Phonotactic frequencies in Marathi

KELLY HARPER BERKSON[1] and MAX NELSON
Indiana University Bloomington, Indiana

ABSTRACT: Breathy sonorants are cross-linguistically rare, occurring in just 1% of the languages indexed in the UCLA Phonological Segment Inventory Database (UPSID) and 0.2% of those in the PHOIBLE database (Moran, McCloy and Wright 2014). Prior work has shed some light on their acoustic properties, but little work has investigated the language-internal distribution of these sounds in a language where they do occur, such as Marathi (Indic, spoken mainly in Maharashtra, India). With this in mind, we present an overview of the phonotactic frequencies of consonants, vowels, and CV-bigrams in the Marathi portion of the EMILLE/CIIL corpus. Results of a descriptive analysis show that breathy sonorants are underrepresented, making up fewer than 1% of the consonants in the 2.2 million-word corpus, and that they are disfavored in back vowel contexts.

1 Introduction

Languages are rife with gradience. While some sounds, morphological patterns, and syntactic structures occur with great frequency in a language, others—though legal—occur rarely. The same is observed crosslinguistically: some elements and structures are found in language after language, while others are quite rare. For at least the last several decades, cognitive and psycholinguistic investigation of language processing has revealed that language users are sensitive not only to categorical phonological knowledge—such as, for example, the sounds and sound combinations that are legal and illegal in one's language—but also to this kind of gradient information about phonotactic probabilities. This has been demonstrated in a variety of experimental tasks (Ellis 2002; Frisch, Large, and Pisoni 2000; Storkel and Maekawa 2005; Vitevitch and Luce 2005), which indicate that gradience influences language production, comprehension, and processing phenomena (Ellis 2002; Vitevitch, Armbrüster, and Chu 2004) for infants as well as adults (Jusczyk and Luce 1994). Given the importance of gradience in various language processing phenomena, then, knowledge about phonotactic probabilities in a language is valuable. It strengthens psycholinguistic and behavioral research by contributing to well-designed production and perception experiments, and enables development of a more nuanced understanding of typological patterns. With this in mind, the present work presents a descriptive analysis of phonotactic frequencies in Marathi, an Indic language which contains typologically rare breathy voiced sonorants.

Typologically rare phenomena are often held up as examples of linguistic diversity—breathy sonorants may be uncommon, the argument goes, but since they *do* occur we must be able to account for them. This is true, but risks glossing over the potentially complex distributional patterns exhibited by these sounds. Frequency data of this type is often unavailable for under-studied languages, however. Marathi, the language reported on herein, is an Indic language spoken in the state of Maharashtra in India (Lewis, Simons, and Fennig 2015). Though it boasts >70 million speakers (2001 India Census), Marathi is

[1] Address correspondence to Kelly Berkson. Email: kberkson@indiana.edu

relatively under-resourced and data related to phonotactic probabilities are sparse.[2] The present research is part of an ongoing effort to address this gap. We report phonological frequency data for Marathi consonants, vowels, and CV bigrams in the EMILLE/CIIL written corpus of 2.2 million Marathi words (described below).

2 Background

To situate the Marathi frequency data presented herein, brief overviews of the consonant and vowel inventories of Marathi are provided in the following sections.

2.1 The Marathi consonant inventory

While scholarship on Marathi exists—notably, an unpublished 1958 dissertation on Marathi phonology and morphology by Ashok Kelkar, a 60-page *Outline of Marathi Phonetics* by Aparna Jha written in 1977, and two recent comprehensive descriptive grammars (Pandharipande 1997; Dhongde and Wali 2009)—commentary on phonetics and phonology remains slim. For example, the combined phonetics/phonology sections in the two most recent volumes are limited in length and detail (30 pages out of >250 in Dhongde and Wali, and about 40 pages out of >600 in Pandharipande). Additional commentary on Marathi phonetics and phonology—as well as reviews of its consonant and vowel inventory—appear in Franklin Southworth's 2000 review of the Pandharipande grammar and Kavadi and Southworth's 1965 Marathi language textbook. Pulling from these and other resources (Ghatage 2013; Masica 1993), we can make the following commentary about the inventory of consonant phonemes found in Marathi. These are illustrated in Fig. 1. Note that parentheses in Fig. 1 indicate what Dhongde and Wali (2009:35) call "marginal phonemes"—in particular the velar nasal, which arises almost exclusively due to place assimilation, the retroflex fricative (/ʂ/) which is retained only in Sanskrit-derived words (Dhongde and Wali 2009:16). Finally, /tsʰ/ occurs in Marathi very rarely, only word medially, and only in words of Sanskrit origin (Dhongde and Wali 2009:15; Jha 1977).

In keeping with Indic relatives such as Hindi and Bengali, Marathi contains a four-way contrast among obstruents composed of plain voiceless, aspirated voiceless, plain voiced, and breathy voiced categories (e.g. /t, tʰ, d, dʱ/). Phonemic breathy voice in obstruents is common in Indic languages, but extension of breathiness into sonorants is not. Marathi contains such contrasts among its nasals (/m, mʱ, n, nʱ, ɳ, ɳʱ/), approximants (/ʋ, ʋʱ/), laterals (/l, lʱ/), and rhotics (/r, rʱ/).

[2] Bhagwat (1961) analyzed Marathi frequency data from 105,000 words in a corpus composed of newspapers, fiction, and children's textbooks. This work, published by the Poona University and Deccan College Publications in Linguistics, constituted an important early contribution but was based on a small data set and is now dated.

Figure (1) Marathi consonant chart

Manner of Articulation	Place of Articulation																	
		Labial		Non-retroflex Apical				Retroflex		Alveo-Palatal		Velar		Glottal	
				Dental		Alveolar									
	Stop	p pʰ	b bʱ	t tʰ	d dʱ			ʈ ʈʰ	ɖ ɖʱ			k kʰ	g gʱ		
	Nasal		m mʱ		n nʱ				ɳ ɳʱ				(ŋ)		
	Affricate					ts (tsʰ)				tʃ tʃʰ	dʒ dʒʱ				
	Fricative					s		(ʂ)		ʃ					h
	Approximant		ʋ ʋʱ		l lʱ				ɭ		j				
	Rhotic										r rʱ				

The UCLA Phonological Segment Inventory Database (UPSID) indexes the inventories from 451 languages. Among them, five languages (1.11% of the total) utilize breathy phonation in vowels and 13 (2.9% of the total) in consonants. Only five of the 13 languages with breathy consonants contain breathy sonorants.[3] It is important to note that UPSID does not exhaustively catalog the Indic and Tibeto-Burman language families, but rather contains just a few examples from each. Phonemic breathiness is quite common in these families, however, particularly in obstruents. PHOIBLE (Moran, McCloy, and Wright 2014), another online database which catalogues phonemic inventories, contains data for 1672 languages. It contains many more Indic and Tibeto-Burman languages than UPSID, and the general trend found in UPSID holds: the labial consonants provide an example, with 13 of the 1672 languages indexed in PHOIBLE reported to contain /bʱ/ and 6 reported to contain /mʱ/. Given this typology, it is perhaps unsurprising that information about the frequency of occurrence of breathy voiced sonorants within a language is rare. This renders the opportunity to study such data in Marathi valuable.

Several comments about Marathi consonants should be made. Masica (1991) notes that the voiceless aspirated labial stop (/pʰ/) is often produced as a labial fricative in Standard Bengali (p. 103), and Dutta (2007) notes that /pʰ/ is undergoing a synchronic change to [f] in Hindi (p. 35). This appears to be happening in Marathi as well: the first author's observation is that while speakers sometimes produce [pʰ] in careful speech, [f] predominates in running speech. In acoustic data reported on in Berkson (2013), ten native speakers of Marathi (five female, five male) produced Marathi words embedded in a carrier sentence. Out of 120 productions of words that could have included [pʰ], fewer than 10% did. Male talkers produced [f] almost exclusively, while each of the five female talkers produced at least one [pʰ] (Berkson 2013:65).

[3] These include Isoko (Niger-Kordofanian), Newari (Sino-Tibetan), Parauk (Austro-Asiatic), and !Xu (Khoisan). Hindi-Urdu is listed in UPSID as a single language and is said to contain /lʱ/ and /mʱ/. It is included in the count of five presented here, to reflect the data presented in UPSID.

Agreement on the status of the labial approximants is lacking. They are categorized as /w/ by Dhongde and Wali (2009), as three distinct phonemes (/f, v, w/) by Kavadi and Southworth (1965) and Pandharipande (1997), as two phonemes (/v, w/) by Jha (1977), and as /ʋ/ by Masica (1993). We follow Masica (1993), characterizing /ʋ/ and /ʋʱ/ as approximants, thus grouping them with the sonorants.

The rhotic consonants included above are categorized as alveo-palatal, but there is debate in the literature as to how to characterize these sounds. Dhongde and Wali (2009) categorize /r/—correctly, in our opinion—as a "voiced alveo-palatal unaspirated flap or short trill" (p. 15). Ghatage (2013) categorizes it as an alveolar flap (p. 109). Finally, while the glottal fricative is included in Fig. 1 it is not categorized as being definitively voiceless or voiced. This sound is described as voiceless by some (Ghatage 2013; Pandharipande 1997) and voiced by others (Dhongde and Wali 2009; Masica 1991; Southworth 2000). Jha (1977) pointedly avoids commentary about whether or not the glottal fricative is voiced. We therefore leave it unspecified with regards to voicing.

