The lexicalisation of HAPPINESS in the Malayic varieties of Indonesia

Gede Primahadi Wijaya Rajeg[1], I Made Rajeg[2]
[12]Universitas Udayana, Indonesia

(https://orcid.org/0000-0002-2047-8621)

Abstracts: This paper explores the lexicalisation of HAPPINESS in the Malayic varieties spoken in the Indonesian archipelago. We specifically investigate (i) the inventory of lexical forms, and the conceptual categories they encode, that refer to the generic concept of HAPPINESS in English, and (ii) how they vary along the quantitative and sociolinguistic dimensions, particularly the regional variation. Our study reveals that HAPPINESS is strongly lexicalised by forms referring to SENANG overall. In addition to this general trend, we also demonstrated that the lexicalisation of HAPPINESS concept can vary, both qualitatively and quantitatively, across the region at a more generic, concept level, and a more specific level of morph types.

Keywords: *HAPPINESS; lexicalisation; colloquial Malayic varieties; corpus linguistics; MPI EVA JFS*

INTRODUCTION

On the Notions of Lexicon and Lexicalisation

One thing that may come up in our mind when hearing the notion **lexicon** is a (set of) word(s) and the meaning, or "information content" (Jezek, 2016, p. 5), its expresses. This idea of the lexicon, which may be seen as an oversimplification, reveals two dimensions that constitute a basic consideration in any semiotic system such as language, namely **form** (e.g., lexical form, such as *earth*) and **content** (i.e., meaning evoked by or associated with the form, such as 'planet in the solar system,' 'substance on the land surface,' etc. associated with the form *earth*) (Jezek, 2016, p. 5). Meaning, when not yet associated with a lexical form, is generically conceived of as a **concept**: "mental categories carrying some information content, which can be said to exist independently from language" (Jezek, 2016, p. 5).

With respect to the aforementioned view, the notion **lexicalisation** attempts to capture the way in which concept is directly linked with a lexical form or word in a language. In fact, lexicalisation receives different interpretation (see Jezek, 2016, pp. 5–13, for an overview) from different areas of the linguistics sub-fields, such as morphology (see Hilpert, 2019) and language change (Brinton & Traugott, 2005). Word-formation processes are at the heart of lexicalisation if we view it as "the process by which new items that are considered 'lexical' (in terms of the theory in question) come into being" (Brinton & Traugott, 2005, p. 32), adding new open-class lexical items into the lexicon (Hilpert, 2019).

In this paper, we adopt the definition of lexicalisation as the lexical encoding or expressions of concept or conceptual categories (Brinton & Traugott, 2005, p. 18), such as the lexicalisation of the concept 'father's or mother's sister' by the English word *aunt* (Jezek, 2016, p. 6). This is a synchronic view of lexicalisation (Brinton & Traugott, 2005, p. 18), and may be related to the static interpretation of lexicalisation as a product/outcome, that is, the lexical inventory to express certain concept (Jezek, 2016, p. 7). For instance, *liberty* and *freedom* can be viewed as two lexicalisations of "being able or allowed to do what one wants to do" (Jezek, 2016, p. 7).

AIM OF THE STUDY

This paper provides an account of the lexicalisation of HAPPINESS in the colloquial Malayic varieties of Indonesia. The main aim is to investigate (i) the inventory of lexical forms, and their conceptual categories, that refer to the generic concept of HAPPINESS in English, and (ii) how they vary along the quantitative and sociolinguistic dimensions, particularly along the regional variation (cf. Geeraerts, 2006, p. 30). Our study contributes different dimensions and new corpus data to previous quantitative corpus-linguistic studies that have investigated HAPPINESS near-synonyms in (mostly written Standard) Indonesian, especially in terms of internal-variation in their metaphorical usages (Rajeg, 2019) and collocational patterns (Rajeg, 2020).

Inherent in this study is the notion of **onomasiology**, which "takes its starting point in a concept or referent and investigates by which different expressions the concept or referent can be designated, or named" (Grondelaers et al., 2007, p. 989). Our study also represents a facet of the broader area of usage-based, (socio)lexicology (cf. Grondelaers et al., 2007) that adopts quantitative, corpus-based approach and incorporates "lectal" variation (Geeraerts, 2006, p. 30). To pursue our aim, we use large collection of colloquial Malay/Indonesian corpora (cf. the next section) and highlight its potential for the study of language variation in the archipelago, given the large body of linguistic and non-linguistic data available in the corpora. With the help of computational data-science tool (i.e., R), we show how descriptive statistical information and geo mapping data can feed into geospatial map for the variation of the lexicalisations.

