Sonority Patterns in Lelepa Onset Clusters

Chao Sun, Rosey Billington

Australian National University

Chao.Sun@alumni.anu.edu.au, rosey.billington@anu.edu.au

Abstract

Lelepa, an Oceanic language of central Vanuatu, has complex syllable structures with onsets of up to three consonants and codas of up to two, which is rare among Oceanic languages. This study examines two-consonant onset clusters using natural speech data, showing a preference for certain clusters and frequent violations of the sonority sequencing principle. Clusters with smaller sonority distances are preferred, which is uncommon crosslinguistically. These findings enhance our understanding of Lelepa's phonotactics and contribute to the broader typology of phonotactic constraints.

Index Terms: syllable structure, consonant cluster, phonotactics, Oceanic

1. Introduction

1.1. Lelepa language

Lelepa[1] is part of the Southern Oceanic linkage, closely related to languages like Nguna, Nafsan, and Eton. It is spoken by around 500 people on Lelepa and Efate islands in central Vanuatu. Lelepa is notable for its complex syllable structure, which is uncommon among Oceanic languages [1]. Previous linguistic description highlights that among all the complex syllable structures observed in Lelepa, two-consonant clusters in the onset position exhibit the most complexity, including some combinations that are uncommon crosslinguistically [2]. In the context of increasing interest in phonotactic typology, further investigation into these consonant clusters can enhance our understanding of structures in Lelepa and syllable typology more broadly.

1.2. Consonant sequencing within syllables

Many languages allow for complex syllable structures such as CVCC and CCV, although CV is widely considered the most basic and prevalent syllable structure [3, 4]. Generally, the complexity of syllables is conceptualised in relation to the overall structures they exhibit, i.e. the more consonants allowed in the onset or coda position of the syllable, the more complex the syllable is; complexity also relates to the specific segments that can occur within the onset, nucleus and coda, and in what combinations [5]. Thus, languages typically exhibit varying restrictions regarding the types of consonants occurring at syllable margins and their relative positions within a complex syllable. One of the most well-known restrictions is the Sonority Sequencing Principle (SSP).

Sonority is a concept widely used in phonological theory to explain the distribution of segments within the syllable (e.g. [6, 7]). It is generally understood to refer to the loudness or perceptual prominence of speech sounds, which correlates with the degree of constriction in the vocal tract and thus is closely associated with the manner of articulation [6]. Numerous studies have explored sonority and attempted to rank consonant categories based on their sonority levels. One commonly cited five-scale version [7] ranks segments as in (1):

$$Glides\ (5) > Liquids\ (4) > Nasals\ (3) > Fricatives\ (2) > Stops\ (1) \quad (1)$$

The SSP is a phonological principle that was first proposed in the late 19th century [8]. The SSP aims to account for reported crosslinguistic tendencies in the organisation of consonants within a cluster based on their sonority levels. A large-scale 1960s study of consonant clusters at the beginnings and ends of words in 104 languages highlights the role of sonority in predicting the frequency of allowed consonant clusters across different languages [9]. According to the SSP, an ideal well-formed sonority pattern exhibits a rising sonority from the left edge (the first consonant in the onset cluster) towards the nucleus, followed by a fall in sonority from the nucleus to the right edge (the last consonant in the coda cluster) [7, 9]. For example, the monosyllabic English word 'crisp' /kɹɪsp/ is considered to have the ideal sonority pattern according to the SSP. Within the onset position, which is the focus of the present study, a sonority rising slope, such as /plV/, is considered preferred, and a sonority falling slope, such as /lpV/, is not. A sonority plateau, such as /ptV/, is usually language-dependent and can be interpreted as SSP-followed or SSP-violated depending on the analysis (e.g. [6, 9]). Previous studies on sonority also note that in addition to a general preference for rising sonority, onset clusters with larger difference in sonority are preferred over clusters with smaller differences in sonority [7, 9]. For example, /plV/ should be preferred over /pnV/ according to the scale in (1).

However, while the sonority account has been influential in explaining typological tendencies in syllable structures, exceptions to the principle are not uncommon. Such instances can be found within languages like Russian, where the onset cluster /lb/ is relatively common, despite violating the SSP by exhibiting sonority reversal [10]. Another case of violation of the sonority sequencing principle is observed in the Georgian language, where sonority reversal in the onset position occurs in clusters such as /rk/ and /md/ [11]. Additionally, in English, the prevalence of onset clusters starting with the alveolar fricative /s/, which goes against the expected sonority pattern, cannot be accounted for by the sonority sequencing principle. Examples such as /st/ in the word 'stress' and /sp/ in the word 'spell' are quite common [6]. A recent large-scale study of lexical data for 496 languages across 58 language families finds that over half of the languages in the database permit clusters violating the SSP either in word-initial or word-final positions [12], indicating that violations of the SSP are not rare crosslinguistically.