2.2 A brief overview of the Marathi vowel inventory

In contrast with the consonants, the Marathi vowel inventory is relatively straightforward. Fig. 2 is constructed with primary reference to Dhongde and Wali (2009:9), though their analysis of the phonemic monophthongs of Marathi is in keeping with those found elsewhere (as in, for example, Ghatage 2013 and Yardi 1998, whose analyses of Marathi monophthongs aligns with Dhongde and Wali's).

Figure (2) Marathi vowel chart

	Front	**Central**	**Back**
High	**i**		**u**
Mid	**e**		**o**
Low-mid		**ə**	
Low		**a**	

In contrast to Pandharipande (1997), and in concurrence with Dhongde and Wali, we identify the distinction between /ə/ and /a/ as being one of quality, not length. Dhongde and Wali (2009:10) further note that Marathi does not contain nasal vowel phonemes, nor does it retain phonemic length distinctions inherited from Sanskrit though long versions of /i/ and /u/ remain as positional variants in final syllables (2009:9). The Marathi orthography nevertheless retains length distinctions even in places where such distinctions are neutralized in spoken language—in other words, spelling has not been adjusted to account for the loss of the length distinction. We provide counts for the long and short versions of these vowels because they are distinguished orthographically, but bear in mind that there is no phonemic distinction between these sounds.

The two monophthongs contained only in borrowed words (/æ/ and /ɜ/) are not addressed in the present study. Nor are diphthongs. Ghatage (2013) notes that there is still some disagreement with regards to the diphthongs included in the Marathi vowel inventory, mentioning that some scholars propose inclusion of /əu/ and /ai/ in addition to the two more commonly recognized diphthongs—/əi/ and /au/—and pointing out that for many Marathi speakers these diphthongs are perceived not as a single unit but as a sequence (2013:111-112). We thus leave the issue of diphthongs for future investigation.

3 Methods

The present study is an analysis of the Marathi portion of the EMILLE/CIIL (Enabling Minority Language Engineering/Central Institute of Indian Languages) corpus. The corpus is a collection of South Asian language corpora totaling 97 million words, the Marathi portion of which consists of 2,210,000 words from various written sources originally compiled by CIIL. In addition to Marathi data, EMILLE also contains language materials for Assamese, Bengali, Gujarati, Hindi, Kannada, Kashmiri, Malayalam, Oriya, Punjabi, Sinhala, Tamil, Telegu, and Urdu. The corpus is available for free for academic and research purposes. Additional details related to the corpus (and to fair and appropriate use thereof) are at http://catalog.elra.info/product_info.php?products_id=696.

Analysis of a spoken corpus will serve as a useful extension of the current work. No spoken corpus of comparable size is available at present, however, and Devanagari—the orthography used for written Marathi—is phonetically transparent. Each phone is represented with a unique character, and orthography aligns closely (though not perfectly) with pronunciation. As such, analysis of a written corpus provides an important first step towards identifying the general phonotactic patterns in the language.

Analysis of the corpus was conducted in order to provide a basic overview of the frequency of occurrence of the 38 Marathi consonants found in Fig. 1, five of the six vowels found in Fig. 2 (schwa is omitted for reasons described in Section 4.2), and the frequency of co-occurrence of each CV combination. Two types of frequency information are reported. **Token frequency** reflects the sheer number of times the consonant, vowel, or bigram of interest appears, and **type frequency** reflects the unique words that contain the targeted consonant, vowel, or bigram. It is good practice to report both types when presenting information about phonotactic probabilities (Fukazawa, Kitahara, and Sano 2015; Leung, Law, and Fung 2004; Tamaoka and Makioka 2004), for both are known to affect language processing (Bethin 2007; Bybee 2003:10-11; Ellis 2002) and are often explored in tandem in empirical psycholinguistic research (Denhovska 2015; Richstmeier, Gerken, and Ohala 2011).

4 Findings

The following sections present consonant, vowel, and bigram data. Full token and type frequency counts separated out by place of articulation are in Appendix A.

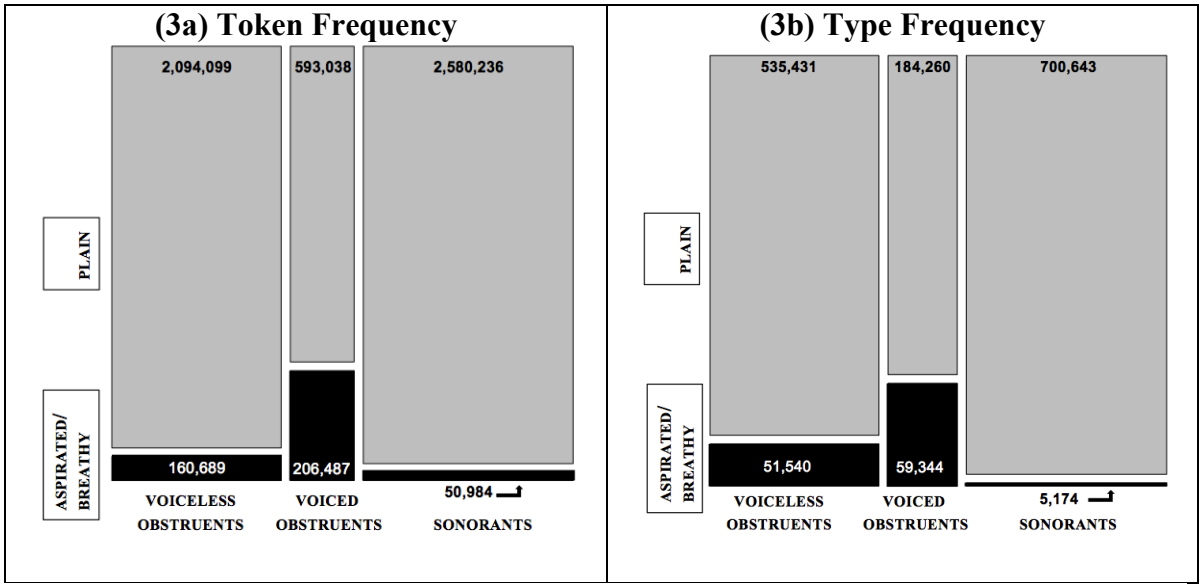
4.1 Consonants – general results

Recall that token frequency refers to the sheer number of times each type of consonant appears in the corpus. Type frequency counts, meanwhile, refer to the number of unique words in which each consonant type appears. As such, they are lower—a single word that contains one instance of a consonant and appears five times in the corpus adds five to that consonant's token count but only one to its type count. Raw data providing an overview of the token and type frequencies of consonants divided by obstruency, voicing, and phonation type are presented graphically in the mosaic plots in Fig. 3 and numerically in Table (1). Token frequency is in (3a) and (1a), while type frequency is in (3b) and (1b).[4]

The mosaic plots, presented first, make it clear that breathy sonorants represent a small proportion of the corpus regardless of how frequency is counted—0.9% of the corpus in terms of token frequency, and 0.3% in terms of type frequency.

[4] Neither the glottal fricative nor the marginal phones are included in the counts here.

Figure (3) Mosaic Plot of Total Consonant Frequency (voicing and obstruency by phonation type)



The data shown in Table (1a) reveals that the most basic observation we can make about token frequency is that 46% of the consonants in the corpus are sonorants, 40% are voiceless obstruents, and 14% are voiced obstruents. This is in keeping with findings from, for example, Romanian (Renwick 2011) and Setswana (Palai and O'Hanlon 2004). Renwick (2011:200) reports that in a Romanian wordlist containing a total of 788,157 characters, 44% of the consonants are sonorants, 40% are voiceless obstruents, and 16% are voiced obstruents. Analysis of conversational Setswana, meanwhile—a Bantu language—revealed that of all the consonants that occurred in the transcribed conversations, 46% were sonorants, 41% voiceless obstruents, and 13% voiced obstruents (Palai and O'Hanlon 2004:133). We also see that 93% of the consonants in the corpus are plain, while just under 7% are aspirated or breathy.

Of these non-modal sounds, 49% (206,487 out of 418,160) are breathy voiced obstruents, 38% are voiceless aspirated obstruents, and 12% are breathy sonorants. Overall, breathy sonorants represent fewer than 1% of the consonants in the corpus. About one out of four (26%) of the voiced obstruents in the corpus are breathy. The proportion of non-modal sounds is very different in the voiceless aspirated and sonorant categories, however: 7% of the voiceless obstruents are aspirated, and 2% of the sonorants are breathy.

Basic values differ only slightly in the type frequency data in (1b): whether counting by type or by token, about 93% of the consonants in the corpus are plain while about 7% are aspirated or breathy. 51% (up from 49% by token) of the non-modal sounds are breathy voiced obstruents, 44% voiceless aspirated obstruents (up from 38% by token), and 4% (down from 12% by token) are breathy sonorants. Non-modal sounds represent about one out of four, or 24%, of the voiced obstruents, about 9% of the voiceless obstruents, and 0.7% of the sonorants. Thus when comparing token with type frequency, the most notable change is in the breathy sonorant category. These sounds account for 12% of the non-modal sounds in the corpus token-wise but 5% type-wise. In terms of the total consonants in the corpus, breathy sonorants represent 0.9% token-wise and 0.3% type-wise.

