Vector Space Models and the usage patterns of Indonesian denominal verbs: A case study of verbs with *meN-*, *meN-/-kan*, and *meN-/-i* affixes

Gede Primahadi Wijaya Rajeg[★], Karlina Denistia[◊] and Simon Musgrave[†]

[★]Universitas Udayana, Indonesia [◊]Eberhard Karls Universität Tübingen, Germany
[†]Monash University, Australia
primahadiwijaya@gmail.com, karlinadenistia@gmail.com, simon.musgrave@monash.edu

This paper demonstrates a computational approach of *Vector Space Model* (VSM), combined with *Hierarchical Agglomerative Clustering*, to identify semantic (dis)similarity and cluster between a set of Indonesian denominal verbs with *meN-*, *meN-/-kan* and *meN-/-i* affixes. We contextualise the study within the hypotheses that some *-kan/-i* verb pairs exhibit indistinguishable as well as distinct semantics. Our VSM-based cluster analysis captures derivational families that do cluster together and those where *-kan*/*-i* pairs are separated, reflecting their distinct semantics. We also found verbs of different roots and morphologies forming coherent semantic clusters (i.e. MOTION, COMMUNICATION, and PSYCH verbs). Our quantitative corpus-based study sheds a new light on how forms with these three morphological affixes differ in their semantic distribution, providing some support to the qualitative view of semantic differences and similarity between *-i* and *-kan* derivatives.

1. Introduction[1]

Most studies of Indonesian (di)transitive verbs with suffix *-kan* and *-i* investigate the role of the suffixes as (i) valency-changing mechanisms (e.g., from intransitive *terbang* ‘fly’ to transitive *menerbangkan* ‘to fly sth. (of plane)’) and (ii) verbal derivational affixes (e.g., from noun root *calon* ‘candidate’ to transitive denominal verb *mencalonkan* ‘to nominate s.o.’) (Kroeger 2007; Arka et al. 2009; Cole & Son 2004). Recent previous works also scrutinise the intricate causative-applicative polysemy/homonymy of *-kan* and *-i* (Kroeger 2007; Arka et al. 2009; Shiohara 2012; Arka 2012). When both *-kan* and *-i* can attach to the same roots (e.g., *menjatuhkan* ‘to drop sth.’ vs. *menjatuhi* ‘to fall on sth.’ based on *jatuh* ‘to fall’), these suffixes are contrasted in terms of semantic roles of their objects. Arka et al. (2009: 5) propose that objects of *-i* verbs are linked to Locative/Goal role while those of *-kan* are linked to Patient/displaced Theme (cf. Kroeger 2007).

Up to now, however, there has been little discussion about the semantics of the verbs, as the *products* of derivational morphology with *-kan* and *-i*, in terms of their words co-occurrences (i.e. collocations) in large collection of texts (cf. §1.2 and towards the end of §2). In this paper, we demonstrate the application of *Vector Space Model* (VSM), clustering analysis, and related corpus-linguistic techniques (e.g., *n*-grams) to examine semantic (dis)similarity between sets of denominal verbs based on their usage co-

[1] **Acknowledgements** For their constructive feedback and comments, we thank the two anonymous reviewers, and the participants of ISMIL 22, 2018 at the University of California, Los Angeles, USA. We also thank Philip Chan (Monash eResearch Centre) for his assistance with MonARCH High Performance Computing Cluster. This research is partly funded by the Monash International Postgraduate Research Scholarships (MIPRS), Monash Graduate Scholarships (MGS), Monash Arts Graduate Research Funding (awarded to Gede Primahadi Wijaya Rajeg), and Indonesian Endowment Fund for Education (LPDP) (awarded to Karlina Denistia). All the remaining flaws are ours.

Rajeg, Gede Primahadi Wijaya, Karlina Denistia and Simon Musgrave, 2019. ‘Vector Space Model and the usage patterns of Indonesian denominal verbs’. In Hiroki Nomoto and David Moeljadi, eds. *Linguistic studies using large annotated corpora*. NUSA 67: 35–76. [Permanent URL: http://repository.tufs.ac.jp/handle/10108/94452] [doi: 10.15026/94452]

occurrences (§1.3, §3). We chose denominal verbs since verbal derivation with noun roots is understudied compared to that with adjectival and verbal roots (Arka et al. 2009: 1).