METHOD

Since 2015, the colloquial Malay/Indonesian corpora have been made public by the Max Planck Institute for Evolutionary Anthropology (MPI EVA) Jakarta Field Station (JFS) (Gil et al., 2015). MPI EVA JFS engaged in a variety of projects that involve documentation, description, and analyses of the languages of Indonesia. The adult language projects focused on varieties of Malayic languages, ranging from Sumatra, West Kalimantan, Jakarta and the neighbouring area (e.g., West Java), and the Eastern varieties of Malay (i.e., Kupang Malay, Ternate Malay, and Papuan Malay). The complete corpora also include the child-language acquisition project (of Jakarta Indonesian), but it is not the focus of this study. The corpora are downloadable as a set of tabular files with comma-separated values (.csv) format. The files are comprised of (i) metadata information, such as language name, project name, region, word counts, inter alia, and (ii) linguistic-proper database, such as the text corpus, tokenised words, morphs[1], word lists, and more. The relationship between the files is indicated by the **key** column-id (see Figure 1). These interlocking keys then allow us to merge programmatically one or more files and, eventually, build up the whole collection of the corpora into tabular database (see Table 1).

Table 1 Snippet of the corpora with selected variables/columns

Morphs text	Gloss	Languoid name	Region
tante girang	aunt happy	Ternate Malay	North Maluku
kəluwarga daʔ bahagia	family NEG happy	Tapan Binjai	West Sumatra
sinan hak- ikat bahagia	there right- bundle happy	Minangkabau, Padang	West Sumatra

The Malayic-related corpora are extracted by filtering the database whose JFS languoid-codes begin with "M," which stands for the Malayic family, or whose languoid name includes the word "Malay." In total, the Malayic-related database amounts to 5,286,633 word-tokens across 1,280 sessions. Now let us dive into some details about the nature of the texts in the corpora (such as their mode, genres, and naturalness of the production). The predominant mode of production of the text in the sessions is via spoken language (88.75%; N=1,136); written texts only constitute 6.56% of all sessions and the remaining bit (4.69%) is not classified (NA). The highest proportion (i.e., 91.33%; N=1,169) of all these sessions is comprised of naturally occurring texts; only 3.98% (N=51) of the sessions are elicited texts and the remaining 4.69% are not classified (i.e., NA). In terms of the genre of the corpora, the lion's share (65.55%; N=839) of the sessions is "conversation" genre, followed by "narrative" (20.39%; N=261), "word list" (3.59%; N=46), and other genres (0.39%; N=5), namely song, celebration, and mixed of narrative and conversation; the remaining 10.08% (N=129) is NA.

[1] According to Trask (1997, p. 144), a **morph** is "[a]ny piece of morphological material which you want to talk about, without committing yourself to any view of its morphological status."