Table (1) Token (1a) and type (1b) frequency of consonants in the EMILLE/CIIL Marathi corpus, divided by obstruency, voicing, and phonation type.

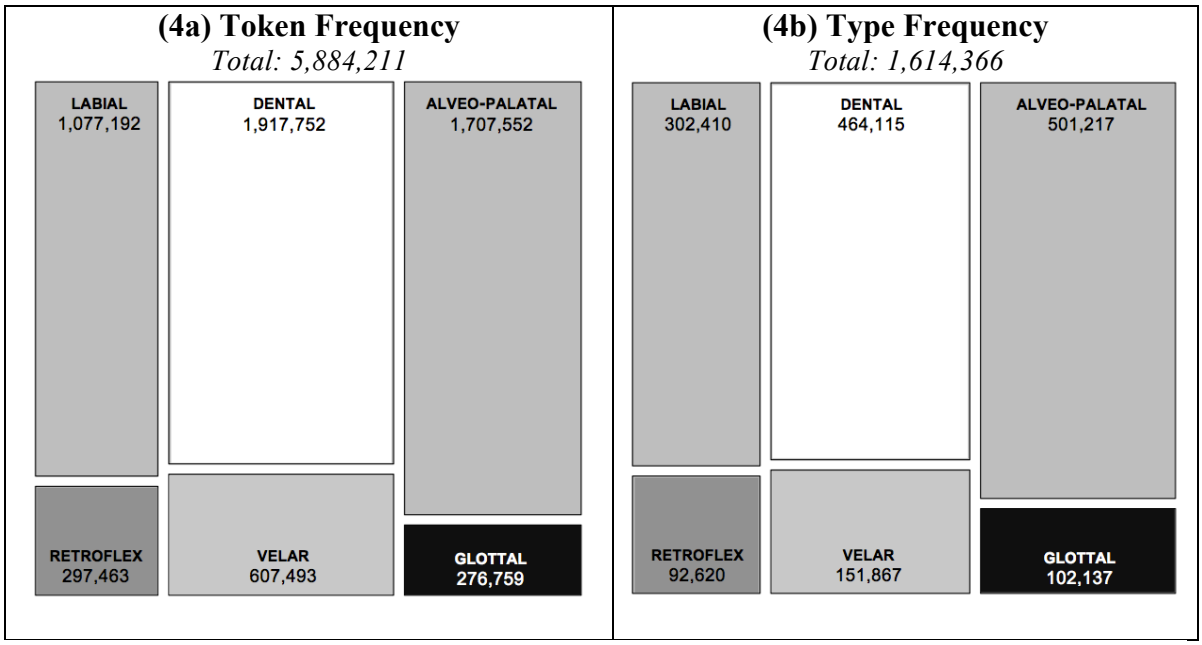
			PLAIN	ASPIRATED/ BREATHY	TOTAL
			count (% of total)	**count** (% total)	**count** (% total)
(1a) Token Frequency	**T**	voiceless obstruent	**2,094,099** (37%)	**160,689** (2.8%)	**2,254,788** (40%)
	D	voiced obstruent	**593,038** (10%)	**206,487** (3.6%)	**799,525** (14%)
	N	sonorant	**2,580,236** (45%)	**50,984** (0.9%)	**2,631,220** (46%)
		total	**5,267,373** (93%)	**418,160** (6.8%)	**5,685,533** (100%)
(1b) Type Frequency	**T**	voiceless obstruent	**535,431** (35%)	**51,540**(3.4%)	**586,971** (38%)
	D	voiced obstruent	**184,260** (12%)	**59,344** (3.9%)	**243,604** (16%)
	N	sonorant	**700,643** (46%)	**5,174** (0.3%)	**705,817** (46%)
		TOTAL	**1,420,334** (92.4%)	**116,058** (7.6%)	**1,536,392** (100%)

The initial picture, then, is one of underrepresentation: breathy sonorants are phonemic in Marathi but they constitute a small proportion of the consonants in the EMILLE/CIIL corpus. Their appearance is often due to a select set of lexical items that shows up repeatedly.

4.1.1 Consonant place of articulation

Grouping the data into broad obstruent and sonorant classes is informative, but there is always the risk that doing so may obscure small-scale patterns within the data. We could find, for instance, that breathy sonorants at the labial or dental place of articulation diverge from the overall pattern. To that end, we also assess the distribution of consonants divided across place of articulation (POA). Those data appear in Fig. 4. Token frequency is in (4a), and type frequency in (4b). Though the two do not differ notably, both are included for consistency.

Figure (4) Mosaic Plot of Total Consonant Frequency (by Place of Articulation)



Overall, there are fewer labials (both by token and by type) than alveo-palatals and dentals. All three outnumber the velars, retroflexes, and the glottal [h].

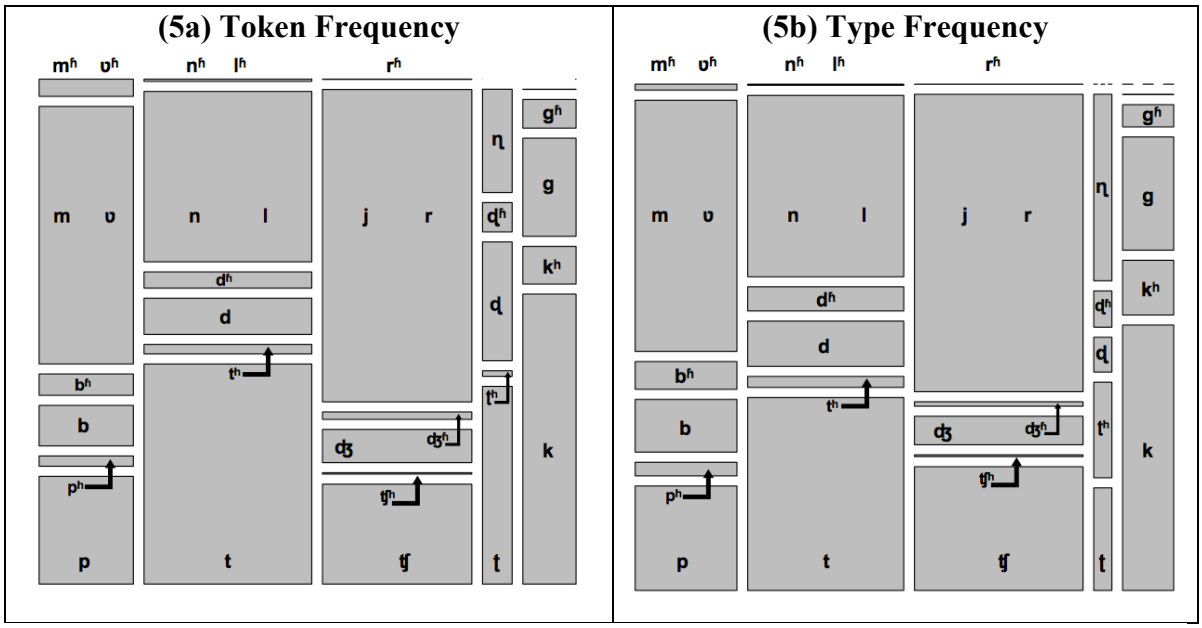
The patterns within each POA with regards to obstruency and phonation type which are of key interest are presented in Fig. 5. As above, token frequency data are in (5a) and type frequency data in (5b). The labeling here is dense, but the trends align completely with what would be expected given the total consonant data reported in section 4.1. Breathy sonorants make up a small proportion of the total corpus, and a small proportion of the sounds at each POA where they occur. Modal sounds are more numerous whether assessing type or token frequency, and it is the plain voiceless obstruents and plain sonorants at the labial, dental, and alveo-palatal POAs that account for the bulk of the consonants in the corpus. Note that exact token and type counts appear in Appendix A, with labials presented first (in Table A1), dentals next (Table A2), and so forth. Data for [h], which is not included in Fig. (5), is included with the dorsals at the end of Appendix A (Table A5).

By token frequency, 60% of the labials are sonorants and 40% are obstruents; for dentals, 38% are sonorants and 62% are sonorants; 69% of the palatals are sonorants and 31% are obstruents; of the retroflexes, 26% are sonorants while 74% are obstruents. As noted, nearly 100% of the velars are obstruents—the velar nasal is not considered a phoneme, but rather is a positional variant of other nasals. It is represented with its own orthographic character, though, and that character appears in the corpus 227 times. The velars are all obstruents, then: 64% of them are plain voiceless [k], 22% plain voiced [g], 8% aspirated voiceless [k^{h}], and 6% breathy voiced [$g^{ɦ}$].

These percentages differ only slightly if calculated for type frequency instead of token frequency. By type, 56% of the labials are sonorants and 44% obstruents; 40% of the dentals are sonorants and 60% obstruents; 65% of the alveo-palatals are sonorants and 35% obstruents; 24% of the retroflexes are sonorants and 76% obstruents. And, 58% of the velars are plain voiceless [k], 25% plain voiced [g], 12% aspirated voiceless [k^{h}], and 5% breathy voiced [$g^{ɦ}$]. Thus when considering consonants only, rather than bigrams,

distribution of consonants in the corpus across obstruency, phonation, and places of articulation is relatively consistent whether assessing by type or by token frequency.