1.1. Hypotheses of *-kan/-i* verb pairs

Sneddon et al. (2010) propose two hypotheses regarding semantic (dis)similarity between *-kan/-i* verb pairs (i.e. *-kan* and *-i* verbs of the same roots). The first hypothesis states that "[m]any verbs can take both **-kan** and **-i**, usually with a clear distinction in meaning, **-kan** marking the object as patient and **-i** marking it as location or recipient" (Sneddon et al. 2010:100); we refer to this as the *distinctive hypothesis*. One of the given examples is the contrast between *menawari* and *menawarkan*, both translated as 'to offer':

(1)	*Mereka*	***menawarkan***	*bantuan*	*kepada*	*saya.*
	3PL	offer	help	to	1SG

'They offered help to me.' (Sneddon et al. 2010: 100)

(2)	*Mereka*	***menawari***	*saya*	*bantuan.*
	3PL	offer	1 SG	help

'They offered me help.' (Sneddon et al. 2010: 100)

Examples (1) and (2) in fact convey the same proposition of offering assistance and the boldfaced verbs co-occur with the same words as the argument fillers, namely *mereka* 'they', *bantuan* 'help', and *saya* 'I'. The so-called "clear distinction in meaning" that Sneddon et al. (2010) suggest concerns with the different mapping of the verbs' syntactic objects onto their semantic roles. Given the hypothesis, *bantuan* 'help; assistance' for the *-kan* verb in (1) is mapped onto the Patient role, while *saya* 'I' in the prepositional oblique phrase is mapped onto the Recipient role. In contrast, in (2), the Recipient *saya* 'I' is mapped onto the first object while the Patient *bantuan* 'help' is mapped onto the second.

The second hypothesis states that:

> "[w]ith a number of words the distinction between **-kan** and **-i** is blurred in common usage. In some cases, both **-i** and **-kan** occur with the same meaning. With some there is a recipient or locative object, while with others the object is the patient." (Sneddon et al. 2010:101)

The examples provided by Sneddon et al. (2010) in favour of this hypothesis, which we refer to as the *similarity hypothesis*, are only the verb pairs, without their actual sentential usages. They include *menamai* vs. *menamakan* 'to name', *mendoai* vs. *mendoakan* 'to pray for', *mengenai* vs. *mengenakan* 'subject to', among others.

1.2. Problem statements

Decontextualised examples illustrating the similarity hypothesis above are problematic and inconclusive. For instance, distinctive collocates analyses (Gries & Stefanowitsch 2004) for one word to the right (i.e. R1 collocates) of *mengenai* and *mengenakan* in one corpus file used in this paper clearly show that these verbs have completely different meanings.[2] *Mengenakan* conveys 'to wear (clothes)' sense indicated by its distinctive

[2] Data and R codes are available. The collocation analysis is performed using the association measure technique called *Distinctive Collexeme Analysis*, member of the family of *Collostructional Analysis* (Stefanowitsch 2013).

collocates overwhelmingly referring to clothing and accessories,[3] while *mengenai* conveys the 'subject to; regarding to' sense proposed by Sneddon et al (2010)[4] (cf. Rajeg & Rajeg 2019 for a similar approach in contrasting the R1 collocates of two causative verbs *memperbesar* and *membesarkan* that are presumed to be near-synonyms). In a way, collocational evidence cannot support the similarity hypothesis, at least for *mengenai/menganakan* pair, albeit their object's semantic roles are hypothesised to be of the same type. Another study by Denistia, Shafei-Bajestan & Baayen (2019) using distributional semantics further shows that Indonesian nominalising prefixes *pe-* and *peN-* are quantitatively discriminable from their semantic distribution, even though they both can express Agent. *pe-* is identified to be more similar to its noun base words compared to *peN-*.

Through this paper, we argue that argument-structures need not, and should not, be *the only* analytical level to characterise the semantic (dis)similarity of *-kan*/*-i* verb pairs. The advance of usage-based approach and corpus linguistics allows us to harness co-occurrence usage patterns of these verb pairs to capture their semantic (dis)similarity. This argument is rooted in the usage-based approach to meaning, stating that the meaning of words (e.g., *-kan*/*-i* verb pairs) can be characterised via their usage patterns (e.g., their words co-occurrences in sentences) as *actually* observed in large language corpora (§3.1), rather than being based on introspective, linguist-constructed, decontextualised evidence. Returning to the example with *mengenai/menganakan* pair, they in fact (i) exhibit "clear distinction in meaning" *if* one adopts the usage-based approach to meaning and (ii) illustrate the case for the *distinctive hypothesis* between *-kan/-i* pairs in usage patterns.