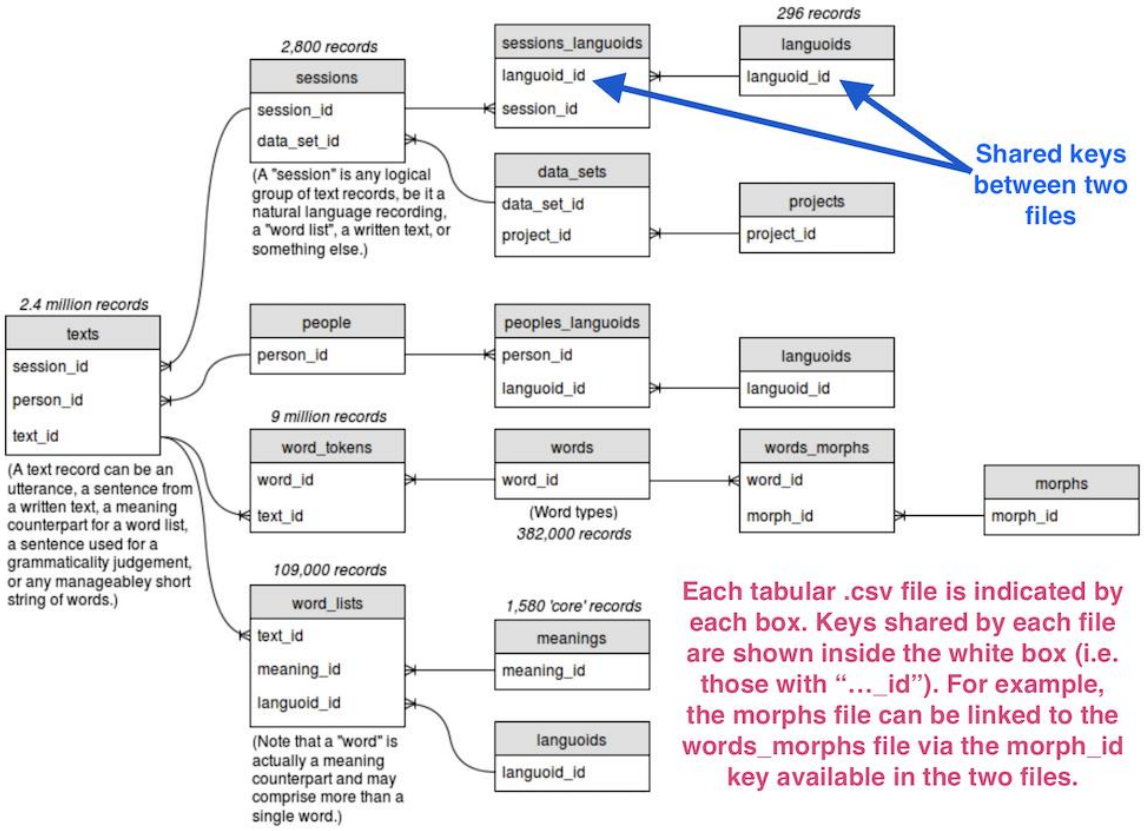

Figure 1: Schematic relationship between files in the MPI EVA JFS corpora
(Available from https://lingweb.eva.mpg.de/archive/jakarta/data.php.html)

Next, we retrieved the morphs glossed as 'happy,' 'glad,' and 'joy(ful).' Note that the values of the retrieved morphs column are the same with those in the word_form column; for this reason and succinctness, we use "morphs" in replace of "word_forms." We excluded the derivatives (e.g., as derived verbs or derived nouns) of the morphs, which will be part of a future work on their morphological profiles. Table 2 shows examples of the database (the Phon. transc.[2] column is the phonetic transcription of the values in the Morphs column).

Table 2 Snippet of the HAPPINESS lexicon database

Phon. transc.	Morphs	Gloss	Languoid name	Region
ribaŋ	ribang	happy	Besemah	West Sumatra
gəmbira	gembira	happy	Kerinci, Sungai Penuh	Jambi Province
riyaʔ	ria	happy	Bekasi	West Java

In the analyses, we group a set of related morphs under the same conceptual categories that are labelled in standard Indonesian. For instance, the conceptual category or concept of SENANG subsumes such morphs as *snang*, *sonaŋ*, *seneng*, *sanang*, inter alia, that differ slightly in terms of their phonemic structures, such as the vowels. Another example is the concept BAHAGIA that can be lexicalised by *bahɐgiɐ*, *bahagié*, *bahagiya*, among others. The quantitative analyses involve descriptive statistics, namely frequency count and percentages for univariate and bivariate designs (Gries, 2013, pp. 102, 136). In the univariate descriptive statistics, we tally the total frequency of morphs lexicalising a given concept; this way, we can assess the relative prominence (statistically speaking) of the Malayic conceptual categories for HAPPINESS (Figure 2). The bivariate analysis captures the frequency distribution of the top-five Malayic HAPPINESS concepts by regions, both at the level of token and type frequencies. The token frequency of a concept by region refers to the total occurrences of the concept in the corpus of the region. The type-frequency of a region refers to the number of different HAPPINESS concepts lexicalised in the given region. Geo spatial visualisations are included to capture the regional distribution of the HAPPINESS concepts and morphs more intuitively. All data pre-processing, statistical analyses, and visualisations are performed in RStudio using the R programming language (R Core

[2] Only 8.1% of the total 691 cases of the HAPPINESS lexicon database contain the phonetic transcription information.

Team, 2021). We wrote the paper fully in R Markdown Notebook that interleaves the regular texts and programming codes to run the quantitative and graphical analyses. The Notebook and the data are published at *Open Science Framework* (OSF) (see Rajeg & Rajeg, 2021 for the download link).

FINDING AND DISCUSSION

Distribution of the concepts of HAPPINESS in colloquial Malayic varieties

To begin with, we discuss the prominent concepts referring to HAPPINESS in the Malayic varieties. As mentioned in the previous section, the concepts are postulated from a set of morphs. Figure 2 shows the relative distribution (in percentages) of these concepts in the database.