These observations suggest that while sonority provides a useful framework, other phonetic, phonological, and language-

[1] ISO 639-3: lpa, Glottocode: lele1267

specific factors also influence cluster patterns in different languages. They also highlight another issue: current research on the SSP and phonotactic typology has predominantly focused on major, well-studied languages and relied primarily on text-based data such as lexicons or written corpora, often without syllabification, limiting analyses to word-initial and word-final clusters. Therefore, there is a need for closer investigation into a more diverse range of languages, including understudied and minority languages, as well as different data types such as spontaneous speech corpora. This approach can lead to a more comprehensive understanding of consonant cluster patterns.

1.3. Syllable structure in Lelepa

Table 1. *Lelepa consonant inventory (based on [2])*

	Bilabial	Labiodental	Alveolar	Palatal	Velar	Labial-velar
Plosive	p		t		k	k͡p^{w}
Nasal	m		n		ŋ	ŋ͡m^{w}
Fricative		f	s			
Trill			r			
Lateral			l			
Glide				j		w

The phoneme inventory of Lelepa includes 14 consonants as in Table 1, with no voicing distinctions among obstruents, which are all phonemically voiceless. In the grammatical description of Lelepa [2], a schema for the syllable structure is introduced, showing the maximum number of consonants allowed in the onset position as three, and two in the coda position. Ten syllable structures are identified in Lelepa, from simple CV to rare forms like CCVCC and CCCVC (see more detail in Table 4, Section 3.4). On this basis, in a recent crosslinguistic study of syllable structures, Lelepa has been classified as exhibiting 'complex' syllable structure [1]. Some uncommon consonant clusters are observed in the language, such as /ntV/, /rkV/ and /fsrV/, which do not follow the ideal form of the SSP [2].

Previous study observes that the presence of complex clusters in Lelepa relates to ongoing vowel deletion processes within the language [2], compared to more conservative related languages and the ancestral language, Proto-Oceanic. Some neighboring languages, such as Nafsan, have also undergone similar processes, seemingly somewhat earlier than in Lelepa [13, 14]. Table 2 shows two examples of the vowel deletion observed in Lelepa and Nafsan compared with Nguna, which is more conservative. Both the /pr/ and /mt/ onset clusters in the words /prau/ 'long' and /mtak/ 'afraid' arise due to the deletion of the word-medial vowel /a/ in the words /parau/ and /mataku/. While the cluster /pr/ is considered ideal according to the SSP, the cluster /mt/ is ill-formed. The presence of consonant clusters like /mt/, which may be dispreferred according to the SSP, has made the syllable structure of the Lelepa language more complex.

Table 2. *Words realised with vowel deletion in Lelepa [2] and Nafsan [13] compared with Nguna [15]*

Lelepa	Nafsan	Nguna	Gloss
prau	pram	parau	'long'
mtak	mtak	mataku	'afraid'

2. Research questions

Given that the complexity of consonant sequencing in Lelepa primarily revolves around the two-consonant onset clusters, this study builds on previous descriptive research [2] to undertake a quantitative investigation of Lelepa onsets based on natural speech data. The following questions are addressed:
1. What is the frequency of different consonant combinations in two-consonant onset clusters?
2. To what extent does the Sonority Sequencing Principle account for the onset patterns in Lelepa?

3. Method

3.1. Corpus and data

This study uses archived spoken language data, collected during fieldwork (2007-2012) in the Lelepa-speaking villages Nataɲao and Mangaliliu and available as an open-access collection [2, 16]. The corpus includes approximately 100 items, totaling 13 hours of recordings. Each item includes audio or video recordings with time-aligned transcriptions in ELAN. Most of the annotation files from the archive include transcription, interlinear glossing, as well as a English freetranslation of the entire sentence.

For this study, a subset of the corpus comprising monologic narratives was selected. The subset was chosen based on genre, length, and recording quality. For genre, a variety of different types were selected to maximize the range of content covered, such as procedural descriptions, folktales, historical stories, and personal life stories. For length, samples of 2-5 minutes were chosen to allow for the analysis of a greater number of samples. Recording quality was assessed based on clarity, absence of ambient noise, and the voice quality of the speakers to ensure that the individual speech sounds could be segmented accurately. The dataset comprises 67 minutes and 48 seconds of recordings from 13 male speakers and 8 female speakers (for further details see [17]).