1.3. Aims

One of the main aims of this paper is to demonstrate the application of semantic *Vector Space Model* (VSM) to test the two hypotheses in §1.1 with respect to co-occurrence patterns of a set of denominal verbs, whose noun roots are attested to occur with *meN-*, *meN-/-i*, and *meN-/-kan* affixes (cf. §2 for the explanation of *meN-*). Justification for how VSM can be applied for that purpose is in order (see §3 for further details on VSM).

VSM is a machine-learning technique that has recently gained serious attention for theoretical linguistic research (cf. Hilpert & Perek 2015; Hilpert & Saavedra 2017, and the references therein). VSM leverages words co-occurrences in large language corpora to generate mathematical model representing the semantic space of each word. Semantic (dis)similarity between words is captured via spatial analogy in which a word is regarded as semantically distant from the other words when they exhibit distinct co-occurrence patterns. The semantic space model produced by the technique can be the input for further analysis, such as *Hierarchical Agglomerative Cluster* (HAC) analysis (§3.3). HAC can be utilised to determine the clustering of words. Words with similar co-occurrence patterns will cluster together and be set apart from those with distinct co-occurrences.

[3] The top-5 distinctive R1 collocates of *mengenakan* in ind_mixed_2012_1M-sentences.txt file are *pakaian* 'clothes' ($N_{mengenakan}$ = 157 vs. $N_{mengenai}$ = 2), *celana* 'pants' ($N_{mengenakan}$ = 83 vs. $N_{mengenai}$ = 2), *baju* 'shirts' ($N_{mengenakan}$ = 83 vs. $N_{mengenai}$ = 4), *gaun* 'dress' ($N_{mengenakan}$ = 37 vs. $N_{mengenai}$ = 0), and *jubah* 'cloak' ($N_{mengenakan}$ = 37 vs. $N_{mengenai}$ = 0).

[4] The top-5 distinctive R1 collocates of *mengenai* in ind_mixed_2012_1M-sentences.txt file are *hal* 'matter' ($N_{mengenai}$ = 233 vs. $N_{mengenakan}$ = 0), *masalah* 'problem' ($N_{mengenai}$ = 94 vs. $N_{mengenakan}$ = 0), *bagaimana* 'how' ($N_{mengenai}$ = 58 vs. $N_{mengenakan}$ = 0), *hubungan* 'relationship' ($N_{mengenai}$ = 49 vs. $N_{mengenakan}$ = 0), and *dampak* 'impact' ($N_{mengenai}$ = 46 vs. $N_{mengenakan}$ = 0).

The two hypotheses for *-kan/-i* verb pairs can be contextualised within the usage-based approach to word meaning. In this respect, they can be tested using VSM and explored through visualisation technique resulting from HAC (cf. Figure 4 in §5):

a) The *distinctive hypothesis* predicts that *-kan* and *-i* pairs for a given noun root will be in different cluster if they have distinct co-occurrence patterns (§5.3).

b) The *similarity hypothesis* predicts that *-kan* and *-i* pairs for a given noun root will cluster together if they have similar co-occurrence patterns (§5.2).

The inclusion of *meN-* base forms adds another layer of comparison regarding the derivational family of a given root as to whether *-kan* and/or *-i* derivatives cluster together with, or are separated from, the *meN-* base. In addition to testing the two hypotheses, we pursue two exploratory aims. Given the VSM-based cluster analysis, we are interested in discovering (i) which denominal verbs (of the given affixes) cluster together (§5.2) and which ones are set apart (§5.3), and (ii) whether the clustered verbs may exhibit some coherent semantic clusters (§5.1). To explore the cluster in more details, we use two additional methods, namely (i) *n-grams* and (ii) *nearest-neighbours* technique (§5.3), the latter of which is based on distance relation in the semantic space model (see Table 1).