Figure (5) Consonant Frequency (voicing and obstruency by phonation type)



There are so few instances of the breathy sonorants that when assessing the full data set they are nearly impossible to see. To gain a better sense of their distribution, the token frequency and type frequency of sonorant pairs is presented in Table (2). Type-Token Ratio (TTR) is also presented in Table (2). TTR is standardly used to reflect the amount of lexical variety in a text. In this case, it provides information about how many novel lexical items each sound of interest appears in. If the token frequency of a hypothetical consonant in the corpus is four, for instance, but the type frequency is one, it means the consonant appears in only one word but that that word appears four times. The closer the TTR is to 1, then, the more words contribute to its token frequency. When the TTR is closer to zero, it means very few lexical items contain the sound in question.

With this in mind, what we can see in Table (2) is that breathy sonorant values—whether considering token frequency, type frequency, or TTR—are lower than plain sonorant values in almost all instances. The lone exceptions are the TTR values for [lʱ] and [ɳʱ]. These two phones boast extremely low token and type frequencies—for [lʱ], 981 and 383, and for [ɳʱ] 89 and 23. Thus while the TTRs for the other breathy sonorants hover close to zero, indicating very little lexical variety (or understood differently, indicating that very few lexical items account for many of the occurrences of each phone), [lʱ] and [ɳʱ] appear to pattern differently but must be interpreted bearing their overall low frequencies in mind.

Table (2) Token frequency, type frequency, and type/token ratios for plain/breathy sonorant pairs

	Token Frequency	Type Frequency	Type/Token Ratio
[m]	257,748	71,352	0.28
[mʱ]	28,952	1491	0.05
[ʋ]	349,220	94,578	0.27
[ʋʱ]	11,742	2457	0.2
[n]	361,848	95,567	0.26
[nʱ]	8315	767	0.09
[l]	355,628	88,885	0.25
[lʱ]	981	383	0.4
[r]	547,782	153,291	0.27
[rʱ]	849	36	0.04
[ɳ]	182,140	40,356	0.22
[ɳʱ]	89	23	0.26

What about the occurrence of breathy sonorants at each POA? Token-wise, breathy sonorants account for 3.8% of the labial consonants, 0.5% of the dental consonants, 0.1% of the alveo-palatal consonants, and 0.03% of the retroflexes. Type-wise, they account for 1.3% of the labials, 0.2% of the dentals, 0.01% of the alveo-palatals, and 0.02% of the retroflexes. These are small percentages all around, but why is it that there are proportionally more breathy sonorants at the labial POA? This is in large part because [mʱ] appears in two function words in Marathi ([amʱi], *'we'* and [tumʱi], *'y'all'*). These two lexical items, and their morphologically-related forms, represent many of the instances of [mʱ] found in the corpus. While this particular breathy sonorant has a relatively high token frequency, then—in comparison with the other breathy sonorants—it has a lower type frequency and a very low TTR. In fact, Table (2) reveals that both of the breathy labial sonorants occur more frequently in the corpus than any of the other breathy sonorants. This is a pattern that bears more investigation in future, for at present we do not have an explanation for the [ʋʱ] data that parallels that of the [mʱ] data.

One final note that we should make about the token frequency of consonants is that the retroflex sibilant, /ʂ/, occurs 57,461 times in the EMILLE Marathi corpus. The type frequency is 12,095. This sound, which is considered to be a "marginal phoneme" by Dhongde and Wali (2009), appears in Sanskrit-derived words. Though it is unlikely that the spelling will be changed to reflect this, Dhongde and Wali report that the retroflex fricative is in fact pronounced as an alveo-palatal [ʃ] by almost all speakers in almost all instances. As such, this is one place where we can predict that the results will differ greatly in a spoken corpus.

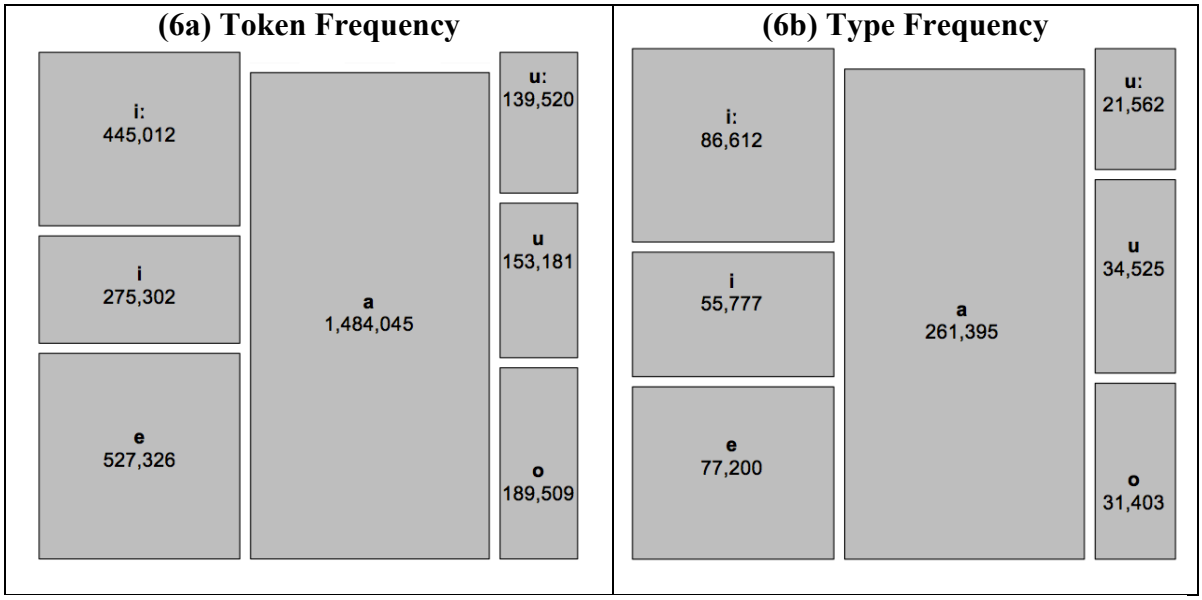
4.2 Vowel Frequencies

Frequency data for the monophthongs of Marathi is presented next. Note that data for /ə/ is excluded here, because /ə/ is not represented with a character in Devanagari. In other words, it is not spelled: rather, every consonant is assumed to have an inherent /ə/ unless some other vowel diacritic has been added on to it. (For example, the Devanagari symbol which represents the plain voiced velar stop is assumed to represent the syllable [gə] unless modified with a symbol representing another vowel.) This inherent /ə/ is

sometimes pronounced and sometimes omitted, particularly in word-final position, and there is no way to know which is the case in a written corpus. As such, an accurate /ə/ count can only be pulled from a spoken corpus, and so procurement of these data must wait until a suitable spoken corpus is available.

Data below are for the remaining monophthongs. Recall that length alternation in the high vowels is allophonic, as discussed previously, but that an orthographic distinction between [i] and [i:] (and between [u] and [u:]) remains. As such, counts for both the long and short versions of these vowels appears in the mosaic plots in Fig. (6).

Figure (6) Token and type frequency of monophthongs.



Several trends are readily visible in Fig. (6). First, the low central vowel [a] occurs far more frequently than any of the remaining vowels, accounting for 46% of the total both by token and by type. Combining the long and short allophones of the high vowels, /i/ accounts for 23% of the total by token and 25% by type while /u/ accounts for 14% of the total by token and 16% by type. The total for /e/ is 16% by token and 14% by type, while /o/ clocks in at a mere 6% by token and by type. This means that the back rounded vowels occur relatively infrequently, accounting for just 20% of the total vowels in the corpus by token and 22% by type. What is also true here is that holistically, the distribution of vowels in the corpus is relatively consistent regardless of whether they are assessed by token or by type frequency. /a/ predominates, and the back rounded vowels are somewhat underrepresented.

