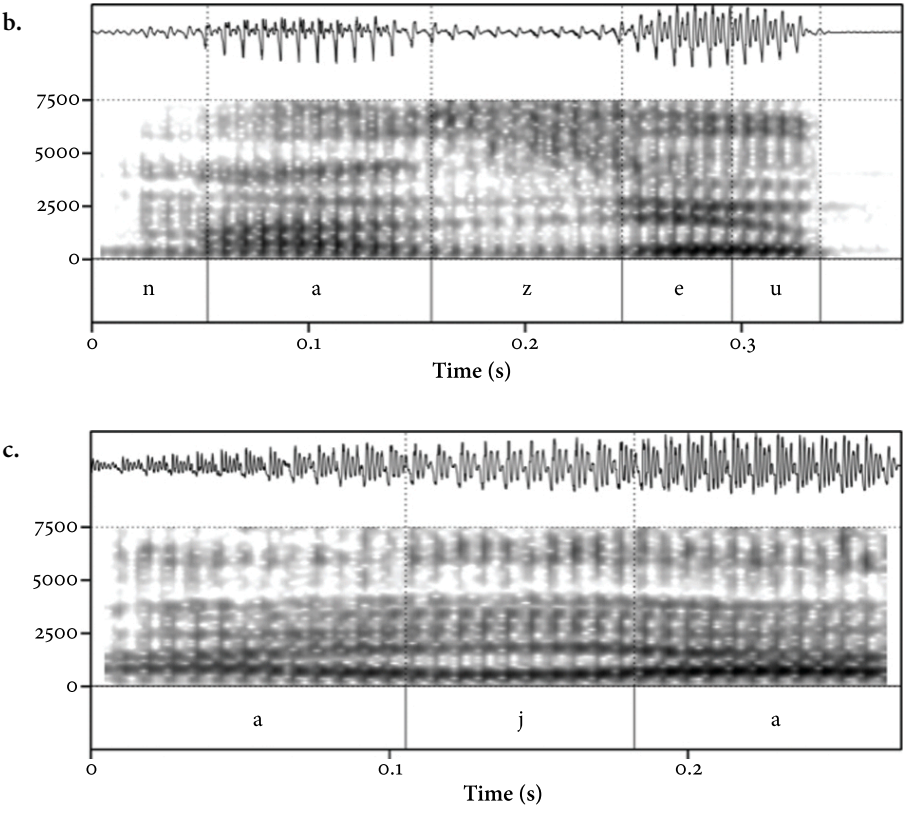


Figure 3. Examples from three speakers of (a) a voiced affricate in /jeu/ 'wife' (Lospalos); (b) a voiced alveolar in /jeu/ 'wife' (Lospalos); (c) a voiced glide in /aja/ 'rain' (Tutuala)

4. Analysis

Our analysis followed two guiding principles. First, given the difficulty of returning to fieldwork due to the COVID-19 global pandemic, we wanted to make maximal use of the available data. Thus, we set out to exclude as little data as possible, despite the clear imbalance of, for example, gender representation across region. Second, we prioritized a "bottom-up" approach to our analysis. This was due to both the lack of empirical data about variation in (j) and a generally poor understanding of what factors (beyond region) might impact the realization of the variable. In keeping with this "bottom-up" approach, we make use of the Boruta random forest classifier algorithm and classification trees to assess variation in the current data. Both techniques, described below, are appealing because they do not make *a priori* assumptions about the way data is distributed, nor do they assume how factors should influence the data (cf. Tagliamonte & Baayen, 2012; Dickson & Durantin, 2019). After delineating variation in the current sample using these methods, we turn to mixed-effects modeling to tease apart patterns in the data. We first discuss how the variants of (j) pattern in terms of raw frequencies, then move on to what factors we were able to operationalize in an analysis of (j).

4.1 Overall frequencies

The voiced coronal exhibits a wide range of variants in the current data, which fall into four broad groups: a glide realization, an elided (null) form, voiced obstruents, and finally, devoiced obstruents.[6] The raw numbers of each variant represented in the entire dataset are shown in Figure 4. The most frequent variant by far is the glide [j] ($n = 349$), accounting for nearly 37% of the data. Voiced obstruent realizations are the next most common, the greatest proportion of which are [dʒ] and [z] variants ($n = 186$ and $n = 162$, respectively); less common are [dz] ($n = 78$), [ʒ] ($n = 60$), and the null variant ($n = 71$), each accounting for under 10% of the data. The least frequent variants are devoiced obstruents: [d̥ʒ̥, d̥z̥, z̥], which together account for less than 5% ($n = 43$) of the data.

6. We characterize obstruents as 'devoiced' rather than 'voiceless' because they were typically produced with some degree of glottal perturbation, which either abated or increased through the duration of the segment. Of note, the plurality of such tokens ($n = 13$) come from a 55-year-old woman from Raça. Additional investigation did not suggest any further systematic patterning.

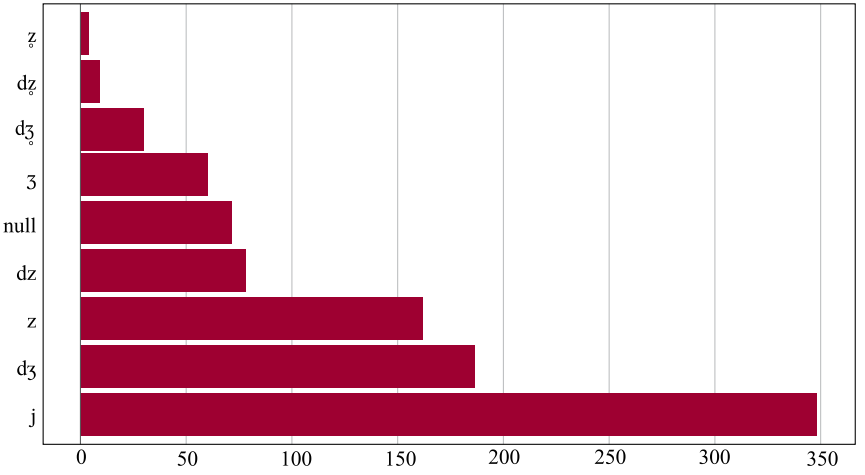


Figure 4. Count of (j) variants

4.2 Consideration of social and linguistic factors

Six factors were considered as potential predictors of variability in (j): region, age, gender, level of education, work type, and the variable's position in the word. Factors were selected for inclusion in the analysis based on how well-represented they were across the sample. Information about parents' region of origin (suggested as a potential source of variability in Heston, 2019) was not well-represented or evenly distributed enough across the participant pool to be included in the analysis. A primary goal of fieldwork was to identify geographic variation in Fataluku, and thus token numbers are relatively balanced across the six field sites, though there is less representation in Lautem and Lore I. Region was assigned based on where participants spent most of their childhood and adolescence.