Table 1. The analytical methods used in the study

Methods	**Purpose**
Vector Space Model (VSM)	To generate mathematical semantic model of words by learning co-occurrences of words in large collection of language corpus (§4.3).
Hierarchical Agglomerative Cluster analysis (HAC)	To perform cluster analysis on the semantic space model of the denominal verbs (§3.3).
Average Silhouette Width (ASW)	To automatically determine the number of clusters assumed from the data to split the studied verbs into different clusters (cf. §3.3).
Dendrogram	To visualise the results of HAC and ASW on the verbs into clustering tree in an empirical, data-driven fashion (Figure 2 and Figure 4).
N-grams/consecutive word sequences	To inspect in more details the argument assignment and collocational patterns of the verbs in the targeted clusters (cf. e.g., §5.1 and §5.2).
Nearest neighbours	To inspect why certain verb pairs are split from the cluster based on words having similar distribution with the target verbs (§5.3). To address orthographic problems in our sources that became apparent in the unsupervised learning process building our semantic model (§5.4).

In general, we seek to demonstrate how VSM and the usage-based approach to meaning sheds a new light on a long-standing issue in Indonesian linguistics, especially on the semantic (dis)similarity between *-kan* and *-i* verb pairs, and how these verbs relate to their *meN-* base forms. To this end, this paper is structured as follows: §2 introduces the verbal derivation in Indonesian, focusing on verbs with noun roots and addressing further the knowledge gap of the topic. §3 provides the details of VSM. §4 presents our methodology and databases. §5 discusses the results and §6 concludes the paper.

2. Verbal derivation in Indonesian

More formal varieties of Indonesian, both written and spoken, have a system of morphology which indicates the semantic role of the subject argument (see e.g., Musgrave 2001, chapter 2, for extensive discussion).[5] The prefix *meN-* is a part of this system. Adding the prefix to a nominal root derives a verb; such verbs vary in their syntactic transitivity and in the meaning relation between the root and the derived form. Several verbs mean ‘to go to [root]’, such as *menepi* ‘to move to the side’ (from *tepi* ‘side’), *melaut* ‘to sail’ (*laut* ‘sea’), *mengudara* ‘to air’ (*udara* ‘air’), *mendarat* ‘to land’ (*darat* ‘land’). Other verbs convey ‘to produce [root] (of sound)’ (e.g., *menjerit* ‘to scream’ from *jerit* ‘scream’) and still others ‘to become/resemble [root]’ (e.g., *membatu* ‘to petrify’ from *batu* ‘stone’; *menggunung* ‘to pile up’ from *gunung* ‘mountain’). The transitive denominal verbs with *meN-* can mean ‘to use/apply [root] to the object’, For instance, *menggunting* ‘to cut (with scissors)’ (*gunting* ‘scissors), *memborgol* ‘to handcuff’ (*borgol* ‘handcuff’) (Sneddon et al. 2010: 69–71, 73).

Denominal verb formation with *meN-* can be extended with *-kan* and *-i* suffixes that derive transitive denominal verbs (Arka & Yannuar 2016: 6). For instance, in our database (cf. §4), we found derivational family based on *susu* ‘milk’ with different frequency of occurrence, namely *menyusu* ‘(of a baby/young animal) to suckle’ (N = 458 tokens), *menyusui* ‘to breast-feed (sb.)’ (N = 2,358), and *menyusukan* ‘to let sb. suckle; to breast-feed sb.’ (N = 35) (cf. example (4) below).

Arka et al. (2009) as well as Arka & Yannuar (2016: 6–7) point out that denominal verbs with *-kan* or *-i* show diversity in the semantic roles of their object noun phrase since the noun root is not argument-taking predicates (such as adjectival or verbal roots), but the argument itself. Arka et al. (2009) propose that *-kan* and *-i* chiefly differ in the semantic roles associated with the direct object of the corresponding verbs. The locative-related roles (e.g., goal, source, or locative) are associated with *-i* (e.g., (3)) while *-kan* is associated with non-locative roles (e.g., theme, beneficiary, recipient, patient). Both suffixes can convey causative and applicative functions for the derived verbs.

(3) *Mereka* *meng-atap-i* *rumahnya.*
3PL AV-root-I house-3SG.POSS

‘They roofed the house.’ (Arka et al. 2009: 14)

Arka et al. (2009: 14) indicates that the noun root *atap* ‘roof’ in (3) is understood as a displaced theme placed on the house. In this sense, the object *rumah* ‘house’ bears a locative role (e.g., Goal) of the roof-placing event indicated by the denominal verb. Searching the *-kan* form with the root *atap* in our corpus (§4.1) yielded zero result. Example (4) illustrates the denominal *-kan* and *-i* verbs based on *susu* ‘milk’ above.