4.3 Type frequency of CV bigrams

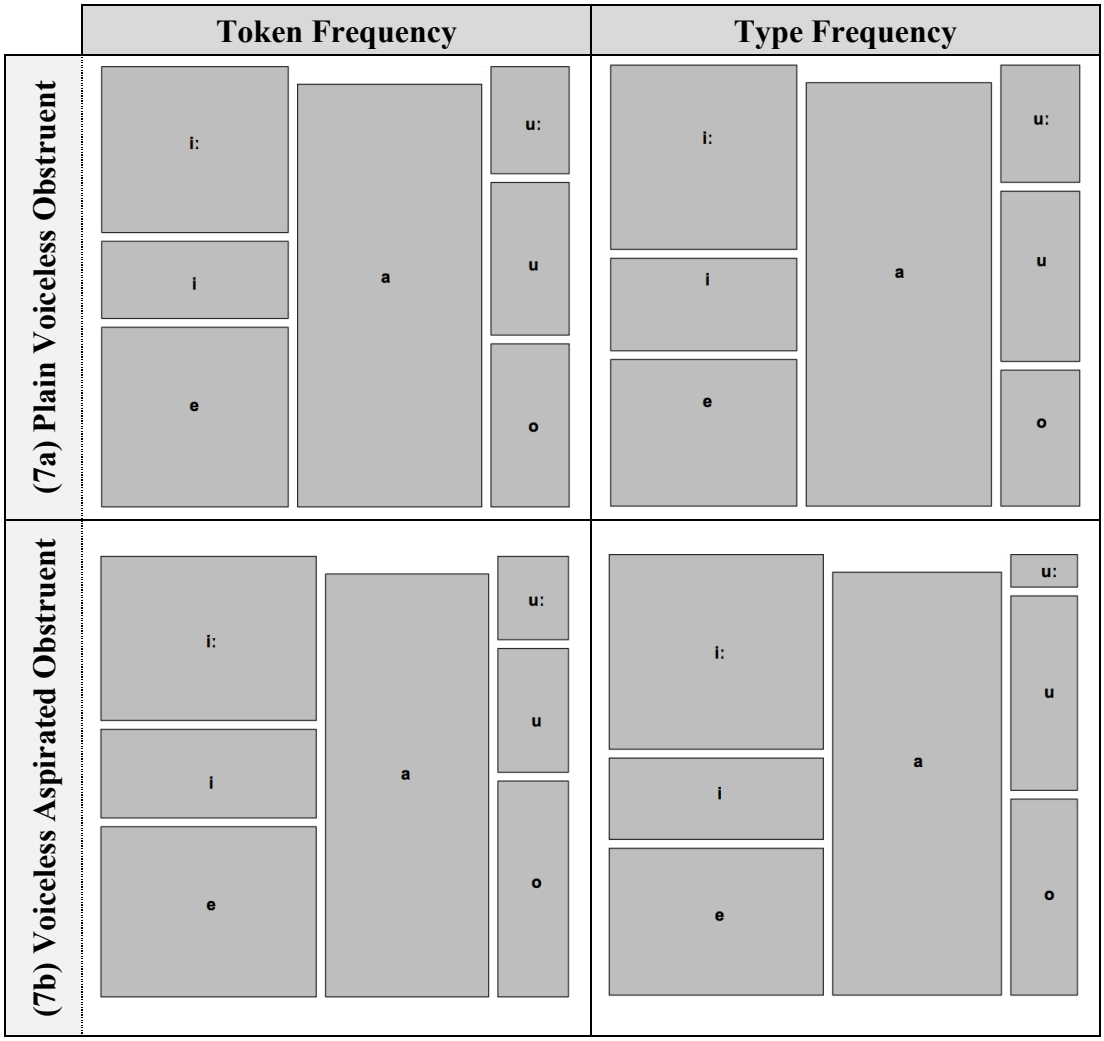
The distributional patterns of CV bigrams are also of relevance here, in part because some evidence suggests that the cues distinguishing plain from breathy consonants may be most robust in low-vowel ([a]) contexts and least robust in high-vowel ([i, u]) contexts. Indeed, acoustic studies of phonation type contrasts in consonants often focus on low-vowel contexts as a way to target the environments where spectral cues will be least affected by formant interference with spectral measures (Dutta 2007, Esposito 2006, Traill

and Jackson 1988). The acoustic benefit of low vowel contexts on phonation type contrasts may be particularly robust for sonorants (Berkson 2013).

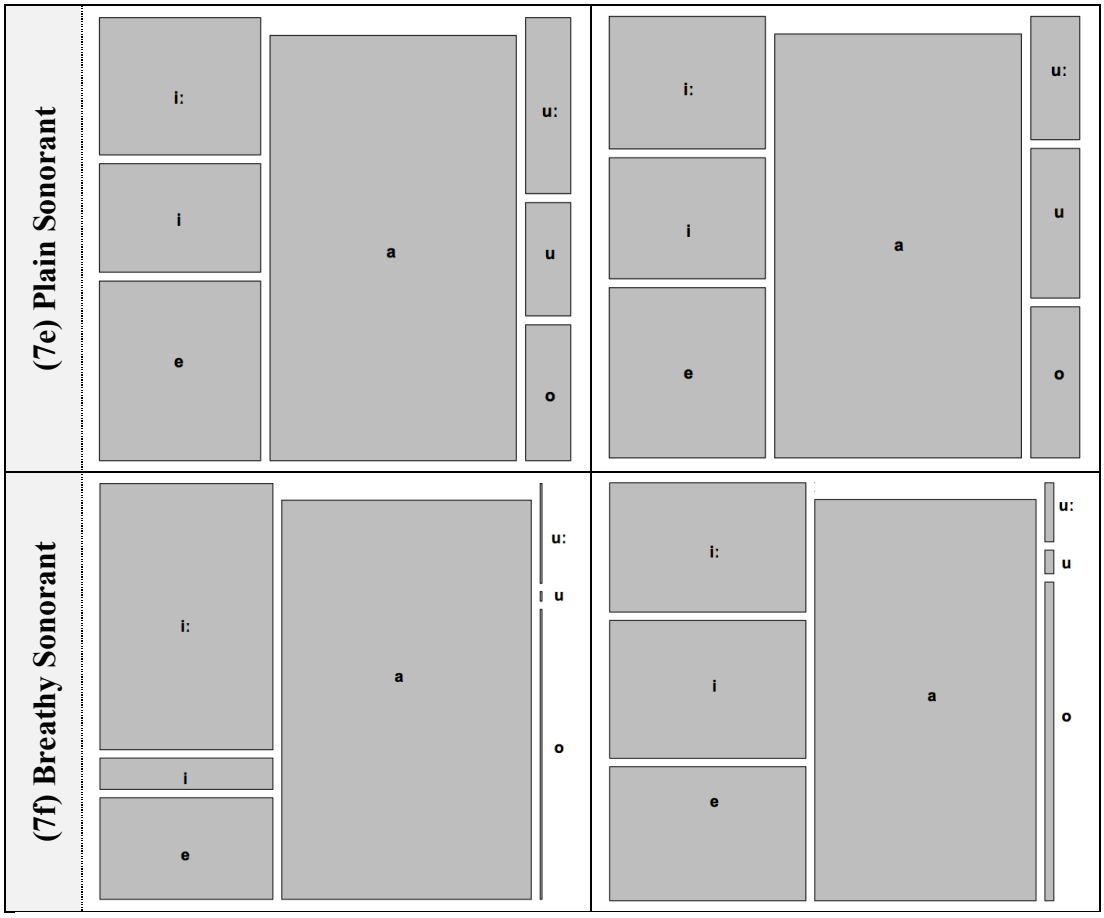
With this in mind, proportional distributions of bigrams across vowel contexts are presented below, with data for each consonant type. Token frequency is in the left panel, and type frequency in the right panel. Data presented are for the plain and aspirated voiceless obstruents (in 7a and 7b respectively), the plain and breathy voiced obstruents (7c and 7d), and the plain and breathy sonorants (7e and 7f).

Recall that approximately 46% of the vowels in the corpus are [a] whether counting by token or by type frequency, while /u/ and /o/ together account for 20% of the vowels by token and 22% by type. As seen in Fig. (7), these trends hold across almost all consonant-types: bigrams involving [a] are always of greater frequency than those involving the back rounded vowels. It also looks like bigrams involving the back rounded vowels hover around 20%, and this observation is largely supported by the percentages included in Table (3). Sonorants constitute a notable exception.

Figure (7) Bigram token and type frequency for plain and aspirated voiceless obstruents (7a and 7b); plain and breathy voiced obstruents (7c and 7d); and plain and breathy sonorants (7e and 7f). See Appendix A for a full overview of all frequency values and type-token ratios.



(7c) Plain Voiced Obstruent
i:
i
e
a
u:
u
o
i:
i
e
a
u:
u
o
(7d) Breathy Voiced Obstruent
i:
i
e
a
u:
u
o
i:
i
e
a
u:
u
o



Whereas bigrams involving back rounded vowels indeed account for about 20% of the data for most of the consonant types, they account for 10% of the plain sonorant-vowel bigrams by token (11% by type) and for just 0.4% of the breathy sonorant-vowel bigrams by token (2% by type). Token-wise, there are 18,871 breathy sonorant-vowel bigrams in the corpus, of which 85 involve back vowels. Type-wise, of the 1,632 breathy sonorant bigrams just 34 involve back vowels.

The data in Table (3) confirm these descriptive generalizations. Here, percentages represent the bigrams formed by combining each consonant type (plain voiceless obstruent, plain voiced obstruent, etc.) with each monophthong. Thus in row 1, which represents token frequency counts, we see that of 24% of the bigrams in the corpus that consist of a plain voiceless obstruents and a monophthong involve /i/, 18% /e/, 41% /a/, 6.7% /o/, and 11% /u/.

What is suggested in the mosaic plots, and confirmed by the data in Table (3), is that some of the consonant types co-occur more frequently with /a/ than would be expected based solely on the fact that /a/ is the best-represented monophthong in the corpus. These include not only the breathy sonorants, but also the breathy obstruents and the plain sonorants. This observation will be more fully investigated in future work. For now, the take-away is that breathy sonorants indeed co-occur more often with [a] than with any of the other monophthongs.

Table (3) Total percentage of CV bigrams divided by consonant-type and monophthong. For all CV bigrams composed of a plain voiceless obstruent and a vowel, for instance, 24% involve /i/, 18% /e/, and so forth. Percentages are calculated for both token and type frequency.