3.2. Data processing and segmentation

The ELAN [18] files for the chosen subset were first manually verified, then exported as Praat textgrids and subsequently separated into utterance-level files using a script in Praat [19]. Semi-automatic phone segmentation and labelling were performed using the online MAUS tools [20, 21], with manual correction of phone boundaries to ensure accuracy. To address the research questions in this study, additional annotations were then added to mark syllable structures.

3.3. Syllabification

In the case of Lelepa, where words can be polymorphemic and polysyllabic, consistent syllabification, particularly of word-medial consonant sequences like ...VCCCV... is crucial for analyzing Lelepa's syllable structure from actual speech data. The approach to syllabification used in this study follows the syllabification previously described for Lelepa word-medial consonant sequences [2]. According to this approach, in word-medial heterosyllabic consonant clusters, which can consist of up to three consonants, the first consonant is considered the coda of the preceding syllable, while the remaining consonants form the onset of the next syllable. This means that the sequence ...VCCCV... is syllabified as ...VC.CCV.... A summary of the approach is shown in (2), which demonstrates the hierarchy of

roles for word-medial consonants in polysyllabic words:

$$singleton\ onset > singleton\ coda > onset\ cluster > coda\ cluster \quad (2)$$

For example, consider the syllabification for the words /ar=msoun/ '3DU.R=want' and /e=salpnot/ '3SG.R=float.come'. In the case of the word /ar=msoun/, there is a word-medial three consonant sequence which is /rms/, and according to the rules mentioned above, the first consonant /r/ functions as a coda consonant following the nucleus vowel /a/, while the others /ms/ form the onset cluster for the next syllable /msoun/. It is the same with the second word; /pn/ forms the onset cluster for the syllable /pnot/. The syllabification of these is summarised in Table 3.

Table 3. *Example of syllabification*

ar=msoun	VCCCVV	→	[ar.msoun]	VC.CCVVC
e=salpnot	VCVCCCVC	→	[e.sal.pnot]	V.CVC.CCVC

3.4. Database and analysis

After the syllabification process using Praat, the Praat textgrid files were converted into a hierarchical database using the EMU Speech Database Management System, accessible through R with the emuR package [22]. This allowed for filtering and extracting syllable structure information, resulting in a dataset of 10,971 syllables. These occur in words of up to seven syllables, but there is a preference for words of three syllables or fewer. 42.8% of words are disyllabic, 25.3% are trisyllabic and 21.8% are monosyllabic. Table 4 shows the frequency of syllable structures found in the dataset by token and type, where 'tokens' indicates the total occurrences of each combination, and 'types' represents the number of unique words containing these clusters. The frequency of different syllable structures is almost identical for both tokens and types. The most frequent syllable structure is CV, with 6092 tokens. The next most frequent structures, CVC, V, and VC, do not allow consonant clusters. Structures with two-consonant onset clusters, namely CCV, CCVC and CCVCC, collectively account for 596 tokens.

Table 4. *Frequency of syllable structures in the dataset*

Syllable structure	Tokens	Percentage (%)	Types	Percentage (%)
CV	6092	55.53	1245	43.27
CVC	2834	25.83	865	30.07
V	887	8.08	325	11.30
VC	526	4.79	208	7.23
CCV	423	3.86	133	4.62
CCVC	159	1.45	80	2.78
CVCC	25	0.23	15	0.52
CCVCC	14	0.13	2	0.07
CCCVC	6	0.05	3	0.10
CCCV	5	0.05	1	0.03

Each consonant in the two-consonant onset clusters was assigned a value according to the sonority scale in (1). The sonority slope and sonority distance can be calculated by subtracting the sonority value of the first consonant in the cluster from that of the second consonant. For example, for a cluster like /tl/, the sonority distance is $4 - 1 = +3$, which indicates a sonority rising slope. For a cluster like /sf/, the sonority distance is $2 - 2 = 0$, which indicates a sonority plateau, and for a cluster like /mt/, the sonority distance is $1 - 3 = -2$, which indicates a sonority falling slope.