[5] The system is largely absent from informal varieties. The data we use in this study is all written material which stays close to the norms of the standard variety and we therefore assume that verb prefixes are used consistently in the data.

(4) a.

meny-(s)usu-kan	*kedua*	*bayi*	*dapat*	*bersama-sama*	(…)
AV-milk-KAN	both	baby	can	simultaneously	

‘*breast-feeding* the two babies can be done simultaneously (…)’
(ind-id_web_2015_3M:1643873)[6]

b.

(…)	*bila*	*seorang*	*ibu*	*meny-(s)usu-i*	*bayinya*	(…)
	if	ART	mother	AV-milk-I	baby.3SG.POSS	

‘(…) if a mother *breast-feeds* her baby (…)’
(ind_newscrawl_2011_1M:771081)

One may analyse that the *-kan* form in (4a) above treats the object as the Patient, in the sense that causing the baby to be breast-fed. In contrast, the exact same object type for the *-i* form in (4b), namely *bayi* ‘baby’, may be understood as the Goal towards which the breast-feeding action is directed, in the sense of giving milk *to* the object.

However, knowing that, for instance, the *-i* and *-kan* forms bear different semantic roles does not fully reveal *how* the verbs are *actually* used in natural discourse (i.e. their usage co-occurrences). Semantic role analysis may posit Locative-related roles for the objects of *-i* verbs such as *menyusui* ‘to breast-feed’, *mengatai* ‘to rebuke’ (from *kata* ‘word’), and *melangkahi* ‘to step over sth.’ (from *langkah* ‘step’), meanwhile the objects of their *-kan* partners (i.e. *menyusukan* ‘to let sb. suckle’, *mengatakan* ‘to say’ and *melangkahkan* ‘to step on sth.’) bear non-Locative roles. Yet, such analysis will not fully reveal that *menyusui* and *menyusukan*, despite their distinct roles for the objects, have similar co-occurrence profiles, thus clustering together (§5.2), but *mengatai* and *melangkahi* have distinct co-occurrence profiles and semantic niches, making them separated from their *-kan* partners (§5.3; cf. §1.2 on *mengenai* ‘subject to’ and *mengenakan* ‘to wear (clothes)’).

As noted in §1, previous works on *-kan* and/or *-i* did not offer detailed discussion for denominal verbs at large, let alone those that *do* occur with *meN-*, *meN-/-kan*, and *meN-/-i*. The gap that this paper aims to fill is a novel way to account for semantic (dis)similarity between the denominal verbs attested with these three affixes in large corpora (cf. §5). We offer a usage-based approach based on distributional semantics (§3.1) in capturing the semantic (dis)similarity and relationship between these denominal verbs. This paper goes beyond the analysis of argument-structures and semantic roles, by using words co-occurrences data. The exclusion of argument-structure and semantic roles in favour of the word co-occurrences may be viewed as the limitation of this paper. Be that as it may, while we do value the significance of argument structure and semantic roles (cf. §5.2), we argue that our approach sheds a new light on a long-standing issue in Indonesian linguistics, harnessing the availability of large corpora, the advance of computational quantitative corpus linguistics, and the usage-based model of language.

3. Vector Space Model

3.1. Underlying assumption

Vector Space Model (VSM) is a computational implementation of venerable insights in linguistics concerning the relationship between meaning of a word and its use. This idea can be traced back to Wittgenstein’s (1953, Section 43) general statement that “the meaning of a word is its use in the language”. More refined implementations of

[6] The English translation of an example is followed by information of the name of the corpus file and the sentence number (after the colon) in which the example is found.

Wittgenstein’s idea are captured by two other classic quotes often given in the VSM literature, namely from (i) John R. Firth (1957: 11), stating that “you shall know a word by the company it keeps” and from (ii) Zellig S. Harris, proposing that:

> if we consider words or morphemes A and B to be more different in meaning than A and C, then we will often find that the distributions of A and B are more different than the distributions of A and C. In other words, difference of meaning correlates with difference of distribution. (Harris 1954: 156)

This theoretical view is also known as the *distributional hypothesis* (DH) and is a key assumption that unites usage-based linguistics and corpus-based methods. A well-known implementation of the DH is the study of words’ collocation to identify the recurrent usage patterns of words and therefore to discriminate the meaning of words. VSMs are built on top of collocation data using the co-occurrence patterns of the words in large corpora to locate each word in relation to every other word in a multi-dimensional space (§3.2). Various kinds of semantic relations are reflected in the spatial relationships between words in a VSM making such models the state of the art in Computational Linguistics for modelling the semantics of human language (Turney & Pantel 2010; Erk 2012).

