VARIATION IN TWO PATTERNS OF WORD-INITIAL DELETION IN JAKARTA INDONESIAN: INSIGHT FROM NATURALISTIC DATA*

Abigail C. Cohn, Rachel C. Vogel

Cornell University
acc4@cornell.edu, rcv44@cornell.edu

ABSTRACT

Based on naturalistic corpus data, we investigate two patterns of phonetic variation observed in Jakarta Indonesian (JI), an emerging variety of colloquial Indonesian spoken in and around Indonesia's capital, Jakarta. Word-initial [s] ~ Ø is observed in grammatical forms and taken to be lexicalized, e.g. *saja ~ aja* 'just', *sampe ~ ampe* 'until'. Word-initial [h] ~ Ø is more pervasive and said to be an optional phonological rule of H deletion, e.g. *hari ~ ari* 'day', *habis ~ abis* 'finished'. We examine the patterns of variation in these two variables for 20 speakers, in terms of lexical properties, frequency, phonological conditioning and socio-indexical factors—sex, education, and age—in order to contribute to a fuller understanding of patterns of inter- and intra- speaker variation in this rapidly developing language variety.

Keywords: Jakarta Indonesian, lexical variation, H deletion, inter-speaker variation, naturalistic corpus.

1. INTRODUCTION

Jakarta Indonesian (JI) is a colloquial variety of Indonesian (Bahasa Indonesia, BI) spoken as a first language by an increasingly large population in and around Jakarta, the capital of Indonesia. It is in some sense an admixture of Betawi Malay, the local variety of Malay historically spoken in the area, and Standard Indonesian (SI), also a variety of Malay, declared as the national language at the founding of the Republic of Indonesia in 1945. (See [7, 8] for review.)

JI fits into the complex, dynamic linguistic landscape of Indonesia, with a population of 260 million people, home to 700 languages spoken across an archipelago of over 14,000 islands. During the second half of the 20th century, Indonesian developed and was developed as a national language for this new nation-state. Indonesian became the second language of an increasingly large percentage of the population, and in recent decades, colloquial varieties have become the native language of a significant population. Based on the 2010 census [1], BI (encompassing various colloquial varieties) has become the second most widely spoken language at home, with 42 million speakers, making it the world's 30th most widely spoken native language [10].

The most significant of these colloquial varieties is JI, increasingly serving as a model for other urban varieties. And yet, to date, little linguistic work has been done on JI. Such work would not only provide much needed documentation, but also offer insight into the linguistic structure of and sociolinguistic variation in a major emerging variety as part of the linguistic ecology of a complex multilingual national capital. It is interesting to consider whether patterns of variation due to socio-indexical factors function in similar ways as in other linguistic landscapes. The relationship JI has to SI is complex and it is not just a simplified form of SI. It is thus especially important that JI be studied in its own right.

Differences between JI and SI mentioned in the literature are described as showing variable realization [4, 7, 8]. This includes grammatical properties and lexical and phonological differences. A clearer understanding of the patterns of variation observed for each of these variables, as well as similarities and differences between the patterns, is sought. Crucially, naturalistic data is needed, and fortunately a well-constructed corpus of colloquial naturalistic data exists: The Max Planck Institute for Psycholinguistics (MPI) Jakarta field station corpus of Betawi Jakarta Indonesian (BJI) [5]. As part of a larger project looking at multiple variables, here we conduct an analysis of two patterns of phonetic variation.

2. VARIABLES BEING STUDIED

We focus our study on two cases of variation between a consonant-initial form, which we refer to as the "C-initial" form, and a vowel-initial "V-initial" form, in which the initial consonant, [h], [s], or [m], is absent. As shown in (1), word-initial [s] ~ Ø is observed in several high frequency grammatical forms. An [m] ~ Ø alternation is also observed in one form.

(1) <u>initial [s]/[m] ~ Ø</u>
saja ~ aja 'just'
sampe ~ ampe 'until'
suda(h) ~ uda 'perfective'
memang ~ emang 'indeed'

We avoid calling the V-initial form a "reduced" form, as it is not clear if there is a synchronic process of reduction. Word-initial [h] ~ Ø, exemplified in (2) is observed in both high and low frequency words. (See [7] for discussion of final-[h].)

(2) initial [h] ~ Ø
hari ~ ari 'day',
habis ~ abis 'finished'
hijau ~ ijo 'green'

There is brief mention of both of these variables in the limited previous work [4, 9]. Ewing [4, p. 229] observes "There are, however, a few frequent variations in Colloquial Indonesian phonology which stand out as salient to speakers themselves. . . . These include . . . unrealized /h/. . . other examples of reduced forms . . . are limited to a small set of specific function words".