Gender was treated as binary (men vs. women). Age was treated as a categorical predictor, split between old and young. Age in the sample is highly skewed by gender; women in the sample are considerably older (median = 49, range = 19–58) than men (median = 25, range = 17–48). In part, this difference stems from the aforementioned data collection challenges, where older women were easier to recruit, especially in more rural locations outside of Lospalos. A cut-off was established in the data at 35 years old, which placed roughly two-thirds of the women in the older category and two-thirds of the men in the younger category.

Level of education was categorized binarily based on the speaker's participation in secondary schooling: (1) individuals who had never attended school, or who completed up to elementary-level schooling, and (2) those who had attended or were attending any secondary school (including those who had either

completed high school and/or continued to university). Work was also categorized binarily based on whether an individual's daily tasks were subsistence-based (e.g., farming, ranching, fishing, domestic work), or commercial-based (e.g., construction, teaching). Many younger speakers reported under work that they were students. These speakers were included with commercial workers, as attending school in Timor-Leste is considered a path to a commercial career.

The sole linguistic effect considered was the position of (j) in the word; only two positions were possible: initial and medial. In both cases, (j) was always followed by a vowel (or, in the case of medial tokens, realized intervocalically), which was one of /a/ (69.3%), /e/ (18.7%), /i/ (11.3%) or /u/ (0.7%). The identity of the following vowel was not diverse enough to operationalize in analyses as a fixed effect, but investigation revealed vowel quality played no obvious role in the conditioning of (j) variant. Given the highly structured nature of data elicitation, no other linguistic effects can be reported here. Word position proves to be key in interpreting the results of our data, and in particular, on our ability to comment on the phonemic status of (j), as we discuss below. The factors, their levels, and token numbers are reported in Table 3.

Table 3. Factors included in the analyses

Factor	Categories (token count)
Region	Cacavei ($n=172$), Lautem ($n=95$), Lore I ($n=102$), Lospalos ($n=185$), Raça ($n=208$), Tutuala ($n=187$)
Age	young [17–34] ($n=444$), old [36–58] ($n=505$)
Gender	women ($n=247$), men ($n=702$)
Work	subsistence ($n=537$), commercial ($n=412$)
Education level	minimal to none ($n=328$), any secondary schooling ($n=621$)
Word position	initial ($n=424$), medial ($n=525$)

4.3 Analysis of factor importance

Despite agreement in the literature that region bears on the realization of (j), we did not want to make *a priori* assumptions about the importance of individual factors. Assuming the pre-eminence of, for example, region might have obfuscated the role played by other factors. As a result, we implemented the Boruta algorithm (Kursa & Rudniki, 2010) – a random forest wrapper – to assess the importance of individual features (i.e., predictors). Boruta has been used suc-

cessfully in variationist sociolinguistic research by Dickson and Durantin (2019), who utilize the algorithm to delineate variation in the Australian Kriol reflexive (see additional implications for use in Villarreal & Grama, 2023). Boruta works by duplicating the dataset and randomly shuffling features, creating so-called 'shadow' features. A random forest classifier is then trained on the data, and the importance of each feature is estimated by calculating the mean decrease accuracy of the model against a model with that feature removed. *Z*-scores of the real features are then compared to those of the shadow features; if the *z*-score of a real feature exceeds that of the shadow features, the feature is validated, and rejected otherwise. Boruta differs from other machine learning algorithms (such as random forests) in that it attempts to solve the so-called "all-relevant problem" (Kursa & Rudnicki, 2010, p. 2), which aims to identify *all* of the attributes relevant for classification (as compared with the "minimal-optimal problem", which seeks to identify the minimal set of features relevant for model performance; see also Nilsson, Peña, Björkegren, & Tegnér, 2007, pp. 601–602). Operationally, there are a number of benefits to Boruta; it can be fit to dependent variables with multiple levels, it is not hampered by covariation across predictor levels, and it is able to accommodate sparse data (Tagliamonte & Baayen, 2012, pp. 171–172; Dickson & Durantin, 2019, p. 198). Here, we run this algorithm using the Boruta package (Kursa & Rudnicki, 2010) in R (R Core Team, 2022); default settings were used, but alpha was lowered to 0.001, and all (j) variants were included as the dependent variable.

All factors included were judged to improve prediction accuracy and were thus retained. Figure 5 plots the importance of each factor, as estimated by the Boruta algorithm. The single most important predictor on (j) realization was position in the word, suggesting a strong effect of phonological conditioning. Region was assessed as the most important social factor, corroborating reports in the literature. All other factors ranked relatively lower in importance, with age and gender outranking level of education and work type. We address each predictor in turn in the subsequent section.