3.2. Constructing VSMs

Having selected a data corpus, the first step in creating a VSM is to construct a matrix tracking the co-occurrences of words in a window of some predetermined size. For each occurrence of each word, the words in the window around it are counted as co-occurring with the target word. For example, looking at six deverbal nouns (as target words) in our corpus (see the rows in Table 2) and six other words as collocates (columns), we can generate Table 2, which shows that even with only six other words considered, the target words display different patterns of co-occurrence. Each cell records the frequency of co-occurrence for the collocates within the span of three-word window either side of the target words.[7]

There are two disadvantages though to such a table, of which one can be seen in Table 2. Firstly, the data are *sparse*, namely most words in a text sample will not co-occur with each other at all and many cells in the table therefore contain a zero. The second problem is that for any usefully large text sample, the table will be very large. Our corpus contains 184,666 word-types, and the co-occurrence table will therefore be 184,666 rows by 184,666 columns.

[7] In this example, we excluded collocates of one-character token and only maintained collocates with alphabetic characters and hyphen (thus excluding numbers). The example data is based on one file of all the Indonesian Leipzig Corpora we used, namely ind_mixed_2012_1M-sentences.txt.

Table 2. Raw co-occurrence frequency for the target words with the context words

	alkitab ‘the Bible’	*dituangkan* ‘be poured’	*halal* ‘halal’	*keras* ‘hard’	*mengandung* ‘to contain’	*penutup* ‘cover’	…
bacaan ‘reading’	31	0	0	0	0	0	…
lukisan ‘painting’	0	0	0	1	0	0	…
makanan ‘food’	0	0	19	12	55	7	…
masakan ‘cooking’	0	0	0	0	1	0	…
minuman ‘drinks’	0	0	0	150	14	0	…
tulisan ‘writing’	16	6	0	2	2	0	…

The second stage in creating a VSM addresses these problems by reducing the dimensions of the table while preserving as much information as possible. The first algorithm developed for this purpose, *word2vec* (Mikolov, Sutskever, et al. 2013), uses neural networks to accomplish this, but subsequent work (Pennington, Socher & Manning 2014) has shown that other implementations are possible; mathematically, the various algorithms all perform a factorization on the original matrix, reducing it to one which has a vector with a prespecified number of dimensions representing each word in the text sample. The raw co-occurrence data is transformed to weighted data at some point in the process; this can be an independent step (as for *Glove* in Pennington et al. (2014)) or part of the factorization process (as for *word2vec* where training the neural network has this effect). The weighting is typically some version of the *Pointwise Mutual Information* measure (Church & Hanks 1990) and has the effect of giving “higher weight to context words that co-occur significantly more often than expected by chance” (Heylen et al. 2015: 156). In our analysis, we used *word2vec* (see §4.3 for further details).

An important property of a VSM is that semantic relations are represented mathematically, and these mathematical relations can be the basis for further analysis. The most common initial approach is determining the semantic (un)relatedness and (dis)similarity of various target words using some measure of distance in the multidimensional space represented by the model. This is typically done using *cosine similarity* as the measure (cf. §5.3), and the results of such analysis can then be the input to further procedures such as cluster analysis (Levshina 2014; Levshina 2015).

3.3. Exploring Vector Space Model with cosine similarity and cluster analysis

The position of each word in a VSM is defined by a vector. This vector also defines a line from the point at which each dimension of the model has a value of zero to the position of the word. One way of quantifying the amount of (dis)similarity between words in the VSM is to calculate the cosine of angles between the words’ vectors. The cosine value between a pair of word is close to 1 when they are semantically more similar, close to 0 when the words are less similar (the vectors are close to orthogonal).[8] Table 3 shows cosine similarities for the sample of words previously discussed.