The [s]/[m] ~ Ø alternations described as being limited to function words and used in casual speech are understood to be markers of an informal register. Most of these forms have been grammaticalized, and it is claimed the original lexical form cannot be realized as the V-initial variant; thus, it is assumed that *sampe* ~ *ampe* is observed for 'until' but only *sampe* is observed for 'arrive'.

The initial [h] ~ Ø resulted from a loss of initial /h/ in Betawi and other closely related varieties of Malay [6]. Synchronically, in JI it is assumed to be an optional rule of H deletion (e.g. Sneddon [9, p. 22] "Very frequently initial *h* is lost.").

Here we investigate the observed patterns of variation for both of these variables in casual speech. We look at the lexemes that show variation in the BJI corpus and consider what structural factors account for the observed variants. We then look at the patterns of inter- and intra-speaker variation for 20 speakers, comparing sex, education, and age, to see whether these variables are used as socio-indexical markers.

3. NATURALISTIC DATA

Recently there has been increased attention to the nature of our data. It had often been assumed that laboratory speech was representative of careful speech, but much recent research leads us to question this assumption (see [2]). There are further limitations of elicitation and self-reporting when we are looking at colloquial speech phenomena, where there might be a significant gap between what speakers think they are doing and what they are actually doing. These issues are even more present in the context of an emerging colloquial variety such as JI.

An alternative approach to eliciting laboratory speech that is gaining wider currency is the analysis of speech from naturalistic corpora. However, this brings with it its own set of challenges and requirements. There needs to be metadata – identifying speakers along with key socio-indexical properties, as well as careful and reliable phonetic transcription and associated acoustic files for further analysis (see [3] for recent discussion). Additionally, such corpora are rarely available for less studied languages. Fortunately, such a corpus exists for JI.

4. METHODOLOGY

4.1. Betawi Jakarta Indonesian corpus

We analyse data from the BJI corpus [5], based on recordings done in informal settings in Jakarta, with 28 hours of recorded speech with a total of 75,079 utterances transcribed by trained native linguists in ELAN based on careful listening and imported into a relational database. The database is searchable by orthography, lexeme, phonetic transcription, morphological structure, and speaker.

4.2. Speakers

We investigate the pattern of variation in initial [s]/[m] and [h]-lexemes, for a group of 20 JI speakers, balanced by sex and educational attainment (Lower = secondary school or less, Higher = some post-secondary education). These speakers range in age from 20-49. The only previous quantitative study including some of these same lexemes [9] was limited to Indonesian "as spoken by educated Jakartan in everyday interactions" (p. 1) and did not present results based on sex.

4.3. Target lexemes

We include in our analysis all forms with s/m ~ Ø alternation and all h ~ Ø forms with at least five tokens in the corpus.

Table 1: List of lexemes ranked by total tokens (#) for 20 JI speakers and observed variants.

Lexeme	Gloss	#	C-initial	V-initial
(h)abis	after, finish	144	habis	abis, abis-nya
(h)ari	day	131	hari, hariq	ari, ariq
(h)aji	Haji	32	Haji, Hajiq	Aji
(h)idup	life	23	hidup	idup
(h)ampir	near	15	hampir	ampir
(h)item	black	10	—	item
(h)ati	liver	10	hati, hatiq	ati, atiq
(h)ijo	green	6	—	ijo, ijoq

(s)udah	PFCT	1197	sudah	udé, udéh, uda, udah, dah udaq
(s)aja	just	614	saja	aja, ajaq, ajé ajéq, aja-lah
(s)ama	with	391	sama, samaq	ama, amaq, amé, améq
(s)atu	one	205	satu, satuq	atu, atuq
(s)ampe	arrive	111	sampéq, sampé,	ampéq
(s)ama	same	92	sama, samaq	—
(s)ampe	until	14	sampéq, sampé	ampéq
(m)emang	indeed	174	mémang, memang,	émang, mang

5. RESULTS

5.1. Results by lexeme [h] ~ Ø, [s]/[m] ~ Ø

In Table 2, showing overall results, we can see very high percentages of V-initial forms for [s]/[m]-lexemes, but a roughly even split for [h]-lexemes.

Table 2: Results for each category of lexemes.

	C-initial	V-initial	Total
h-initial	46%	54%	362
s-initial	12%	88%	2532
m-initial	13%	87%	174

Looking at the results by lexeme shown below in Fig. 1, we see that %V ranges from 100% to 0%; thus some are categorically either V-initial or C-initial, while some show significant variation. Among [h]-lexemes, no [h]-initial forms are observed for *ijo* and *item* suggesting these are restructured, consistent with their forms in closely related Malay varieties. %V-initial is also very high for some of the high frequency items and function words, *habis, hampir*, but not consistently so, cf. *hari*.[1]