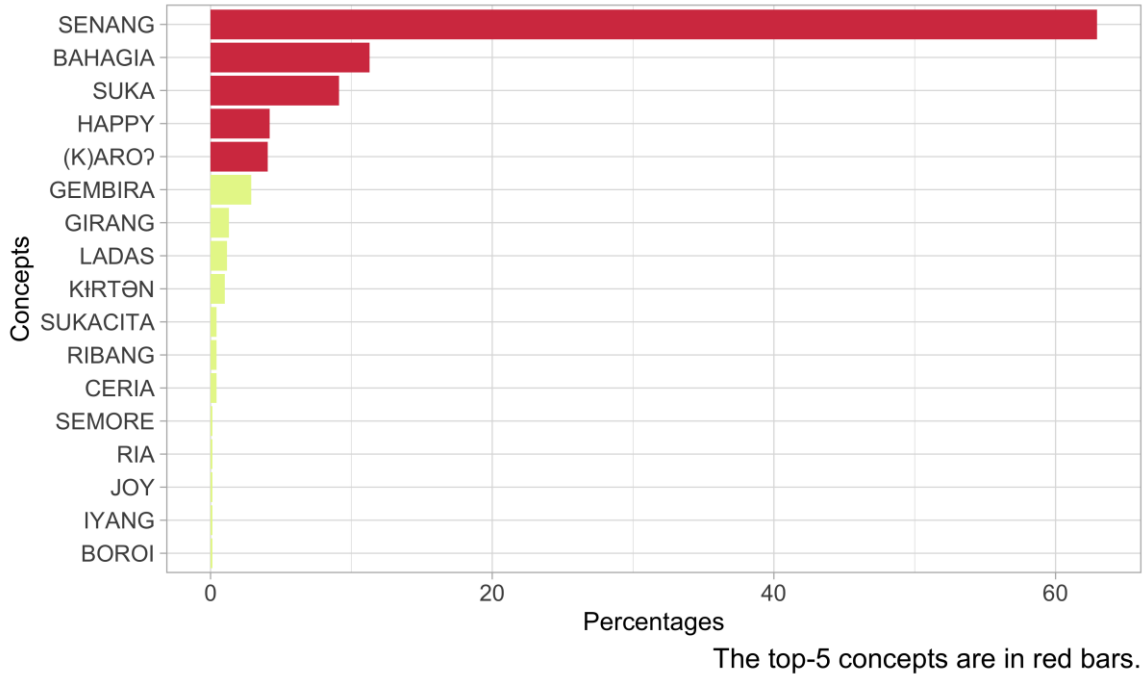

Figure 2: Percentages of the HAPPINESS concepts in colloquial Malay/Indonesian corpora

It can be observed that morphs lexicalising the concept SENANG are the most prominent for the Malayic varieties in Indonesia (N=435; 62.95% of the total tokens of all concepts); it is followed by BAHAGIA (N=78; 11.29%), SUKA (N=63; 9.12%), HAPPY (N=29; 4.2%), and (K)ARO? (N=28; 4.05%). The concept (K)ARO? is attested in the Kerinci, Pulau Tengah and Kerinci, Sungai Penuh languages spoken in the region of Jambi Province (cf. Figure 3), and postulated from these morphs: *aro?, araw?, karóq*. The prominence of SENANG and BAHAGIA in the colloquial Malayic varieties corroborates findings from previous experimental study on Indonesian emotion lexicon (cf. Shaver et al., 2001, p. 208), showing that SENANG and BAHAGIA are two of the five prototypical conceptual categories for HAPPINESS in Indonesian (together with GEMBIRA, CERIA, and RIANG).

Distribution of the top-five HAPPINESS concepts by regions

This sub-section discusses the distribution of the top-five concepts across the regions in the archipelago. This is visualised in Figure 3. The points represent certain areas within the region. If the database provides specific areas/location for a given region, but missing the geo mapping data (i.e., latitude and longitude) for these areas (indicated by NA in the database), we identified their approximate latitude and longitude information via Google Maps. For instance, data for the Jambi Province subset includes the geo mapping information for these specific projects: Tanjung Raden; Mudung Darat; Jambi City; Kerinci, Pulau Tengah. However, the other projects in the same region do not have such geo mapping data: Kerinci, Sungai Penuh; Kerinci, Tanjung Pauh Mudik; Sarolangun; Rantau Panjang. For these projects lacking the geo mapping data, we retrieved them from Google Map by searching the specific areas mentioned as the project and/or data set names.