4. Results

4.1. Sonority slope

Table 5. *Onset two-consonant clusters grouped by manner of articulation. The frequency of 'types' is the number to the left, and the frequency of 'tokens' is the number in brackets.*

		Consonant 2				
		PLO	FRI	NAS	LIQ	GLI
Consonant 1	PLO	11 (20)	7 (8)	3 (5)	26 (38)	3 (3)
	FRI	19 (106)	3 (12)	2 (2)	21 (67)	1 (1)
	NAS	25 (72)	14 (21)	27 (146)	27 (50)	3 (27)
	LIQ	9 (11)	-	5 (7)	-	-
	GLI	-	-	-	-	-

In the dataset, 50 unique consonant combinations were identified in the 596 syllable tokens with two-consonant onset clusters, occurring within 206 different words. These clusters are categorized by their manner of articulation and are presented in Table 5. In this table, 'Consonant 1' refers to the consonant in the syllable-initial position, while 'Consonant 2' refers to the consonant closer to the nucleus. Each combination shows the count in tokens and types as mentioned earlier: the number on the left represents the number of unique words containing these clusters (types), and the number in parentheses indicates the total occurrences (tokens). For example, the entry '7 (8)' in the first row signifies that the plosive-fricative onset cluster appears eight times across seven different words. Table 5 shows that a range of consonant combinations are possible, and sequences starting with a nasal followed by either another nasal or a liquid are more common, each observed 27 times, followed by plosive-liquid clusters (26 times) and nasal-plosive clusters (25 times).

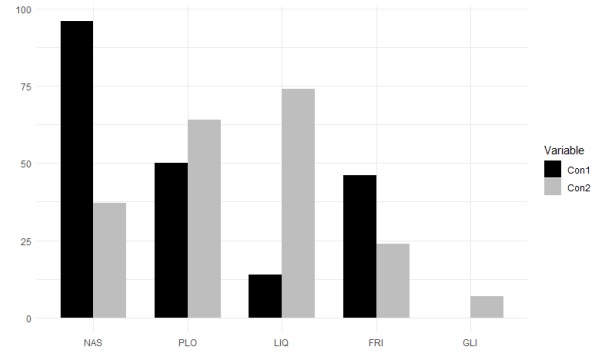

Figure 1: *Consonant distribution of onset two-consonant clusters (types)*

From Table 5, we can compare the frequency of each group of consonants in the Consonant 1 or Consonant 2 position in two-consonant onset clusters. Figure 1 shows this comparison more clearly. According to Figure 1, nasals are the most frequent consonants in the Consonant 1 position, occurring in 96 unique words, while liquids are the most frequent in the Consonant 2 position, occurring in 74 words. Glides do not occur in the Consonant 1 position.

Table 5 also indicates the change in sonority within the clusters. The gray regions represent a sonority plateau, where the sonority change is 0 as introduced in Section 3.4, such as in a nasal-nasal sequence. These gray regions also divide the table

into three parts: the regions to the upper right indicate a sonority rise, such as in a plosive-fricative sequence, and the regions to the lower left indicate a sonority fall, such as in a nasal-plosive sequence. This demonstrates that all three patterns of sonority slopes can be observed. Figure 2 further illustrates the frequency of different sonority slopes in two-consonant onset clusters. Most clusters exhibit a sonority rising slope (45.1%), and while 19.9% of clusters exhibit a plateau in the onset position, 35.0% of the clusters show a sonority falling slope, which violates the Sonority Sequencing Principle.