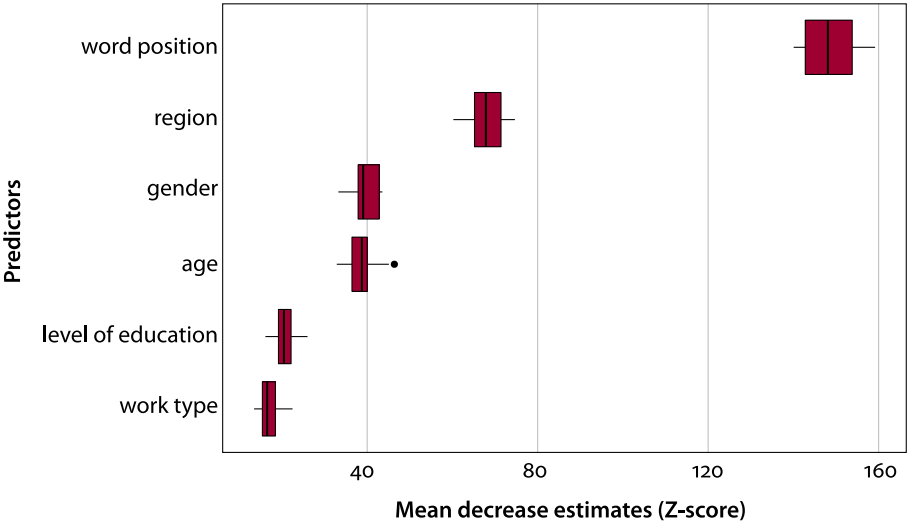


Figure 5. Importance estimates derived from runs of the Boruta algorithm

4.4 Investigating the network of effects

Word position was the strongest contributor to variation in (j). Figure 6 plots the frequency of each variant across position in the word (initial v. medial). Immediately evident is the split between variant identity and word position. In initial position, (j) exhibits considerable heterogeneity; the affricate [dʒ] dominates ($n = 147$), with the null variant ($n = 82$), [z] ($n = 71$), and [dz] ($n = 62$) together forming a secondary cadre; [j] and two voiceless obstruent variants [dʑ, ʑ] are minority variants. Closer inspection of the distribution of these variants reveals two important lexical patterns. First, the null form is highly lexically constrained, occurring only in the lexemes *jia* 'leg', *i'a* 'road', and *iapata/jampata* 'path'. This latter item is also the only one in which the glide variant occurs initially ($n = 11$).[7] By contrast, medial position shows a strong preference for [j] ($n = 338$). While all variants (except for null) have at least some representation in medial position, they together are just over half as frequent as the glide ($n = 187$).

This distribution of forms suggests three things that guide our focus moving forward. First, both the null form and the initial glide are highly lexically specific; (j) in these words appears to fall outside the variable context, and these words are thus excluded from the forthcoming discussion. Second, devoiced realizations of (j) are clearly a minority feature. To the extent that they appear, devoiced obstruents are largely restricted to word-initial contexts. This may be the result

7. Participants identified *jampata* 'road' as a loan from Indonesian *jembatan* 'bridge'.

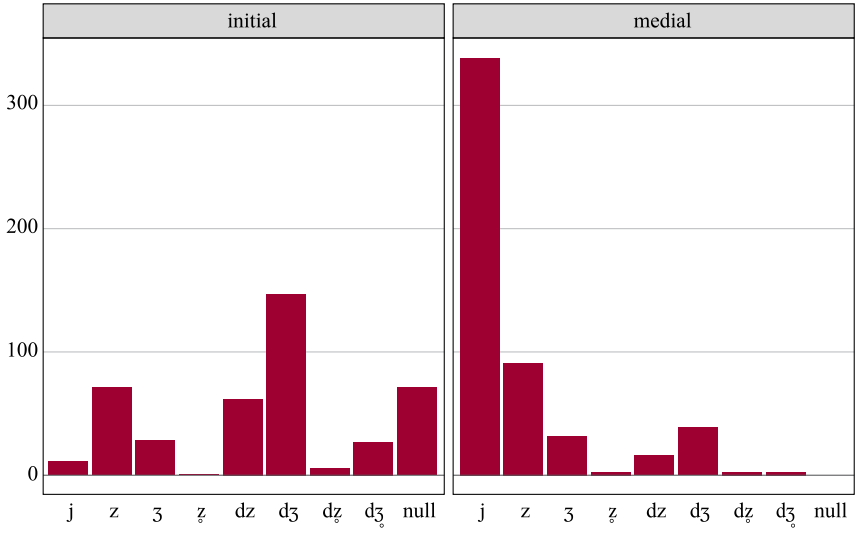


Figure 6. Frequency of (j) variants in initial and medial position in order of decreasing sonority

of domain-initial strengthening in Fataluku, in concert with the greater attention paid to speech involved in the elicitation task (see Cho & Keating, 2001). Third, and in line with observations in Heston (2019), there is a clear asymmetry across word position, with glide variants heavily favored in medial position and affricates preferred in initial position. Fricatives, by contrast, are more generally represented across word positions.

To help tease apart the distribution of these variants moving forward, we grouped all obstruent realizations (regardless of voicing or manner of articulation) together into a single category separate from the glide variant. We then fit a classification tree using ctree in partykit (Hothorn, Hornik, & Zeileis, 2006) to the dataset that included all factors in Table 3, given that none were rejected by Boruta. The significance level for variable selection was set at 0.01, with a maximum depth of 4, and the minimum number of observations at the leaf node at 50 to avoid overfitting. The tree is presented in Figure 7. Corroborating the observations from both the raw data and Boruta, the principal split is across word position; (j) variants in initial position are categorically obstruents and the glide is strongly preferred in medial position. Within medial position, a secondary split is identified for region; Tutuala patterns separately from all other regions in showing near-categorical preference for the glide variant (the lone exception being a token of [ʒ] in *aja* 'rain'). For regions outside of Tutuala, gender surfaces as an important predictor; women produce a higher proportion of obstruent realizations (62%) than men (45%). A further split across education indicates that men

with minimal education produce fewer obstruent variants (24%) than those with any secondary education (46%).

To account for variance across speaker and word, a logistic mixed-effects model was fit using lme4 (Bates, Mächler, & Bolker, 2015) to a subset of the data that included medial (j) produced by speakers outside of Tutuala; speakers from Tutuala were not included as they showed categorical use of the glide. The dependent variable was the manner of articulation of (j), broadly construed (obstruent v. glide), with region, gender, age, education, and work as fixed effects; speaker and word were random intercepts. Interactions were simplified if they did not improve model fit (by comparing AIC via ANOVA). Alpha was set at 0.05. To aid model fit, region was simplified to a three-way factor (Central/West [Lospalos-Cacavei], North [Lautem-Raça], and South [Lore I]) based on the macro-dialect regions identified in the literature (see Section 2.3). For region, we also used sum contrast coding, which compares each level of a factor to the overall group mean. Post-hoc comparisons were conducted with emmeans (Lenth, 2022).

The model identifies some variability within the non-Tutuala locales. Specifically, speakers from Lore I show significantly lower obstruent rates in medial position compared with the overall mean. Post-hoc comparisons further reveal that obstruent rates on the whole in Lore I (i.e., the South) are significantly lower than all other grouped locales; no difference emerges between the Central/West and North dialects. The model also identifies a gender difference, as well as a significant effect of education, depicted in Figure 8. Women produce significantly higher obstruent rates than men, but this effect is restricted to those men with minimal to no education. Post-hoc comparisons with emmeans with false discovery rate adjustment demonstrate that among speakers with any secondary education, this gender difference does not surface ($p = 0.735$). To reduce the possibility of overfit, a more parsimonious model was fit that lacked non-significant factors (work and age); this model corroborates the effects from the larger model reported here for education ($\beta = 3.81$, $z = 2.28$), gender ($\beta = 7.38$, $z = 3.06$), the interaction between education and gender ($\beta = -8.28$, $z = -2.47$), and region, specifically, Lore I ($\beta = -2.07$, $z = -1.83$).