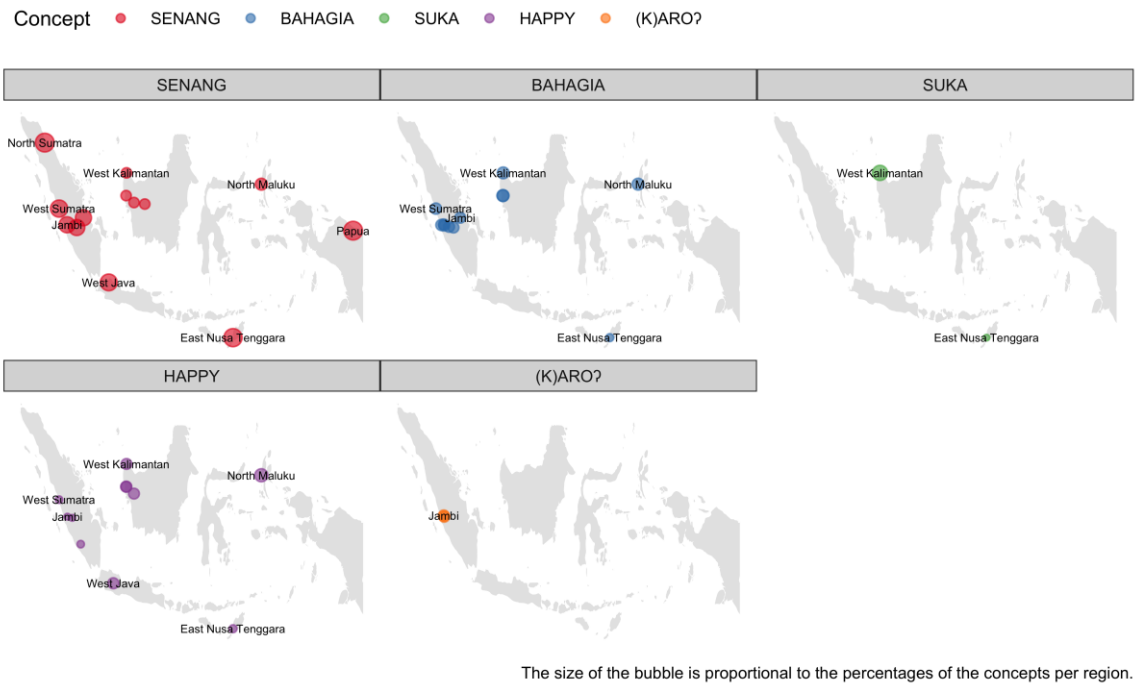

Figure 3: Top-five HAPPINESS concepts in the Malayic varieties across the regions.

In addition to being the most frequent concept overall, SENANG is also the most frequent for six out of the eleven studied regions (Figure 3). The frequency of SENANG is the highest in three of the six regions, namely in East Nusa Tenggara (N=191), Jambi Province (N=123), and West Sumatra (N=97)[3]. The English concept HAPPY is also distributed rather widely, namely in six out of the total eleven regions. In terms of its proportion, HAPPY is relatively the most frequent in North Maluku (N=3; that is, 30% of the total 10 tokens from all concepts in this region) compared to the other five regions (i.e., West Java [14.29%; N=1]; West Kalimantan [13.22%; N=16]; West Sumatra [1.56%; N=2]; East Nusa Tenggara [2.38%; N=5]; and Jambi Province [0.99%; N=2]). The presence of morphs referring to English HAPPY among the top-five concepts suggests that the lexicalisation incorporates foreign words. Another interesting finding from Figure 3 is the distribution of SUKA, in which 61 (96.83%) of its total 63 tokens glossed as 'happy' are used in Sambas Malay in West Kalimantan (cf. the panel for SUKA in Figure 3); morphs evoking SUKA are not attested to be glossed as 'happy' or 'glad' in the corpora for the other regions, except in West Kalimantan and East Nusa Tenggara.

Out of curiosity, we further checked if morphs referring to the concept SUKA are exclusively used to encode 'happy'/'glad' in the corpora from the two regions and the rests. The motivation for this is that in the standard Indonesian language, SUKA tends to lexicalise verbal concept of LIKING. To verify this assumption, we searched for the morphs glossed as 'like' and check if the morphs include those referring to HAPPINESS concepts in the top-five list. We found that only morphs referring to SENANG (in addition to SUKA) are attested to be glossed as 'like'[4]. We then calculated the relative frequency (i.e., in percentages) of 'to like' and 'happy' to be lexicalised by morphs for SUKA and SENANG in the two regions and the other regions in the database. The results are visualised in Figure 4.