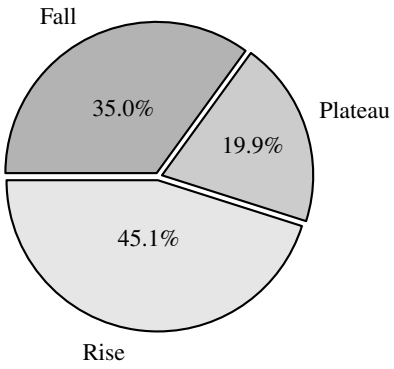

Figure 2: *Sonority slopes of onset two-consonant clusters (types)*

4.2. Sonority distance

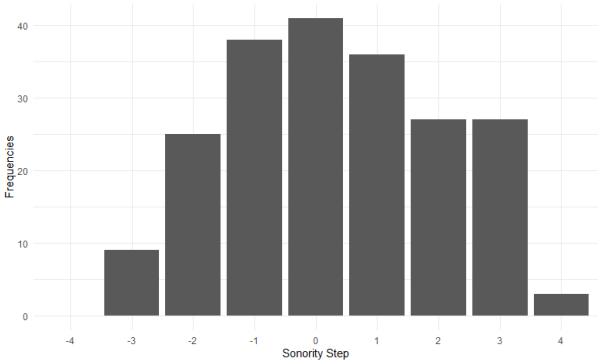

Figure 3: *Sonority distances of onset two-consonant clusters (types)*

Based on the five-scale sonority rank mentioned in (1), Figure 3 shows the frequency of different sonority distances for the two-consonant onset clusters. The sonority distance ranges from -3 (e.g., liquid-plosive) to +4 (e.g., plosive-glide). The most common sonority distance is 0, such as in nasal-nasal clusters, occurring 41 times. This is followed by a sonority distance of -1, occurring 38 times, and +1, occurring 36 times. This shows that onset clusters with small sonority distances are preferred. Although sequences with a large sonority distance, such as +4, are found in the database, they are the least common, occurring only 3 times.

5. Discussion and conclusions

This study focuses on the sonority pattern of two-consonant onset clusters in Lelepa. The results, based on both syllable tokens and types, show that Lelepa allows for most consonants in both positions of the onset cluster, except that glides only occur in the Consonant 2 position. The frequency of clusters in Table 5 shows that the most common clusters are nasal-nasal clusters and nasal-liquid clusters, each accounting for 13.1%. The frequency of consonants in Figure 1 shows that the most common consonants in the Consonant 1 position of the cluster are the nasals, accounting for 46.6%, while the most common consonants in the Consonant 2 position are the liquids, accounting for 35.9%. The prevalence of initial nasals likely relates to the very common nominal marker /n(V)/ that occurs in Lelepa and other central Vanuatu languages (discussed further below) [23].

The data shows that, as expected, all three types of sonority slopes - rising, plateau, and falling - are allowed in the onset position in Lelepa. According to the Sonority Sequencing Principle (SSP) introduced in Section 1.2, clusters with a sonority rising slope are expected to be preferred in the onset position. However, the frequency of clusters in Figure 2 shows that less than half exhibit sonority rising slopes (45.1%). The rest exhibit either sonority plateaus (19.9%) or sonority falling slopes (35.0%), which would typically be considered a violation of the SSP. However, as mentioned in Section 1.2, there is increasing evidence that violations of the SSP are not rare crosslinguistically [12]. The present study lends support to these observations, and highlights the need for more intensive exploration of phonotactic typology with reference to factors beyond sonority.

The SSP also suggests that the frequency of different clusters can be related to the degree of sonority distance between the consonants within the cluster, with a greater sonority distance being preferred in the onset position. However, according to Figure 3, the most common clusters are those with 0 sonority distance, namely the sonority plateau clusters. This result does not align with the sonority distance preferences of the SSP. As mentioned earlier in Section 1.3, vowel deletion in Lelepa mainly drives the formation of consonant clusters. One possible explanation for the lack of large sonority distances in onset clusters is that the (relatively recent) deletion process allows a high degree of flexibility in which consonants can be adjacent to one another. Additional clues can be found in other languages of central Vanuatu, such as Nafsan [13] and Neve'ei [24], which also permit consonant clusters violating the SSP and have also undergone medial vowel deletion processes. In all three languages, the clusters that violate the SSP often begin with the alveolar nasal /n/, which aligns with the frequency distribution of consonants in Figure 1, despite the crosslinguistic rarity of /n/ serving as the initial consonant in SSP-violating clusters [12].

The findings of this study confirm the complexity of Lelepa syllables, and suggest the need to consider language-specific factors and how they might interact with previously observed crosslinguistic tendencies, such as sonority sequencing. This study also highlights the need to continue expanding the range of languages analyzed in research on phonotactic typology, particularly focusing on segment sequencing in clusters. It would be beneficial for crosslinguistic research in this area to incorporate data from languages representing a broader range of language families and linguistic structures, with a greater focus on understudied languages.

6. Acknowledgements

We wish to acknowledge the Lelepa speakers who contributed to the corpus this study draws on, and Sébastien Lacrampe for providing access to the Lelepa materials via the ELAR archive.

7. References

[1] Easterday, S. *Highly Complex Syllable Structure: A Typological and Diachronic Study*. Berlin: Language Science Press, 2019.

[2] Lacrampe, S. "Lelepa: Topics in the grammar of a Vanuatu language," Ph.D. Thesis, The Australian National University, Canberra, 2014.

[3] Jakobson, R. *Selected Writings I: Phonological Studies*, The Hague: Mouton, 1962.