Table 3. *Cosine Similarity* matrix between the deverbal nouns (values are rounded)

	bacaan 'reading'	*lukisan* 'painting'	*makanan* 'food'	*masakan* 'cooking'	*minuman* 'drinks'	*tulisan* 'writing'
bacaan 'reading'	1.00					
lukisan 'painting'	0.04	1.00				
makanan ' food'	0.02	0.01	1.00			
masakan 'cooking'	0.03	0.03	0.03	1.00		
minuman 'drinks'	0.02	0.04	0.03	0.04	1.00	
tulisan 'writing'	0.06	0.04	0.00	0.02	0.02	1.00

It can be seen from Table 3 that *tulisan* 'writing' and *bacaan* 'reading' has the highest similarity score, while *tulisan* 'writing' and *makanan* 'food' is not similar. Of course, each word is similar to itself, hence their cosine similarity of 1 along the diagonal.

Such matrices in Table 3 become difficult to interpret when there are large numbers of words to examine. It is then useful to employ more objective methods than simply eyeballing the numbers in the table to make conclusion. One such method is to perform cluster analysis, such as *Hierarchical Agglomerative Clustering* (HAC) (Levshina 2014; for details on HAC implementations in R, see Gries 2013: 336; Levshina 2015: Ch. 15; Desagulier 2017: 276). HAC takes as input a matrix of distance between words (i.e. the inverse of the similarity matrices) that can be derived from the cosine similarity matrices (Levshina 2015: 330). The output of HAC can be visualised as dendrogram (see Figure 2 and Figure 4) in which the objects (i.e. the target words) are represented as "branches of a clustering tree" (Levshina 2015: 309).

The optimal number of cluster solution is determined using the *Average Silhouette Width* (ASW) score (Gries 2013: 348; Levshina 2015: 311). ASW "quantifies how similar elements are to the clusters which they are in relative to how similar elements are to the other cluster" (Gries 2013: 348). ASW indicates "average well-formedness of the clusters in a given solution", meaning that "the members of one cluster are close to one another and far away from the members of the other clusters" (Levshina 2015: 311). Given the six target words, we tested two- up to five-cluster solutions;[9] the cluster solution with highest ASW score, in this case two clusters, is preferred (see Figure 1). ASW score

[8] Mathematically, vectors can also have opposite directions, and then the cosine similarity approaches –1. We assume that text based VSMs are positive spaces and that cosine similarity for them ranges between 0 and 1.

[9] The one-cluster solution (all six target words form one cluster) and the six-cluster solution (each word forms a one-member cluster) were not tested.

ranges from 0 (no clustering) to 1 (perfect separation/clustering). The rule-of-thumb for assuming substantial clustering is ASW ≥ 0.2 (Levshina 2015: 311).

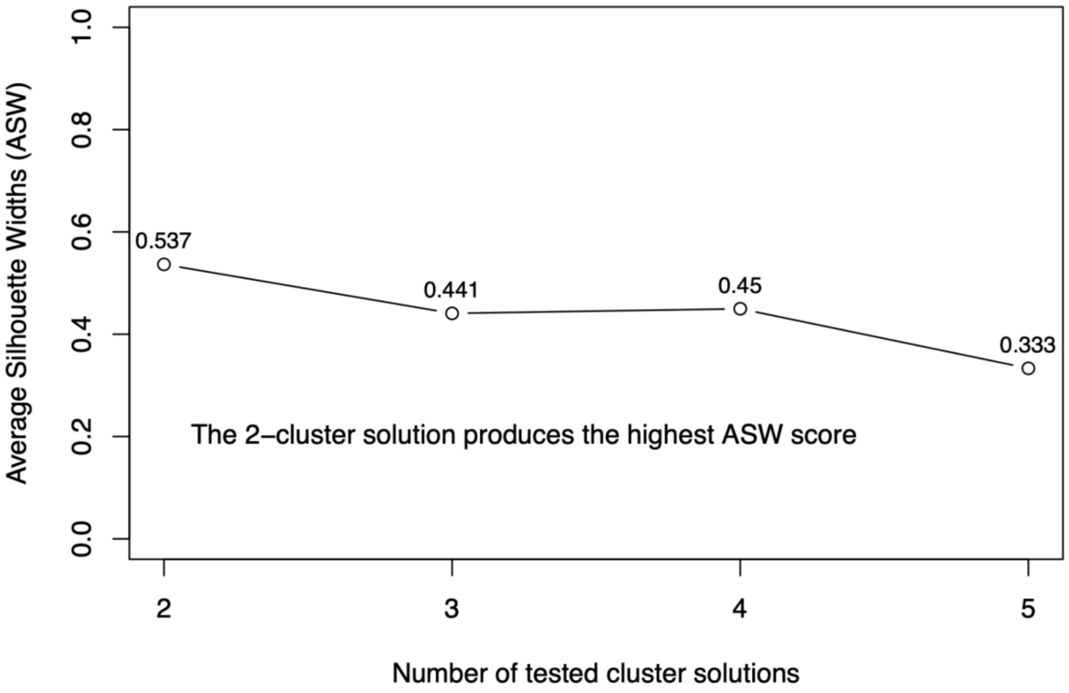


Figure 1. ***Average Silhouette Width*** **(ASW) scores (y-axis) for the possible cluster solutions (x-axis) for the six deverbal nouns data**