[3] The other three regions where SENANG is the most frequent concept are West Java (N=5), North Sumatra (N=2), and Papua (N=2).

[4] We found that the morphs *naksir* and *nakser* are glossed with 'to like,' especially in a romantic sense. However, they occur very rarely (N=4), once in each of the following languages: Ternate Malay; Besemah; Tapan Binjai; Kupang Malay. We excluded *naksir* and *nakser* for the analysis reported in Figure 4 since they are not found to be glossed as 'happy/glad/joyful' in the database.

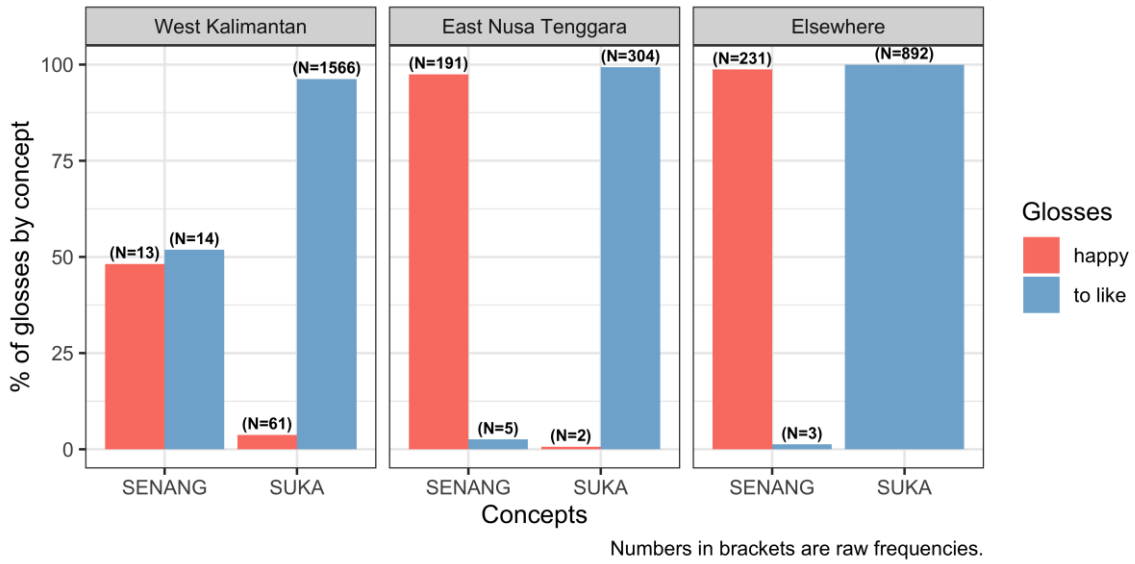

Figure 4: Distribution of morphs for the concepts SENANG and SUKA that express 'happy' and 'to like' in West Kalimantan, East Nusa Tenggara, and Elsewhere

As can be seen from all panels in Figure 4, morphs referring to SUKA are overwhelmingly used to express 'to like' than 'happy' in Kupang Malay (East Nusa Tenggara), the Malayic varieties of West Kalimantan (i.e., Pontianak Malay, Sambas Malay, Ketapang Malay, and Dayak Keninjal), and in the other regions coded in the database[5]. It is also now clear that the glossing of SUKA as 'happy' in West Kalimantan is a minority compared to its predominant gloss of 'to like,' even though the proportion of SUKA as 'happy' is still the highest in West Kalimantan. Figure 4 also shows that morphs referring to SENANG are found to be glossed as 'to like.' However, it is rare for Kupang Malay and the other regions, but is not the case for the Malayic varieties in West Kalimantan since the morphs referring to SENANG appears to have relatively similar proportion in lexicalising 'happy' (48.15%) and 'to like' (51.85%). The question as to whether this could be variation between people who did the coding of the glossing for morphs encoding SUKA and SENANG (i) remains unclear, and (ii) requires in-depth study for the usage contexts of SUKA and SENANG as to when/why they were coded as 'happy' vs. 'to like.'

The next analysis to report in this sub-section is the type-frequency analysis, measuring the number of different HAPPINESS concepts attested in each region. We found that, of the eleven regions, the five regions exhibiting the highest type-frequency are Jambi Province (eight different types of the HAPPINESS concepts), East Nusa Tenggara (seven types), and three other regions (i.e., North Maluku, West Kalimantan, West Sumatra) with ties (six types).