In medial position, then, a picture of how this variable is socially distributed in Fataluku is beginning to emerge. Tutuala represents the clearest evidence of a functionally categorical allophonic split in (j), with initial obstruent variants and a medial glide. Importantly, Tutuala is both relatively isolated and of particular cultural import, a point we return to in the discussion. It is worth noting that of all non-Tutuala locales, Lore I shows the greatest preference for medial glides, though not nearly to the same extent (cf. van Engelenhoven's (2009, pp. 334–335) attestation that Lore I and Tutuala pattern in similar ways with respect to (j) realization). Outside of Tutuala, men and women show differentiated patterns that

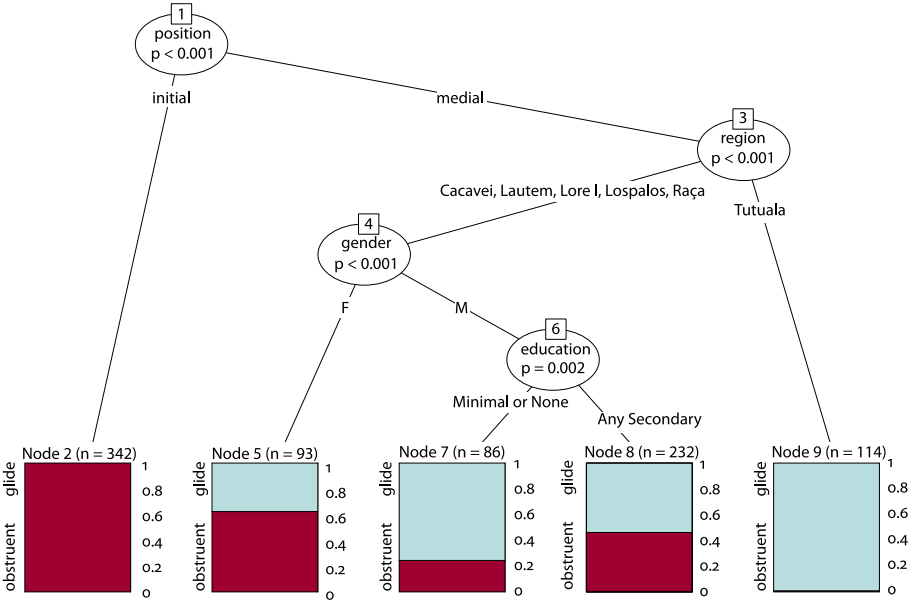


Figure 7. Classification tree fit to (j) realizations; bar plots at each terminal node reflect the proportion of obstruent (dark) and glide (light) variants

are mediated by education. Among speakers with limited formal education, men produce obstruents at far lower rates than women, a distinction that disappears among speakers with more formal schooling.

Of course, the variable (j) is complex in phonological form, and our inferences thus far have largely been drawn from its treatment as a binary variable. We achieve greater resolution into (j)'s social patterning when we account for different manners of articulation. This is done in Figure 9, which plots the proportions of affricate, fricative, and glide variants of (j) across region. Focusing first on initial position, we observe that most regions exhibit a higher proportion of affricates (hovering around 75%), with Raça showing the highest such proportions (around 87%). Lautem, by contrast, shows a much more even distribution between affricate and fricative realizations. In medial position, the effects of region are largely organized along the obstruent-glide dimensions already discussed above. Most regions show roughly equal proportions of the fricative and affricate variants; only Lore I and Lautem show a preference for fricatives, a relationship that is clearest in Lautem. We take this to mean that the findings from our classification tree above have adequately captured variation in medial position for (j).

To further delve into these patterns, we fit a logistic mixed-effects model to (j) in initial position. The dependent variable was affricate v. fricative realizations,

Table 4. Summary of logistic mixed-effects model predicting obstruent (vs. glide) (j) in medial position ($n = 411$; 45.3% obstruent); factors with significant effects highlighted; Marginal $R^2 = 0.242$; Conditional $R^2 = 0.832$; variance(speaker) = 5.16 ($SD = 2.27$); variance(word) = 6.40 ($SD = 2.53$); AIC = 322.2, BIC = 362.4, logLik = −151.1

Predictors	Log odds	SE	z	p	n	% obstruent
(Intercept)	−4.85	1.72	−2.83	< 0.01		
Region						
Central/West (Lospalos-Cacavei)	1.66	0.90	1.85	0.064	190	44.2%
North (Lautem-Raça)	1.04	0.80	1.31	0.191	163	49.1%
South (Lore I)	−2.71	1.10	2.55	< 0.05	58	37.9%
Gender (ref=Male)						
Female	7.57	2.26	3.35	< 0.001	93	62.4%
Age (ref=Old)						
Young	−2.23	1.49	−1.49	0.136	237	40.1%
Education (ref=Minimal to none)						
Any secondary	4.33	1.69	2.55	< 0.05	271	46.5%
Work (ref=Subsistence)						
Commercial	1.91	1.23	1.55	0.120	170	47.1%
Education:Gender						
Any secondary:Female	−8.16	3.12	−2.62	< 0.01	39	48.7%

with region, age, gender, education, and work as predictors; speaker and word were included as random intercepts. As above, region was expressed as a three-level factor using sum contrast coding (here with Tutuala grouped with Lore I under a South/East group, following van Engelenhoven, 2009, pp. 334–335). Alpha was set at 0.05. No interactions reached significance. Table 5 presents the model output.

Age is the lone factor to reach significance in the model, providing evidence for a change over time. Younger speakers are significantly more likely to produce affricate realizations than older speakers, a result that also arises from a more parsimonious model containing only age as a predictor ($\beta = 2.04$, $z = 2.57$). The three macro-regions show some variation in terms of changes in raw proportion over time; the North shows the greatest proportional increase in initial affricates (30.4%), followed by the South/East (20.9%), and Central/West (11.3%). However, these differences do not reach significance, suggesting the change is operative in the broader Fataluku community.