		/i/	/e/	/a/	/o/	/u/	back rounded
Plain Voiceless Obstruent	Token	24%	18%	41%	6.7%	11%	17%
	Type	27%	14%	41%	5.6%	12%	18%
Aspirated Voiceless Obstruent	Token	29%	19%	36%	8.0%	7.7%	16%
	Type	31%	17%	38%	6.9%	7.9%	15%
Plain Voiced Obstruent	Token	23%	18%	38%	9.2%	12%	21%
	Type	24%	15%	40%	8.1%	13%	21%
Breathy Voiced Obstruent	Token	19%	15%	52%	4.6%	9.7%	14%
	Type	22%	9%	46%	8.0%	14%	22%
Plain Sonorant	Token	21%	15%	54%	3.2%	6.8%	9.9%
	Type	21%	14%	55%	3.9%	7.0%	11%
Breathy Sonorant	Token	30%	10%	59%	0.3%	0.1%	0.4%
	Type	30%	15%	52%	1.7%	0.4%	2.1%

We posited that co-occurrence of breathy sonorants and high vowels might be restricted, but in fact that is not the pattern that is evident. Rather, it is the co-occurrence of breathy sonorants and back rounded vowels that appears to be restricted. Possible explanations for these patterns, future avenues of inquiry, and remaining questions are detailed in Section 5 below.

5 Conclusion and Future Directions

We have presented a descriptive analysis of phonotactic frequencies in Marathi, with data drawn from the Marathi portion of the EMILLE/CIIL corpus. The results indicate that in Marathi the typologically rare breathy sonorants are under-represented language internally. They account for fewer than 1% of the consonants in the corpus. About 15% of the consonants in the corpus are voiced obstruents (14% by token, 16% by type), meanwhile, and a quarter of these—4% of all consonants in the corpus by token and by type—are breathy.

As mentioned previously, breathy voiced obstruents are also uncommon crosslinguistically, though not to the same degree as breathy sonorants. We can consider the frequency information presented above in tandem with a pattern that has been observed in other languages. There are a number of languages in Nepal that at one time had phonemically breathy voiced obstruents and sonorants in their inventories, and a pattern that has been observed in them involves retention of breathy voiced obstruents and loss of breathy sonorants. Newari, which is Tibeto-Burman, is one of these: while all known dialects of Newari have retained breathy voiced obstruents, many have lost breathy sonorants (Genetti 2005). Only one dialect, Kathmandu Newar, is attested to retain breathy sonorants (Hargreaves 2003). Interestingly, one dialect (Dolakha Newar) is reportedly losing phonemic breathiness in obstruents as well (Genetti 2003; Hargreaves 2005). Breathy voiced obstruents have been retained in only a few lexical items in Dolakha Newar, and even in those the breathiness is reportedly neutralized in running speech (Genetti 2007). The Kiranti languages in Eastern Nepal pattern similarly. Camling is the

only one of 32 extant Kiranti languages that has retained breathy sonorants, though many retain breathy voiced obstruents (Ebert 2003). The frequency data from Marathi highlight the under-representation of breathy sonorants language-internally. The pattern from Newari and the Kiranti family suggests that breathiness is more likely to be retained in obstruents than in sonorants. Considered together, these facts are suggestive.

The data presented here also hint at co-occurrence restrictions between sonorants and back rounded vowels. The breathy sonorants in particular appear to be more heavily disfavored in these contexts than elsewhere: back rounded vowels account for about 20% of the monophthongs in the corpus (by type), and bigrams involving back-rounded vowels account for 15 - 20% of the total for most of the other consonant types. They account for about 10% of the CV sequences involving sonorants, however, and only about 2% of the CV sequences involving breathy sonorants contain /o/ or /u/. The questions of *why* this should be the case remains open at present, a matter for future investigation. We know of no acoustic work investigating the correlates of breathy sonorants in back rounded vowel contexts, nor of work focused on the perception of sonorant phonation-type contrasts in these vowel contexts. We return to this in a moment.

We had in fact considered the possibility that breathy sounds would be underrepresented in high vowel contexts in general. Previous acoustic studies of breathiness have sometimes restricted their investigation to the low vowel context of /a/ (Blankenship 1997; Dutta 2007: Esposito 2006). This is because the potential for formant frequency and bandwidth interference in spectral measures (Simpson 2012) is minimized in low vowel contexts like /a/ because they have a high F1. There is evidence, too, that the acoustic correlates of breathiness in sonorants are more robust in low vowel than in mid vowel contexts (Berkson 2013). This prediction was not borne out, however: the unrounded high front /i/ accounts for 23% of the monophthongs in the corpus by token, and 25% by type. Bigrams involving /i/ account for 30% of breathy sonorant data in the corpus. Several possible explanations for this deviation from the expected pattern could be investigated: as noted previously, some functional morphemes feature a breathy sonorant-high vowel sequence. The lexical items [amʱi] *'we'* and [tumʱi] *'y'all'* and morphologically-related words account for many of the instances of [mʱ] found in the corpus. Furthermore, a number of inflectional and derivational morphemes take the form of suffixal [-i], a morphological fact which could also account for the relatively high percentage of breathy sonorant-[i] sequences.

It is also the case, more broadly, that phonemic breathiness and high vowels co-occur regularly in other languages: breathy obstruents and sonorants in Marathi occur with high vowels, as in [bʱiti] *'fear'*, [dʱir] *'courage'*, and [tumʱi] *'you all'*. The same is true in other languages with breathy consonants—Sumi, for instance, contains words such as [nʱi] *'marry/betroth'*, [mʱi] *'clouded sky'*, [alʱi] *'transaction'*, and [alʱu] *'flea'* (Harris 2009:81). In Gujarati, which contains modal and breathy vowels, breathy high vowels occur as in words like [ṳlːəʈ] *'riot'* (Esposito and Khan 2012:5). Falling tone in White Hmong is produced with breathy phonation, and co-occurs with high vowels in items such as [di̤ː] *'probe/dig with a stick'* (Esposito and Khan 2012:10). Thus, the co-occurrence of breathiness and high vowels is attested both in Marathi and crosslinguistically.

Open questions about the restricted appearance of back rounded vowels with sonorants (to some extent) and breathy sonorants in particular in Marathi remain, then. As

noted, investigation of the acoustic and perceptual characteristics of breathy sonorant-high vowel sequences in Marathi constitutes an important future step.

Though the exact mechanisms underpinning the patterns of co-occurrence reported herein remain open to investigation, one clear take-away is that breathy sonorants are both typologically rare and under-represented language-internally. Given the fact that some function words (such as the examples provided above, [amʱi] *'we'* and [tumʱi] *'y'all'*) contain these sounds, it is unlikely that Marathi will follow the path illustrated by Newari and the Kiranti languages. Breathy sonorants as a class probably aren't vanishing from Marathi, at least not in the near future.

Several avenues of investigation present themselves as obvious next steps in the pursuit of a more nuanced understanding of the typology of phonation type contrasts in sonorants. Collection of data about phonotactic frequencies in other languages that contain phonation contrasts in sonorants—such as Sumi, a Tibeto-Burman language that contains breathy nasals and laterals (Harris 2009); Tsonga, a Bantu language that contains breathy nasals (Traill and Jackson 1988); or the other languages listed in UPSID (Isoko, Parauk, !Xu)—is desirable. Acoustic and perceptual investigation of the cueing of phonation type contrasts in sonorants preceding back rounded vowels is also high on the list of priorities, for the under-representation of such sequences in the EMILLE corpus is notable. Acoustic analysis of some of the function words mentioned herein will also prove enlightening: it is possible that although these items are still written with the Devanagari characters representing breathy sonorants, the breathy contrast is in fact neutralized in running speech, akin to the phenomenon reported for Dolakha Newar (Genetti 2007).

Perhaps the take home message in the present study, however, is that it is not sufficient to merely cite Marathi as a language that bucks the typological trend by containing phonation type contrasts in sonorants. Doing so fails to account for the language-internal patterning and distribution presented herein, and it is probably most accurate to say that while these sounds are indeed phonemic in Marathi their functional load is light. Breathy sonorants are both typologically rare and under-represented language-internally.