[4] Zec, D. "The syllable," in *The Cambridge Handbook of Phonology*, 2007, pp. 161–194.

[5] Maddieson, I. "Syllable structure," in *The World Atlas of Language Structures Online (v2020.3)*. Zenodo, 2013. doi: 10.5281/zenodo.7385533.

[6] Blevins, J. "The syllable in phonological theory," in *The Handbook of Phonological Theory*, 1st ed. Cambridge: Blackwell, 1995, pp. 206–244.

[7] Clements, G. N. "The role of the sonority cycle in core syllabification," in *Papers in Laboratory Phonology I: Between the Grammar and the Physics of Speech*, Cambridge: Cambridge University Press, 1990, pp. 283–333.

[8] Sievers, E. *Grundzüge der Phonetik: Zur Einführung in das Studium der Lautlehre der Indogermanischen Sprachen*. Leipzig: Breitkopf und Härtel, 1881.

[9] Greenberg, J. H. "Some generalizations concerning initial and final consonant sequences," Linguistics, vol. 3, no. 18, pp. 5–34, 1965, doi: 10.1515/ling.1965.3.18.5.

[10] Pouplier, M., Marin, S., Hoole, P., and Kochetov, A. "Speech rate effects in Russian onset clusters are modulated by frequency, but not auditory cue robustness," *Journal of Phonetics*, vol. 64, pp. 108–126, 2017. doi: 10.1016/j.wocn.2017.01.006.

[11] Crouch, C., Katsika, A., and Chitoran, I. "Sonority sequencing and its relationship to articulatory timing in Georgian," *Journal of the International Phonetic Association*, vol. 53, no. 3, pp. 1–24, 2023. doi: 10.1017/S0025100323000026.

[12] Yin, R., van de Weijer, J., and Round, E. R. "Frequent violation of the Sonority Sequencing Principle in hundreds of languages: How often and by which sequences?," *Linguistic Typology*, vol. 27, no. 2, pp. 381–403, 2023. doi: 10.1515/lingty-2022-0038.

[13] Billington, R., Thieberger, N., and Fletcher, J. "Nafsan," *Journal of the International Phonetic Association*, vol. 53, no. 2, pp. 511–531, 2023. doi: 10.1017/S0025100321000177.

[14] Billington, R., Thieberger, N., and Fletcher, J. "Phonetic evidence for phonotactic change in Nafsan (South Efate)," *Italian Journal of Linguistics*, vol. 32, no. 1, pp. 125–150, 2022. doi: 10.26346/1120-2726-151.

[15] Schmidt, H. *Nguna Dictionary*, Neuendettelsau: Erlanger Verlag für Mission und Ökumene, 2023.

[16] Lacrampe, S. "Possession in Lelepa, a language of Central Vanuatu," Master Thesis, University of the South Pacific, Suva, 2009.

[17] Sun, C. "An investigation of syllable structure in Lelepa," Master Thesis, The Australian National University, Canberra, 2023.

[18] The Language Archive, *ELAN*. (version 6.5) Nijmegen: Max Planck Institute for Psycholinguistics, 2023. Accessed: May 20, 2023. [Computer program]. Available: https://archive.mpi.nl/tla/elan.

[19] Boersma P. and Weenink, D., *Praat: Doing Phonetics by Computer*. (version 6.3.10), 2023. Accessed: May 20, 2023. [Computer program]. Available: http://www.praat.org/

[20] Reichel, U. D. "PermA and Balloon: Tools for string alignment and text processing," in *Proceedings of INTERSPEECH 2012*, ISCA, 2012, pp. 1874–1877.

[21] Kisler, T., Reichel, U. D., and Schiel, F. "Multilingual processing of speech via web services," *Computer Speech & Language*, vol. 45, pp. 326–347, 2017. doi: 10.1016/j.csl.2017.01.005.

[22] Winkelmann, R., Harrington, J., and Jänsch, K. "EMU-SDMS: Advanced speech database management and analysis in R," *Computer Speech & Language*, vol. 45, pp. 392–410, 2017. doi: 10.1016/j.csl.2017.01.002.

[23] Lynch, J. "Article accretion and article creation in Southern Oceanic," *Oceanic Linguistics*, pp. 224–246, 2001. doi: 10.1353/ol.2001.0019.

[24] Musgrave, J. *A Grammar of Neve'ei, Vanuatu*. Canberra: Pacific Linguistics, 2007.