Distribution of morphs across the region

In the previous two sub-sections, we focus on generic, abstract level of conceptual categories, referred to by the specific morphs. It is also possible now to explore concrete, lexicalisations of each concept across the relevant regions, and reveal the phonemic variation of morphs for a given concept both qualitatively and quantitatively. For the reason of space, we illustrate this idea with SENANG[6]. Figure 5 maps out all morphs referring to SENANG; the morphs labels are available from the morphs.csv file (cf. Figure 1) and incorporated into the Morphs column (cf. Table 2).

[5] Regions constituting the "Elsewhere" category in Figure 4 are as follows: Banten; Jakarta; Jambi Province; Java; North Maluku; North Sumatra; Papua; Riau; West Java; West Sumatra

[6] Within the `figures` folder in the supplementary materials repository (Rajeg & Rajeg, 2021), we included two plots showing distribution of the morphs evoking BAHAGIA and HAPPY as in Figure 5.

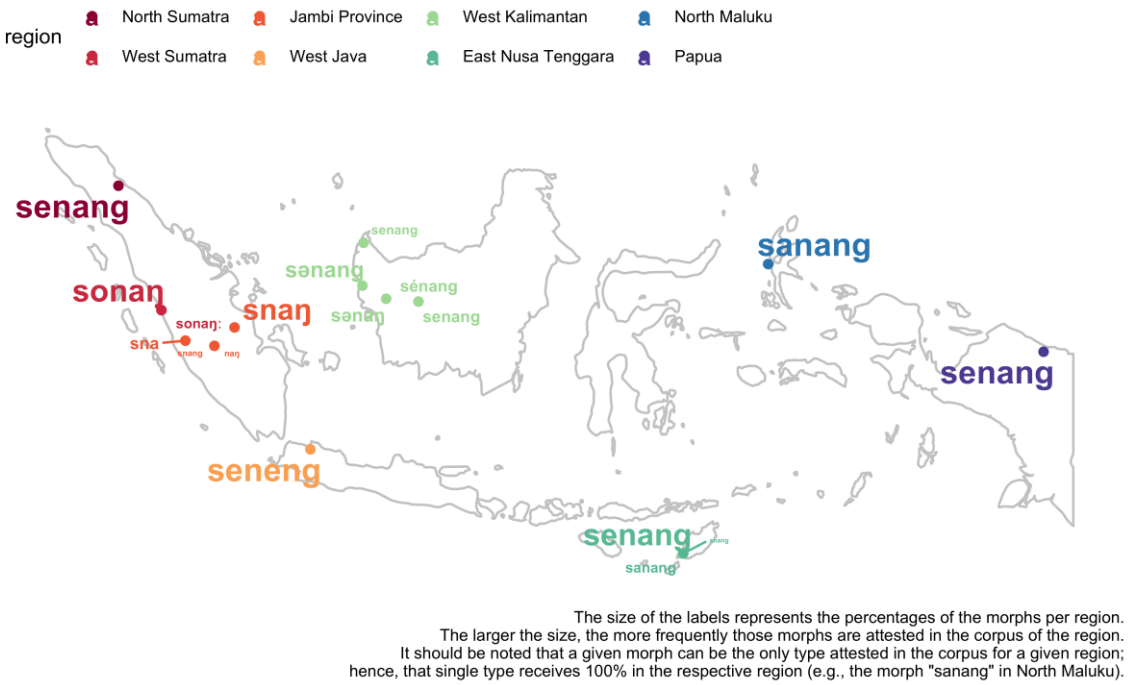

Figure 5: Distribution of morphs realising the concept SENANG across the regions

One qualitative observation that we can make from Figure 5 is that phonemic variation for the morphs lexicalising SENANG appears in the first and second syllables, such as the use of different vowel phonemes. For instance, SENANG is realised as *sənang*, *sənaŋ*, *senang*, and *sénang* in West Kalimantan. Meanwhile in Jambi Province, the morphs can vary in terms of omission of phonemes in the first syllable, such as (i) the omission of vowel leading to initial consonant cluster (e.g., *snaŋ*, *sna*, *snang*) or (ii) the omission of the initial *s* consonant (e.g., *naŋ*) and final consonant (*sna*).