Appendix A: Bigram Data - Type and Token Frequencies, Type/Token Ratios

Table A1: LABIALS

		p	**pʰ**	**b**	**bʱ**	**m**	**mʱ**	**ʋ**	**ʋʱ**	**Total**
	token	2140	354	2795	1621	15659	3360	14568	382	40879
i:	type	674	138	815	241	1311	66	2345	106	5696
	TTR	**0.31**	**0.39**	**0.29**	**0.15**	**0.08**	**0.02**	**0.16**	**0.28**	
	token	4638	2148	2852	2930	15099	0	38318	594	66579
i	type	1532	602	1112	842	2517	0	8663	251	15519
	TTR	**0.33**	**0.28**	**0.39**	**0.29**	**0.17**		**0.23**	**0.42**	
	token	6652	1511	2694	2945	6187	8	22117	1247	43361
e	type	2261	491	905	562	1807	6	3033	169	9234
	TTR	**0.34**	**0.32**	**0.34**	**0.19**	**0.29**		**0.14**	**0.14**	
	token	156056	9582	51575	15837	124525	23036	194196	4281	579088
ə	type	48979	5801	23624	10896	47010	1221	65789	1593	204913
	TTR	**0.31**	**0.61**	**0.46**	**0.69**	**0.38**	**0.05**	**0.34**	**0.37**	
	token	53504	5820	25155	20799	62100	2505	73707	5238	248828
a	type	9070	746	5789	3958	11633	181	13972	338	45687
	TTR	**0.17**	**0.13**	**0.23**	**0.19**	**0.19**	**0.07**	**0.19**	**0.06**	
	token	9277	0	1914	2943	3932	17	5611	0	23694
u:	type	1653	0	334	752	837	2	575	0	4153
	TTR	**0.18**		**0.17**	**0.26**	**0.21**		**0.10**		
	token	14834	2932	2742	985	20612	0	303	0	42408
u	type	2984	810	1044	357	4474	0	85	0	9754
	TTR	**0.20**	**0.28**	**0.38**	**0.36**	**0.22**		**0.28**		
	token	5605	1737	5628	1808	9634	26	400	0	24838
o	type	1495	350	1225	655	1763	15	116	0	5619
	TTR	**0.27**	**0.20**	**0.22**	**0.36**	**0.18**	**0.58**	**0.29**		
Grand Total	token	252706	24084	95355	49868	257748	28952	349220	11742	
	type	68648	8938	34848	18263	71352	1491	94578	2457	

Note that here and in the remaining tables, TTR stands for Type/Token Ratio. TTR is not calculate when type or token frequency is in the single digits, as in those instances the raw numbers are more informative.

Table A2: DENTALS

		t	tʰ	d	dʱ	n	nʱ	l	lʱ	TOTAL
	token	61183	1853	9300	8421	19541	1341	38795	4	140438
iː	type	9517	353	1816	1873	4512	64	5105	2	23242
	TTR	**0.16**	**0.19**	**0.20**	**0.22**	**0.23**	**0.05**	**0.13**		
	token	18423	4238	17883	4942	24283	8	7052	1	76830
i	type	3087	568	1738	827	4939	5	2060	1	13225
	TTR	**0.17**	**0.13**	**0.10**	**0.17**	**0.20**		**0.29**		
	token	52545	5438	17166	1079	36369	202	58705	59	171563
e	type	3474	679	2721	389	7333	45	6743	24	21408
	TTR	**0.07**	**0.12**	**0.16**	**0.36**	**0.20**	**0.22**	**0.11**	**0.41**	
	token	346267	14395	66738	35890	200184	4308	134938	336	803056
ə	type	77942	6888	31655	16737	63359	435	58031	240	255287
	TTR	**0.23**	**0.48**	**0.47**	**0.47**	**0.32**	**0.10**	**0.43**	**0.71**	
	token	69621	10446	22992	12772	72377	2431	101761	573	292973
a	type	9769	2261	5312	3253	13109	209	14469	110	48492
	TTR	**0.14**	**0.22**	**0.23**	**0.25**	**0.18**	**0.09**	**0.14**	**0.19**	
	token	11063	528	2860	2235	719	0	3794	1	21200
uː	type	2595	38	569	677	200	0	293	1	4373
	TTR	**0.23**	**0.07**	**0.20**	**0.30**	**0.28**		**0.08**		
	token	15066	95	7576	1716	5412	2	644	0	30511
u	type	1731	60	1401	476	1275	2	305	0	5250
	TTR	**0.11**	**0.63**	**0.18**	**0.28**	**0.24**		**0.47**		
	token	28765	2704	7432	1090	2963	23	9939	7	52923
o	type	1670	245	725	236	840	7	1879	5	5607
	TTR	**0.06**	**0.09**	**0.10**	**0.22**	**0.28**		**0.19**		
	token	602933	39697	151947	68145	361848	8315	355628	981	1589494
Grand Total	type	109785	11092	45937	24468	95567	767	88885	383	376884

Table A3: RETROFLEXES

		ʈ	ʈʰ	ɖ	ɖʱ	ɳ	ɳʱ	ɭ	TOTAL
	token	10013	12362	8184	2357	12968	38	10237	56159
iː	type	2355	2737	2209	329	2470	1	1650	6659
	TTR	**0.24**	**0.22**	**0.27**	**0.14**	**0.19**		**0.16**	
	token	4313	2021	2856	99	19852	0	99	29240
i	type	1328	194	928	63	487	0	74	1552
	TTR	**0.31**	**0.10**	**0.32**	**0.64**	**0.02**		**0.75**	
	token	5858	7167	10707	3287	18060	11	13156	58246
e	type	1483	603	3028	383	3060	6	2598	9075
	TTR	**0.25**	**0.08**	**0.28**	**0.12**	**0.17**		**0.20**	
	token	52892	14698	45510	12045	84518	21	36833	246517
ə	type	20148	6734	22981	2991	25914	11	14725	66622
	TTR	**0.38**	**0.46**	**0.50**	**0.25**	**0.31**	**0.52**	**0.40**	
	token	11913	3994	10765	2258	36639	17	12052	77638
a	type	3068	811	2906	368	7596	4	2438	13312
	TTR	**0.26**	**0.20**	**0.27**	**0.16**	**0.21**		**0.20**	
	token	1476	471	5998	1515	9245	2	4252	22959
uː	type	307	36	1105	49	443	1	434	2032
	TTR	**0.21**	**0.08**	**0.18**	**0.03**	**0.05**		**0.10**	
	token	978	45	410	40	685	0	91	2249
u	type	258	23	215	20	323	0	42	600
	TTR	**0.26**	**0.51**	**0.52**	**0.50**	**0.47**		**0.46**	
	token	2299	473	3927	184	173	0	174	7230
o	type	1328	137	709	73	63	0	39	884
	TTR	**0.58**	**0.29**	**0.18**	**0.40**	**0.36**		**0.22**	
	token	89742	41231	88357	21785	182140	89	76894	
Grand Total	type	30276	11275	34081	4276	40356	23	22000	

Table A4: ALVEO-PALATALS

		ʧ	**ʧʰ**	**ʤ**	**ʤʱ**	**r**	**rʱ**	**ʃ**	**j**	**TOTAL**
	token	38492	44	9930	1395	43040	4	14713	1834	109452
i:	type	9832	26	2336	132	7683	3	3156	757	23925
	TTR	**0.26**	**0.59**	**0.24**	**0.09**	**0.18**		**0.21**	**0.41**	
	token	7660	347	4751	399	18532	2	13279	589	45559
i	type	1891	79	991	207	4659	1	3224	184	11236
	TTR	**0.25**	**0.23**	**0.21**	**0.52**	**0.25**		**0.24**	**0.31**	
	token	38551	499	13828	1079	21468	429	6834	28947	111635
e	type	9889	110	1009	201	4507	1	1379	4085	21181
	TTR	**0.26**	**0.22**	**0.07**	**0.19**	**0.21**		**0.20**	**0.14**	
	token	126615	1583	56390	6099	341815	75	42667	95197	670441
ə	type	65470	1181	21612	2917	109825	24	21656	75742	298427
	TTR	**0.52**	**0.75**	**0.38**	**0.48**	**0.32**	**.32**	**0.51**	**0.80**	
	token	55221	2000	30524	16964	84310	333	20118	298637	508107
a	type	12692	429	3870	746	18982	5	3944	48338	89006
	TTR	**0.23**	**0.21**	**0.13**	**0.04**	**0.23**		**0.20**	**0.16**	
	token	1364	21	4340	39	26186	1	453	2386	34790
u:	type	256	10	408	17	4068	1	163	669	5592
	TTR	**0.19**	**0.48**	**0.09**	**0.44**	**0.16**		**0.36**	**0.28**	
	token	1716	24	1840	570	1962	0	1774	6422	14308
u	type	483	16	429	242	762	0	525	1849	4306
	TTR	**0.28**	**0.67**	**0.23**	**0.42**	**0.39**		**0.30**	**0.29**	
	token	1121	512	2961	1124	10469	5	2784	9561	28537
o	type	369	60	606	328	2805	1	566	1702	6437
	TTR	**0.33**	**0.12**	**0.20**	**0.29**	**0.27**		**0.20**	**0.18**	
	token	270740	5030	124564	27669	547782	849	102622	443573	
Grand Total	type	100882	1911	31261	4790	153291	36	34613	133326	

Table A5: DORSALS

		k	kʰ	g	gʱ	h	TOTAL
	token	14737	3553	5009	181	55594	79074
i:	type	2149	996	1092	77	13599	17913
	TTR	**0.15**	**0.28**	**0.22**	**0.43**	**0.24**	
	token	12572	1054	3849	223	17689	35387
i	type	1538	334	770	47	5321	8010
	TTR	**0.12**	**0.32**	**0.20**	**0.21**	**0.30**	
	token	24095	4212	9401	9612	48201	95521
e	type	2651	1323	939	412	5018	10343
	TTR	**0.11**	**0.31**	**0.10**	**0.04**	**0.10**	
	token	220424	22498	74330	17143	64322	398717
ə	type	63082	11436	26047	4949	54643	160157
	TTR	**0.29**	**0.51**	**0.35**	**0.29**	**0.85**	
	token	89576	13156	22160	8335	43063	176290
a	type	13807	3025	5155	1217	16886	40090
	TTR	**0.15**	**0.23**	**0.23**	**0.15**	**0.39**	
	token	4303	2000	1879	897	3378	12457
u:	type	498	137	257	60	1043	1995
	TTR	**0.12**	**0.07**	**0.14**	**0.07**	**0.31**	
	token	7875	1388	5966	559	2417	18205
u	type	2059	410	1853	250	1169	5741
	TTR	**0.26**	**0.30**	**0.31**	**0.45**	**0.48**	
	token	12649	2395	7121	1240	41604	65009
o	type	2300	538	1393	365	4331	8927
	TTR	**0.18**	**0.22**	**0.20**	**0.29**	**0.10**	
	token	386231	50256	129715	38190	276268	
Grand Total	type	88084	18199	37506	7377	102010	

References

Berkson, K. (2013). *Phonation types in Marathi: An acoustic investigation* (Doctoral dissertation). University of Kansas, Lawrence.