Turning to [s]/[m]-lexemes, the %V-initial is very high, particularly for *saja, sudah, sama* 'with' and *memang*. Sneddon [9], the only prior quantitative study, reports variations for some of the same [s]/[m]-lexemes, finding similarly high %V-initial for *sama, sudah* and *memang*. Regarding whether there is a difference between grammaticalized forms and their lexical sources, this seems to be categorically the case for *sama* 'same' vs. *sama* 'with', since there are no V-initial 'same' forms. However, we would also predict that *sampe* 'until' but not 'arrive' would show V-initial variants, but this is not borne out (50% vs. 43%). Taking token frequency in the BJI corpus as a rough measure, we see that for [s]/[m]-lexemes, higher frequency items show higher %V, but this is not the case for [h]-lexemes. We consider two additional factors—phonological conditioning and socio-indexicality—as explanations for the observed variation. We exclude *sama* 'same' and *satu* 'one' from further analysis, since they show little variation, and also *(m)emang,* as the only [m]-initial form.

5.2. Phonological conditioning

We might expect that some of the observed variation is due to phonological conditioning. Specifically, we would predict that the V-initial variants would be more likely to occur following a consonant and the C-initial variants following a vowel. We also might expect a stronger phonological effect for [h] forms than for [s] forms, if V-initial forms for the latter are indeed more lexicalized. To test these predictions, we look at the preceding environment for those forms that show the most variability, as shown in Table 3:

Table 3: % V-initial variants by preceding segment, predicted phonological conditioning shaded.

Word Form	C # _	V # _
hidup	55%	45%
idup	56%	44%
hati	100%	0%
ati	67%	33%
sampe (until)	71%	29%
ampe (until)	53%	47%
sampe (arrive)	45%	55%
ampe (arrive)	59%	41%

We see clearly that the effect is not categorical, but there is a slight tendency for a higher percentage of V-initial forms following a preceding C, but no such tendency for C-initial forms following a preceding V. We also do not see a greater effect for [h]-lexemes. Thus our predictions are not borne out.

5.3. Socio-indexical factors

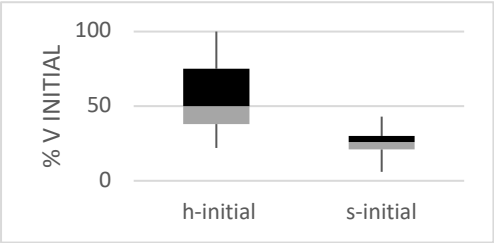
Turning to socio-indexical factors, we present % V-initial by speaker for [h] and [s]-lexemes. Based on [7], we would predict higher % V-initial for males and speakers of lower education. However, looking at Table 4, we see little difference due to either sex or education. (Based on t-tests none of these differences were found to be statistically significant at the .05 level.[2])

Table 4: Percentage of V-initial variants by sex and educational level.

	% V-Initial h	% V-Initial s
Female	56%	77%
Male	54%	80%
Lower Ed.	57%	86%
Higher Ed.	55%	88%

In Fig. 2 below, results are presented by speaker, from youngest to oldest. We predict younger speakers might show a higher percentage of [h]-initial C-variants, due to increasing influence from SI. While inter-speaker variation is seen for [h]-lexemes, no regular pattern tied to age is observed. The most striking observation is the relative consistency of V-initial forms for [s]-lexemes, as compared to [h]-lexemes shown in Fig. 3.

Figure 3: Box plots for 20 JI speakers [h]-initial and [s]-initial lexemes.



This suggests that the role of V-initial forms as a marker of colloquial speech does not interact with socio-indexical properties of age, sex, or education level at least in casual speech. Since only informal speech is available to us, it is not clear whether differences due to such interactions would emerge in more formal speech.

6. DISCUSSION

In conclusion, we see variable realization of both [h] and [s]/[m]-lexemes. In the case of [s]/[m]-lexemes, our results are consistent with previous literature [4, 9] where these are described as a marker of colloquial speech associated with JI. For [h]-lexemes, more variability across speakers is observed, suggesting that this variation is less clearly tied to colloquial speech. No systematic phonological conditioning was found and in neither case were differences observed tied to age, sex, or educational level. This is in contrast with results found for other phonological variables in the BJI corpus, where effects of both educational level and sex were seen [7]. In future work, we plan to compare observed results for JI speakers with Betawi Malay speakers and delve further into additional variables to better understand both the phonological patterning and socio-indexical marking of multiple variables within a single linguistic system.

Figure 1: Percent V-initial forms by lexeme.