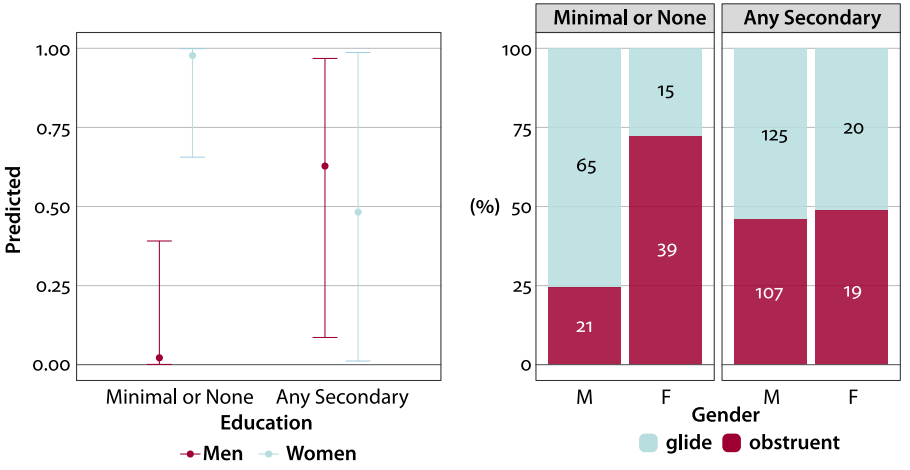


Figure 8. Model estimates (left) and raw proportions (right) of obstruent likelihood by education and gender

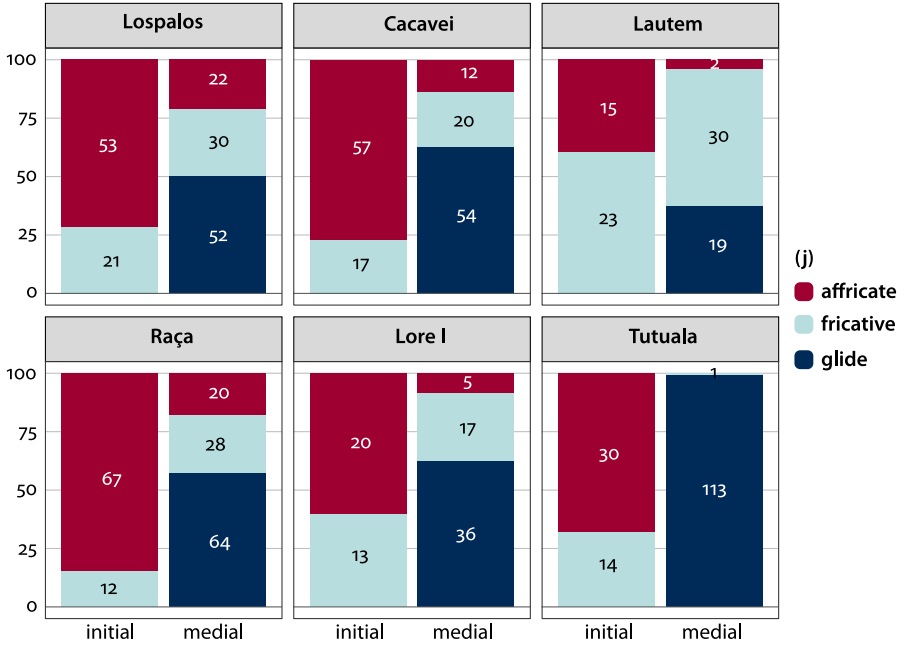


Figure 9. Proportion of affricate (red), fricative (light blue), and glide (dark blue) variants of (j) by word position and geographic region; token numbers in stacked bars

Table 5. Summary of logistic mixed-effects model predicting affricate (vs. fricative) (j) in initial position ($n = 342$; 70.8% affricate); factors with significant effects highlighted; Marginal $R^2 = 0.159$; Conditional $R^2 = 0.563$; variance(speaker) = 2.93 ($SD = 1.71$); variance(word) = 0.11 ($SD = 0.33$); AIC = 352.1, BIC = 386.6, logLik = −167.0

Predictors	Log odds	SE	*z*	*p*	n	% affricate
(Intercept)	0.64	0.76	0.84	0.402		
Region						
Central/West (Lospalos-Cacavei)	0.28	0.53	0.54	0.591	148	74.3%
North (Lautem-Raça)	−0.50	0.57	−0.87	0.383	117	70.1%
South/East (Lore I-Tutuala)	0.21	0.58	0.37	0.713	77	64.9%
Gender (ref=Male)						
Female	0.11	0.86	0.13	0.896	80	68.8%
Age (ref=Old)						
Young	2.64	0.97	2.71	< 0.01	197	79.7%
Education (ref=Minimal to none)						
Any secondary	−0.82	0.91	−0.90	0.366	231	71.4%
Work (ref=Subsistence)						
Commercial	−1.11	0.83	−1.34	0.181	146	69.9%

Also of note, work type did not reach significance in any analyses conducted on (j) in either phonological position, despite not being rejected by Boruta. We believe there are two reasons for this. First, and most importantly, work type is highly correlated with education in our sample. Of the 33 speakers in our sample, roughly half ($n = 15$) work in commercial sectors, and only three of these have limited formal education. Those whose daily tasks are subsistence based, by contrast, show a much more even split across education ($n = 10$ speakers with some secondary schooling; $n = 8$ with minimal schooling). Breaking this down by gender reveals even greater disparities; all but one of the five women with any secondary education were classified as having a commercial-based occupation. While these correlations among social factors are perhaps not surprising given that education is viewed as a pathway to a career in Timor-Leste, they make explicit testing of the effect of work type difficult. Second, because Boruta attempts to solve the aforementioned 'all-relevant problem', work type may have been boosted in importance as a result of its intersection with other social factors (despite being the lowest ranked feature). Despite this, there is some evidence that work type is relevant to (j)'s patterning among men (who show more balanced speaker numbers across work type and education). Among men engaged in com-

mercial work and who have some secondary education, medial (j) shows higher proportions of obstruent variants (56.0%) than among men engaged in commercial work with minimal formal education (11.4%). In fact, men with minimal formal education in commercial sectors more closely parallel the medial obstruent rates of men in subsistence work, irrespective of education (any secondary education = 27.8%; minimal education = 22.8%). However, neither the classification tree, nor the logistic mixed-effects models identified work type as a significant predictor, either alone or in interaction with gender or education type. We submit, then, the suggested split across work type and education must be subjected to further scrutiny and data collection.

5. Discussion and conclusion

We are now in a position to discuss the implications of the findings presented here. First, we consider the value of applying mixed methods to the documentation of under-documented language settings. Second, we contextualize these findings in the unique social setting of Timor-Leste. Finally, we discuss how our findings bear on questions of the phonemic status of the voiced coronal in Fataluku.