Bethin, C. Y. (2012). On paradigm uniformity and contrast in Russian vowel reduction. *Natural Language & Linguistic Theory*, *30*(2), 425-463.

Bhagwat, S. V. (1961). *Phonemic Frequencies in Marathi and their Relation to Devising a Speed-script* (Vol. 1). Deccan College Post-graduate and Research Institute.

Blankenship, B. (1997). *The time course of breathiness and laryngealization in vowels* (Doctoral dissertation). University of California, Los Angeles.

Bybee, J. (2003). *Phonology and language use* (Vol. 94). Cambridge University Press: Cambridge, UK.

Census of India. (2001). *Statement 1: Abstract of speakers' strength of languages and mother tongues.* Census of India Website: Office of the Registrar General & Census Commissioner, India. Online resource available at <http://censusindia.gov.in/Census_Data_2001/Census_Data_Online/Language/Statement1.htm>. Accessed September 20, 2011.

Denhovska, N. (2014). *The role of frequency in implicit learning of a second language* (Doctoral dissertation). University of Manchester, Manchester, United Kingdom.

Dhongde, R. V. and K. Wali. (2009). *Marathi.* John Benjamins Publishing Co.: Amsterdam.

Dutta, I. (2007). *Four-way stop contrasts in Hindi: An acoustic study of voicing, fundamental frequency and spectral tilt* (Doctoral dissertation). University of Illinois at Urbana-Champaign, Urbana-Champaign.

Ebert, K. (2003). Kiranti languages: An overview. in G. Thurgood and R. J. LaPolla (eds.), *The Sino-Tibetan Languages*. Routledge: New York, NY, pp. 505-517.

Ellis, N. C. (2002). Frequency effects in language processing. *Studies in second language acquisition*, *24*(02), 143-188.

EMILLE/CIIL. Enabling Minority Language Engineering project (Lancaster/Sheffield Universities, UK and Central Institute of Indian Languages, Mysore). Information at http://www.lancaster.ac.uk /fass/projects/corpus/emille/.

Esposito, C. M. (2006). *The Effects of Linguistic Experience on the Perception of Phonation* (Doctoral dissertation). University of California, Los Angeles.

Esposito, C. M. and S. D. Khan. (2012). Contrastive breathiness across consonants and vowels: A comparative study of Gujarati and White Hmong. *Journal of the International Phonetic Association, 42*(2), 123-143.

Frisch, S. A., Large, N. R., and Pisoni, D. B. (2000). Perception of wordlikeness: Effects of segment probability and length on the processing of nonwords. *Journal of memory and language*, *42*(4), 481-496.

Fukazawa, H., Kawahara, S., Kitahara, M., and Sano, S. (2015). Two is too much: geminate devoicing in Japanese. *Unpublished article, available at http://user.keio.ac.jp/~kawahara/pdf/Fukazawa_etal_Revised.pdf*

Gathercole, S. E., Frankish, C. R., Pickering, S. J., and Peaker, S. (1999). Phonotactic influences on short-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*(1), 84-95.

Genetti, C. (2003). Dolakhā Newār. in G. Thurgood and R. J. LaPolla (eds.), *The Sino-Tibetan Languages*. Routledge: New York, NY, pp. 355-370.

Genetti, C. (2005). Preliminary notes on the Tauthali Dialect of Newar. Handout from the *Himalayan Languages Symposium*, December 2005. Bangkok, Thailand.

Genetti, C. (2007). *A Grammar of Dolakha Newar.* Mouton de Gruyter. Berlin, Germany.

Ghatage, M. M. (2013). Pronunciation problems of the Marathi speakers. *Language in India*, *13*(4), 107-115.

Hargreaves, D. (2003). Kathmandu Newar (Nepāl Bhāśā). in G. Thurgood and R. J. LaPolla (eds.), *The Sino-Tibetan Languages*. Routledge: New York, NY, pp. 371-384.

Harris, T. (2009). *An acoustic and articulatory study of sonorant phonation in Sumi.* Unpublished honours thesis; The University of Melbourne School of Languages and Linguistics. Melbourne, Australia.

Jha, A. (1977). An outline of Marathi phonetics. Deccan College Press: Pune, India.

Jusczyk, P. W., and Luce, P. A. (1994). Infants′ sensitivity to phonotactic patterns in the native language. *Journal of Memory and Language*, *33*(5), 630-645.

Kavadi, N. B., and Southworth, F. C. (1965). *Spoken Marathi, Book 1.* University of Pennsylvania Press: Philadelphia, PA.

Kelkar, A. (1958). *The phonology and morphology of Marathi* (Doctoral dissertation). Cornell University, Ithaca.

Leung, M. T., Law, S. P., and Fung, S. Y. (2004). Type and token frequencies of phonological units in Hong Kong Cantonese. *Behavior Research Methods, Instruments, & Computers*, *36*(3), 500-505.

Lewis, M. P., Simons, G. F., and Fennig, C.D. (eds.). (2015). *Ethnologue: Languages of the World*, 18th edition. Dallas, Texas: SIL International. Online version: http://www.ethnologue.com.

Masica, C. (1991). *The Indo-Aryan languages.* Cambridge University Press: Cambridge, UK.

Moran, S., McCloy, D., and Wright, R. (eds.). (2014). PHOIBLE Online. Leipzig: Max Planck Institute for Evolutionary Anthropology. (http://phoible.org, Accessed 11.20.15.)

Palai, E. B., and O'Hanlon, L. (2004). Word and phoneme frequency of occurrence in conversational Setswana: a clinical linguistic application. *Southern African Linguistics and Applied Language Studies*, *22*(3-4), 125-142.

Pandharipande, R. V. (1997). *Marathi.* Routledge: London, UK.

Pitt, M. A., and McQueen, J. M. (1998). Is compensation for coarticulation mediated by the lexicon?. *Journal of Memory and Language*, *39*(3), 347-370.

Renwick, M. E. (2011). Phoneme Type Frequency in Romanian. *University of Pennsylvania Working Papers in Linguistics*, *17*(1), 194-204.

Richstmeier, P., Gerken, L. and Ohala, D. (2011). Contributions of phonetic token variability and word-type frequency to phonological representations. *Journal of Child Language*, *38*(05), 951-978.

Simpson, A. P. (2012). The first and second harmonics should not be used to measure breathiness in male and female voices. *Journal of Phonetics, 40*(3), 477-490.

Southworth, F. (2000). Review of *Marathi* by R. Pandharipande. *Lingua*, *110*(3), 215-224.

Storkel, H. L. (2001). Learning New Words: Phonotactic Probability in Language Development. *Journal of Speech, Language, and Hearing Research*, *44*(6), 1321-1337.

Storkel, H. L. (2003). Learning New Words II: Phonotactic Probability in Verb Learning. *Jrnl of Speech, Language, and Hearing Research*, *46*(6), 1312-1323.

Storkel, H. L., and Maekawa, J. (2005). A comparison of homonym and novel word learning: The role of phonotactic probability and word frequency. *Journal of Child Language*, *32*(04), 827-853.

Tamaoka, K., and Makioka, S. (2004). Frequency of occurrence for units of phonemes, morae, and syllables appearing in a lexical corpus of a Japanese newspaper. *Behavior Research Methods, Instruments, & Computers*, *36*(3), 531-547.

Traill, A. and Jackson, M. (1988). Speaker variation and phonation type in Tsonga nasals. *Journal of Phonetics, 16*(4), 385-400.

UCLA Phonological Segment Inventory Database. http://web.phonetik.uni-frankfurt.de/cgi-bin/upsid_sounds.cgi.

Vitevitch, M. S., and Luce, P. A. (1998). When words compete: Levels of processing in perception of spoken words. *Psychological science*, *9*(4), 325-329.

Vitevitch, M. S., and Luce, P. A. (1999). Probabilistic phonotactics and neighborhood activation in spoken word recognition. *Journal of Memory & Language*, *40*(3), 374-408.

Vitevitch, M. S., and Luce, P. A. (2005). Increases in phonotactic probability facilitate spoken nonword repetition. *Journal of memory and language*, *52*(2), 193-204.

Vitevitch, M. S., Armbrüster, J., and Chu, S. (2004). Sublexical and lexical representations in speech production: effects of phonotactic probability and onset density. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(2), 514-529.

Yardi, V. V. (1998). *English Pronunciation for Marathi Speakers.* Saket publications: Aurangabad.