Another aspect of variation that can be explored is quantitative in nature. One example is the average length (measured by the number of character) of the morphs, such that the morphs for SENANG in Jambi Province is on average shorter ($\tilde{x}$[7]=3.5; IQR=1.25) than those in West Kalimantan ($\tilde{x}$=6; IQR=0.25). We can also measure the number of different variants of morphs per region (i.e., the type-frequency) and the relative frequency of occurrence (i.e., relative token-frequency, such as percentage) of a given variant over the others (both of these measures are graphically represented in Figure 5). For example, North Sumatra, West Java, North Maluku all have only one type of morph evoking SENANG. Meanwhile Jambi Province and West Kalimantan have the highest number of morphs, namely four types, with East Nusa Tenggara and West Sumatra have respectively three and two types. We can also calculate the relative token-frequency for each morph attested in the corpus of a given region to reveal the prominence of a given variant of morph compared to the others. For instance, the three variants of morphs attested in East Nusa Tenggara (i.e., *senang*, *sanang*, and *snang*) have different rate of occurrences: *senang* is the most dominant one (as shown by its largest size in Figure 5) (N=169; 88.48%), followed by *sanang* (N=20; 10.47%) and *snang* (N=2; 1.05%).

CONCLUSION

This corpus-based study reveals that HAPPINESS is most predominantly lexicalised by morphs referring to the concept SENANG in the colloquial Malayic varieties in Indonesia, corroborating a previous experimental study on HAPPINESS conceptual categories in Indonesian (Shaver et al., 2001, p. 208). It is also demonstrated that each region varies in the distribution of each HAPPINESS concepts (e.g., variation in terms of the type and token frequencies of the concepts per region). At a more specific level, we show that the morphs lexicalising a given concept (e.g., SENANG) in certain regions also exhibit qualitative (e.g., the phonemic structure) and quantitative variations (e.g., the number of different morphs attested in a region and the morphs' token frequencies). To conclude, we hope to have provided a glimpse of insights from exploring the open-

[7] The "$\tilde{x}$" (i.e., the 'x tilde') symbol refers to a measure of central tendency called median, while "IQR" stands for Interquartile Range, which is a measure of dispersion for median.

access corpora of colloquial Malayic varieties in the Indonesian archipelago for the study of emotion lexicalisation in the context of regional variation.

REFERENCES

Brinton, L. J., & Traugott, E. C. (2005). *Lexicalization and language change*. Cambridge University Press.

Geeraerts, D. (2006). Methodology in Cognitive Linguistics. In G. Kristiansen, M. Achard, R. Dirven, & F. J. Ruiz de Mendoza Ibáñez (Eds.), *Cognitive linguistics: Current applications and future perspectives* (pp. 21–49). Mouton de Gruyter.

Gil, D., Tadmor, U., Bowden, J., & Taylor, B. (2015). *Data from the Jakarta Field Station, Department of Linguistics, Max Planck Institute for Evolutionary Anthropology, 1999-2015*. https://lingweb.eva.mpg.de/archive/jakarta/data.php.html

Gries, S. Th. (2013). *Statistics for linguistics with R: A practical introduction* (Second). Mouton de Gruyter.

Grondelaers, S., Speelman, D., & Geeraerts, D. (2007). Lexical variation and change. In D. Geeraerts & H. Cuyckens (Eds.), *The Oxford Handbook of Cognitive Linguistics* (pp. 988–1011). Oxford University Press.

Hilpert, M. (2019). Lexicalization in Morphology. In *Oxford Research Encyclopedia of Linguistics*. Oxford University Press. https://doi.org/10.1093/acrefore/9780199384655.013.622

Jezek, E. (2016). *The lexicon: An introduction*. Oxford University Press.

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Rajeg, G. P. W. (2019). *Metaphorical profiles and near-synonyms: A corpus-based study of Indonesian words for HAPPINESS* [PhD Thesis, Monash University, Australia]. https://doi.org/10.26180/5cac231a97fb1

Rajeg, G. P. W. (2020). Linguistik korpus kuantitatif dan kajian semantik leksikal sinonim emosi bahasa Indonesia. *Linguistik Indonesia*, *38*(2), 123–150. https://doi.org/10.26499/li.v38i2.155

Rajeg, G. P. W., & Rajeg, I. M. (2021). Supplementary materials for *The Lexicalisation of HAPPINESS in the Malayic Varieties of Indonesia*. *Open Science Framework (OSF)*. https://doi.org/10.17605/OSF.IO/Y42F6. https://github.com/gederajeg/malayic-happiness

Shaver, P. R., Murdaya, U., & Fraley, R. C. (2001). Structure of the Indonesian emotion lexicon. *Asian Journal of Social Psychology*, *4*(3), 201–224. https://doi.org/10.1111/1467-839X.00086

Trask, R. L. (1997). *A Student's Dictionary of Language and Linguistics*. Arnold ; Distributed by St. Martin's Press.