5.1 Mixing methodologies

In this article, we drew from three major methodologies available to variationist sociolinguists today: random forest classifiers, classification trees, and generalized mixed-effects regressions. The Boruta algorithm – a random forest wrapper – proved an effective way to statistically explore the influence of a range of factors in the current dataset. We had little idea of how factors (outside of region and word position) would play a role in the realization of (j), a complex variable with nine different attested forms and highly variable token counts. As Dickson and Durantin (2019, p. 198) argue, a chief benefit to Boruta is its ability to accommodate exactly these issues, and is "particularly valuable for smaller datasets such as those that are typically used in studies of smaller languages". We echo their sentiments here, underscoring that Boruta is particularly well-suited for an *exploration* of the importance played by multiple predictors on complex sociolinguistic variables. Even when all factors are selected as important, as was the case here, Boruta allows for the quantification of the relative importance of each predictor and their prioritization in subsequent analyses. Our use of Boruta then differs somewhat from Dickson and Durantin's, as we implement it as a pipeline to other analytical methods (i.e., classification trees and regression analysis) to character-

ize the interaction between different predictors. The classification tree allowed us to identify that obstruent variants were largely opposed with glide variants of (j) in medial position. This corroborated Boruta's assessment of variable importance; position in the word and region were the chief factors influencing (j) variants. All other factors (except for work type) were selected at various depths in the classification tree, laying out "in panoramic relief" the variable grammar hierarchy (Tagliamonte & Baayen, 2012, p. 172). Finally, mixed-effects regression helped to identify finer-grained variation in both medial and initial contexts. We therefore amplify the argument put forth by Dickson and Durantin (2019) that adding Boruta to the front end of the variationist analysis pipeline is a useful tool for teasing apart the role played by multiple (sometimes highly correlated) factors.

5.2 Contextualizing the interplay of social factors

With the exception of work type, all tested factors had some effect on (j). The strongest effect was that of word position; obstruent realizations are favored in initial position, with glides favored in medial position. Region was the most relevant external predictor of (j). In particular, Tutuala shows functional categoricity of the glide in medial position, and all other areas show greater variability in medial position.

In medial position, the regional differentiation we observe is crucially mediated by gender and education level. Among speakers with little formal education, women are more likely than men to produce obstruent realizations; however, this gender disparity disappears among speakers with more formal education. In initial position, by contrast, region is a poor predictor of variation in (j). Instead, a change in progress was identified, whereby affricate realizations are becoming more common among younger speakers, indicating that the Fataluku youth are converging on a shared preference for [dʒ] in initial position.

Proper contextualization of the broader implications of the above effects requires taking account of the socio-historical context of Timor-Leste – namely, the backdrop of violent societal upheaval of the Fataluku people. The months following Portugal's hasty 1975 withdrawal from Timor-Leste left a power vacuum, which neighboring Indonesia exploited by launching a military invasion. Indonesia immediately proscribed a new social order, which included the forced dislocation of rural Timorese communities to destabilize Timorese guerrilla resistance fighters (Stead, 2012). The effects of these edicts were most stringently felt in the easternmost regions of the country where opposition was the most coordinated. To avoid the initial waves of violence, ordinary people fled their villages to hide in the denser terrain of the mountainous interior. Those who were captured or eventually surrendered were placed into long-term observation areas and later

relocated into heavily guarded and surveilled settlements throughout the country (*ibid*). These resettlements severely disrupted agricultural production, causing widespread famine (Boviensiepen, 2014). During this period, the roles of languages were also strictly regulated. Tetun was tolerated as the colloquial language of the Timorese and, later, of the Catholic church, but Indonesian was imposed as the official language of the provincial administration, to the extent that children were punished for speaking their home languages at school (CAVR, 2006; Ross, 2017, p. 295). By 1999, most East Timorese had experienced some degree of displacement and, subsequently, contact with other Timorese languages and speakers (CAVR, 2006, p. 73).

Despite these assimilatory pressures, variation in Fataluku appears vibrant today, a finding that is in line with the importance placed on ancestral knowledge (see McWilliam, 2007). The strongest regional finding from this study is that in Tutuala there is striking consistency in the distribution of the glide variant in word-medial position. According to one Fataluku tradition, when their ancestors landed near Tutuala at the sacred cave *Ili Kere Kere*, their boats turned to stone, and the spirits inhabiting that area instructed them on where to live and how to settle. Even during Portuguese and Indonesian colonial regimes, Tutuala was promoted as a stronghold of tradition, feeding into "local ideas about ontological precedence and what they consider to be the ultimate source of sovereign authority" (Pannell, 2006, p. 210). One of the country's most influential figures, José Alexandre "Xanana" Gusmão – a former resistance fighter, political prisoner, and independent Timor-Leste's first president and two-time Prime Minister – is said to have personally requested assistance from the Lord of the Land spirit there (*mu'a ōcawa*) for the expulsion of foreign powers. Popular resistance mythology holds that Xanana was then imbued with the thoughts and power of the *tei* of Tutuala (a wild, ravenous spirit), which played a part in 'saving' Timor-Leste. Tutuala, then, is of iconic importance not only in Fataluku country, but in the entire nation.

That Tutuala serves as an ideological reference point for the speakers of Fataluku is also clear in remarks by participants, who identify the variety spoken there as noteworthy (Heston, 2019, p. 73). Indeed, our findings suggest that glides in medial position enjoy greater representation in both Tutuala and Lore I, two of the more difficult to reach, rural areas in the district of Lautém.[8] Obstruent realizations, by contrast, are highly represented elsewhere, including in more urban

8. It is important to note that relative to Lautem, Tutuala is not 'rural' in the same way that Lore I is. By characterizing Tutuala as 'rural' here, we underscore its connection to ancestral practice and suggest that while the village of Tutuala is not any less urban than Lautem, it represents the *ideal* of rurality.

areas like Lospalos and Lautem. Importantly, these realizations outside of Tutuala are further niched by gender and education; among those with limited formal education, men exhibit an increased likelihood to produce the glide variant, while those with any secondary schooling produce more obstruent realizations and show no differentiation by gender. More formal education, then, is linked with realizations that are less in line with patterns that typify Tutuala speakers. This profile suggests the possibility that obstruent realizations of medial (j) are a marker of (urban) mobility. Glide variants of (j) may index rurality, liminality, or tradition, an opposition that is strengthened, reinforced, and potentially even derived from its allophonic distribution in Tutuala. Obstruent realizations would then be connected to participation in more urban, outwardly facing endeavors, including (but perhaps not limited to) participation in formal education. Importantly, given the limitations of our samples, it is impossible to know whether the observed patterns for men are simply not shared by women or if women are not well-represented enough in the sample for such effects to surface. However, we argue that the weight of evidence suggests a strong connection between medial (j) and connection to place.