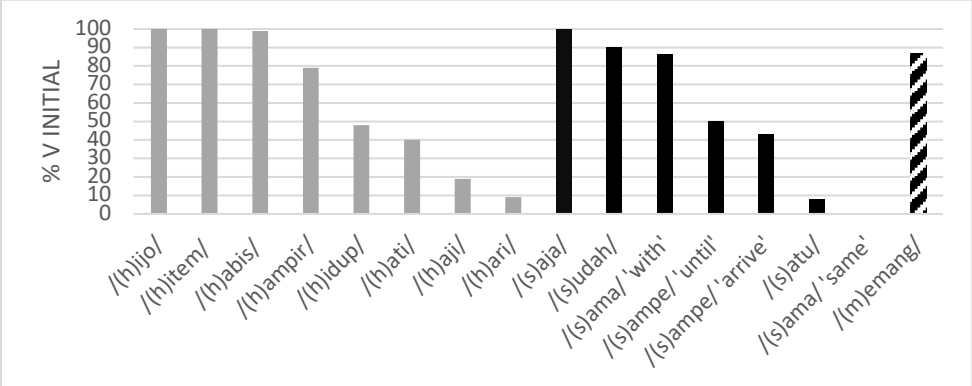
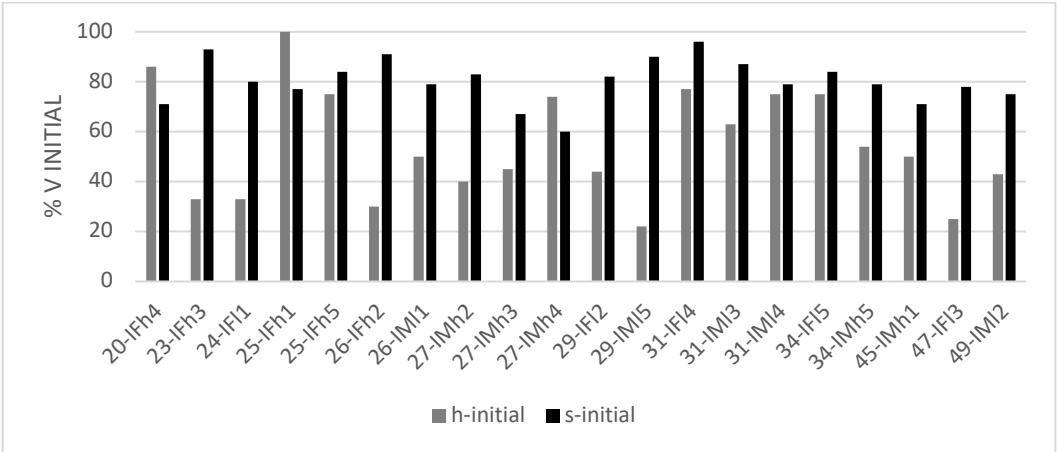


Figure 2: Percentage of V-initial forms by speaker and lexeme category.



7. REFERENCES

[1] Ananta, A., Arifin, E. N., Hasbullah, M. S., Handayani, N. B., Pramono, A. 2015. *Demography of Indonesia's Ethnicity*. Singapore: Institute of Southeast Asian Studies.

[2] Cohn, A., Fougeron, C., Huffman, M. 2018. Laboratory Phonology. In: Bosch, A., Hannahs, S. J. (eds) *The Routledge Handbook of Phonological Theory*. London: Routledge 504-529.

[3] Cohn, A. Renwick, M. 2019. Doing phonology in the age of big data. Cornell and University of Georgia ms.

[4] Ewing, M. C. 2005. Colloquial Indonesian. In: Adelaar, K.A., Himmelmann, N. (eds) *The Austronesian Languages of Asia and Madagascar (*pp.). London: Routledge, 227-258.

[5] Gil, D., Tadmor, U. 2015. *The MPI-EVA Betawi-Jakarta Database*. A joint project of the Department of Linguistics, Max Planck Institute for Evolutionary Anthropology and the Center for Language and Culture Studies, Atma Jaya Catholic University.

[6] Ikranagara, K. 1980. *Melayu Betawi Grammar*. *NUSA Linguistic Studies in Indonesian and Languages in Indonesia*, Vol. 9. Jakarta: Atma Jaya University.

[7] Kurniawan, F. 2018. Phonological Variation in Jakarta Indonesian: An Emerging Variety of Indonesian. Cornell PhD Dissertation.

[8] Sneddon, J. N. 2003. *The Indonesian Language: Its History and Role in Modern Society*. Sydney: UNSW Press.

[9] Sneddon, J. N. 2006. *Colloquial Jakartan Indonesian*. Canberra: Pacific Linguistics.

[10] Wikipedia https://en.wikipedia.org/wiki/List_of_languages_by_number_of_native_speakers, accessed 12/2/18.

* Acknowledgements: Thanks to two anonymous reviewers for their helpful comments and to Okki Kurniawan for guiding us in the corpus analysis.

[1] Our analysis here is based on the phonetic transcription provided in the corpus, but as suggested by a reviewer, it would be interesting to see whether the presence or absence of [h] is gradient.

[2] V-initial h words and gender, $p = 0.57$; V-initial s words and gender, $p = 0.38$; V-initial h words and education, $p = 0.45$; V-initial s words and education, $p = 0.49$.