Initial realizations of (j), by contrast, appear to be unconnected to rurality, instead showing evidence for a change in progress towards affricate realizations. Older speakers conserve initial fricative forms, while younger Fataluku speakers prefer affricate variants. The precise reasons for this change are unclear, but it is reasonable to ascribe it, in part, to increased geographical mobility among young Fataluku people, who are more likely to have experienced formal schooling and spent long periods in Dili (cf. McWilliam, 2007, p. 360). Weak support for this can be found in Lospalos and the surrounding areas, which show lower proportions of initial fricatives compared to Lore I and Tutuala. Although none of the field sites emerged as significantly different from any other locale, it is reasonable to suspect that Lospalos's status as an urban hub in the broader district plays some role in the changing social profile of Fataluku speakers. As young Timorese increasingly seek employment opportunities outside of rural areas, Lospalos may serve as a proving ground for emerging local norms. Further research into the indexicalities of initial (j) variants will undoubtedly provide insight into this change.

5.3 The phonemic status of (j)

In the literature on Fataluku phonology, researchers are split on whether to treat these phones as separate phonemes or allophones. The current analysis corroborates reports that obstruent variants are more common initially and glide realizations are essentially restricted to medial position (van Engelenhoven, 2009; van Engelenhoven & Huber, 2020); however, we demonstrate that the dialect regions

considered here exhibit far greater heterogeneity in (j) than previously reported (see also Heston, 2019). Importantly, we demonstrate quantitatively that the regularity of positional allophony in Fataluku (j) is regionally mediated, with the clearest split in Tutuala; other regions show greater preference for medial obstruents. While these phones can thus be analyzed as a single voiced coronal, speakers' variable application of phonological processes derive their variants, which are mediated by region, gender, and education.

Funding

Open Access publication of this article was funded through a Transformative Agreement with University of Duisburg-Essen.

Acknowledgements

We gratefully acknowledge the comments from several anonymous reviewers that helped strengthen and streamline the arguments in the paper. Thank you to Mateus (Meti) who provided invaluable assistance in data collection, and Felix Maia for providing a translation of the abstract into Tetun. Fieldwork and project support was funded by a Franklin Research Grant from the American Philosophical Society and a grant from the research office of Payap University. Most of all, we thank the project participants for contributing their time and language to this study.

References

Bates, Douglas, Mächler, Martin, & Bolker, Ben (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software 67*(1), 1–48.

Boersma, Paul, & Weenink, David (2022). Praat: doing phonetics by computer [Computer program]. Version 6.3.03, retrieved 17 December 2022, from http://www.praat.org

Boon, Danielle, da Conceição Savio, Edegar, Kroon, Sjaak, & Kurvers, Jeanne (2021). Adult literacy classes in Timor-Leste and diverse language values and practices across the regions: implications for language policy-making. *Language Policy*, *20*, 99–123.

Bovensiepen, Judith (2014). Installing the insider "outside": House reconstruction and the transformation of binary ideologies in independent Timor-Leste. *American Ethnologist 41*(2), 290–304. http://www.jstor.org/stable/24027445.

Campagnolo, Henri (1973). La langue des Fataluku de Lórehe (Timor Portugais). Unpublished doctoral dissertation: Université René Descartes.

CAVR [Comissão de Acolhimento, Verdade e Renonciliação de Timor-Leste / Commission for Reception, Truth and Reconciliation in East Timor] (2006). *Chega! Final report of the Commission for Reception, Truth and Reconciliation in East Timor*. Dili: CAVR.

Cho, Taehong, & Keating, Patricia A. (2001). Articulatory and acoustic studies on domain-initial strengthening in Korean. *Journal of Phonetics*, *29*(2), 155–190.

da Conceição Savio, Edegar, Kurvers, Jeanne, van Engelenhoven, Aone, & Kroon, Sjaak (2012). Fataluku language and literacy uses and attitudes in Timor-Leste. In M. Leach, N. Canas Mendes, A. B. da Silva, B. Boughton, & A. da Costa Ximenes (Eds.), *Peskiza foun kona ba/ Novas investigações sobre / New research on/ Penelitian baru mengenai Timor-Leste* (pp. 355–361). Hawthorne: Swineburn Press.

Dickson, Greg & Durantin, Gautier (2019). Variation in the Reflexive in Australian Kriol. *Asia-Pacific Language Variation 5*(2), 171–207.

Fox, James J. (2003). Tracing the path, recounting the past: Historical perspectives on Timor. In J. J. Fox & D. B. Soares (Eds.), *Out of the ashes: Destruction and reconstruction of East Timor* (2nd ed., pp. 1–27). Canberra: ANU E Press.

Greksáková, Zuzana (2018). Tetun in Timor-Leste: The role of language contact in its development. Unpublished doctoral dissertation, Universidade de Coimbra.

Heston, Tyler M. (2019). Variation in the Voiced Coronals of Two Fataluku-Speaking Villages. *Journal of the Southeast Asian Linguistics Society 12*(2), 71–9. http://hdl.handle.net/10524/52456

Heston, Tyler M. (2015). The segmental and suprasegmental phonology of Fataluku. Unpublished doctoral dissertation, University of Hawai'i at Mānoa.

Hothorn, Torsten, Hornik, Kurt, & Zeileis, Achim (2006). Unbiased Recursive Partitioning: A Conditional Inference Framework. *Journal of Computational and Graphical Statistics 15*(3), 651–674.

Hull, Geoffrey (2001). O mapa linguístico de Timor Leste: Uma orientação dialectológica. *Estudos de Línguas e Culturas de Timor-Leste 4*, 1–19.

Hull, Geoffrey (2005). *Fataluku*. Dili: Instituto Nacional de Linguistica Universidade Nacional Timor Lorosa'e.

Kursa, Miron B., & Rudnicki, Witold R. (2010). Feature Selection with the Boruta Package. *Journal of Statistical Software 36*(11), 1–13.

Langford, Katrina (2014). *Local language council/EMBLI team – Fataluku: Summary report of workshop on Fataluku orthography and materials development*. Unpublished manuscript, Dili, Timor-Leste.

Lenth, Russell V. (2022). emmeans: Estimated marginal means, aka least-squares means. R package version 1.7.2. https://CRAN.R-project.org/package=emmeans

McWilliam, Andrew (2007). Austronesians in Linguistic Disguise: Fataluku cultural fusion in East Timor. *Journal of Southeast Asian Studies 38*(2), 335–375.

Meyerhoff, Miriam (2019). Unnatural bedfellows? The sociolinguistic analysis of variation and language documentation. *Journal of the Royal Society of New Zealand*, *49*(2), 229–241.

Nácher, Alfonso (2012). *Léxico Fataluco-Português*. Dili: Salesianos de Dom Bosco Timor-Leste.

Nilsson, Roland, Peña, José M., Björkegren, Johan, & Tegnér, Jesper (2007). Consistent feature selection for pattern recognition in polynomial time. *Journal of Machine Learning Research*, *8*, 589–612.

Pannell, Sandra (2006). Welcome to the Hotel Tutuala: Fataluku Accounts of Going Places in an Immobile World. *The Asia Pacific Journal of Anthropology 7*(3), 203–209.

R Core Team (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria.

RDTL [República Democrática de Timor-Leste] (2002). *Constituição da República Democrática de Timor-Leste*. Dili: República Democrática de Timor-Leste.

Ross, Melody A. (2017). Attitudes Toward Tetun Dili, a Language of East Timor. Unpublished doctoral dissertation, University of Hawai'i at Mānoa.

Schapper, Antoniette, Huber, Juliette, van Engelenhoven, Aone (2014). The relatedness of Timor-Kisar and Alor-Pantar languages: A preliminary demonstration. In Marian Klamer, (Ed.), *The Alor-Pantar languages: History and typology* (pp. 99–154). Berlin: Language Science Press.

Stead, Victoria (2012). Embedded in the land: Customary social relations and practices of resilience in an East Timorese community. *The Australian Journal of Anthropology 23*, 229–247.

Tagliamonte, Sali A. & Baayen, Herald R. (2012). Models, forests and trees of York English: Was/were variation as a case study for statistical practice. *Language Variation and Change 24*(2), 135–178.

Valentim, Justino (2002). *Fata-Lukunu i Disionariu / Dicionário Fataluku / Fataluku Dictionary*. Dili: Timor Loro Sa'e-Nippon Culture Center.

van Engelenhoven, Aone & Huber, Juliette (2020). Chapter 6: East Fataluku. In A. Schapper (Ed.) *The Papuan Languages of Timor, Alor and Pantar Volume 3* (pp. 347–426). Boston: De Gruyter Mouton.

van Engelenhoven, Aone (n.d.). Regra preliminaro ba ortografia [Preliminary orthographic rules]. Unpublished manuscript. Retrieved from https://webarchive.loc.gov/all/20100609005922/http://www.fataluku.com/orthography/ Accessed. 01 Nov 2022.

van Engelenhoven, Aone (2009). On derivational processes in Fataluku, a non-Austronesian language in East-Timor. In W. L. Wetzels (Ed.), *The Linguistics of Endangered Languages, Contributions to Morphology and Morpho-Syntax* (pp. 331–362). Utrecht: Netherlands Graduate School of Linguistics.

Villarreal, Dan, & Grama, James (2023). Modeling social meanings of phonetic variation amid variable co-occurrence: A machine-learning approach. In R. Skarnitzl & J. Volín (Eds.), *Proceedings of the 20th International Congress of Phonetic Science* (pp. 3745–3749). Guarant International.

Williams-van Klinken, Catherina, & Williams, Rob (2015). *Mapping the Mother Tongue in Timor-Leste: Who Spoke What Where in 2010?* Dili: Dili Institute of Technology.

Appendix

Table A. Target lexical items in broad phonetic transcription

Gloss	broad transcription (anticipated)	Gloss	broad transcription (anticipated)
'ice'	zelu	'cockatoo'	kaja
'Jaco island'	zako	'cousin'	vajan
'leg'	zia	'juice'	vaja
'self'	zen hin	'mango'	pajah
'wife'	zeu	'necklace'	paja
'plane'	zatu	'net'	kajalau
'road'	zampata	'bedroom'	tajan alivana
'plantain'	azan muʔu	'tears'	inavaja
'rain'	aza	'ship'	loojasu
'year'	azaʔira		

Abstract (Tetun)

Artigu ida ne'e hanesan esforsu investigasaun ba variasaun iha lian koronal (j) iha Fataluku – dalen Papua ida iha Timor-Leste. Iha artigu ida ne'e ami implementa algoritmu naran mak 'Boruta' iha inisiu husi faze análiza atu bele sukat importansia husi 'prediktor estatística' – katak ita siik kedas saida mak ema sira sei temi sai bazeia ba liafuan saida mak sira uza – hafoin ami uza 'estrutura klasifikasaun' – estrutura katak fahe sasaan tuir sira-nia kategoria – no regresaun efeitu mistura atu bele komprende didiak efeitu sira ne'ebé ami obzerva ona. Ami nia análize sujere katak pozisaun husi lian sira (iha liafuan) influensia tebes bainhira pronunsia (j), iha ne'ebé 'semivogais' – lian sira ne'ebé rona ba kabeer – akontese iha liafuan nia klaran, no 'frikativu' – lian sira ne'ebé rona ba groseiru – akontese iha liafuan nia oin. Rejiaun influensia tebes ba prediktor sira; hanesan koalia-nain sira iha Tutuala iha variasaun bainhira pronunsia [j], signifika katak iha 'alofonia' – pronunsia fonému ho maneira oi-oin. Iha Tutuala nia liur, 'elementu medial' – lian sira ne'ebé akontese iha liafuan nia klaran – diferente entre jéneru no nível edukasaun; koalia-nain sira ho nível edukasaun ne'ebé limitadu, mane sira uza semivogais barak liu fali feto sira; sira ne'ebé iha edukasaun sekundáriu, uza liafuan frikativu sira aas liu no ne'e hanesan entre feto ka mane. Elementu inisial, pelu kontrariu, komesa iha mudansa ba utilizasaun frikativu nian. Ami intepreta rezultadu peskiza iha kontestu ema Fataluku sira iha Timor-Leste, liu-liu iha area Tutuala nian.

Address for correspondence

James Grama
Sociolinguistics Lab
Institut für Anglophone Studien
Universität Duisburg-Essen
R12 S04 H24
45151, Essen
Germany
james.grama@uni-due.de

Co-author information

Tyler M. Heston
Department of Linguistics
University of Kansas
tyler.m.heston@gmail.com

Melody Ann Ross
Sociolinguistics Lab
Institut für Anglophone Studien
Universität Duisburg-Essen
melody.ross@uni-due.de

Publication history

Date received: 1 March 2023
Date accepted: 25 July 2023
Published online: 11 January 2024