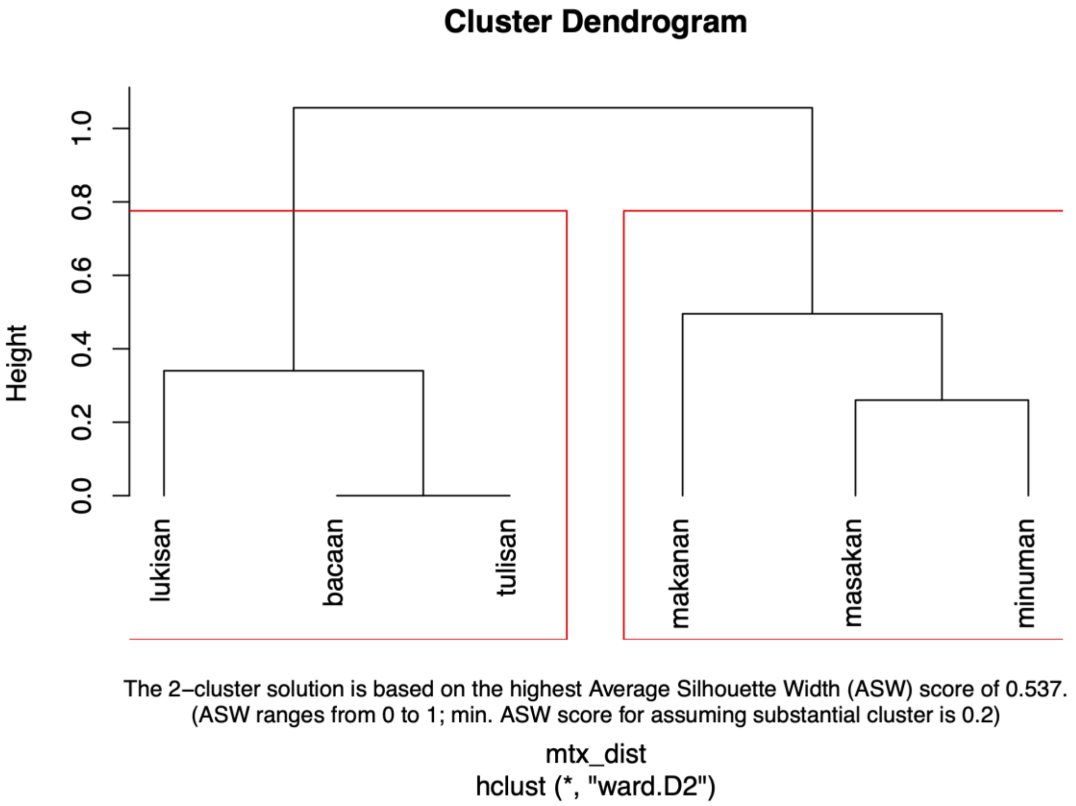


Figure 2. HAC clustering for the deverbal nouns data

Figure 2 shows the two-cluster solution (indicated by the red boxes) as well as further clustering within each of the two. In the left-hand cluster, *bacaan* ‘reading’ and *tulisan* ‘writing’ are merged first as can be seen through their shortest branch showing most similarity. They are less similar to *lukisan* ‘painting’, despite their clustering. Similarly, *masakan* ‘cooking’ and *minuman* ‘drinks’ are more similar to each other than either is to *makanan* ‘food’. Note that the two-cluster solution appears to show semantically interpretable classes. One is nourishment-related words (right cluster) and the other (left cluster) is words related to visible representations of information.

In §5, we use HAC and ASW for detecting and visualising the clustering of the denominal verbs. All analyses are performed in R (ver. 3.6.0). For the HAC, we use the `hclust()` function from the `stats` package in the base R with the `"ward.D2"` clustering method that “usually produces compact and interpretable clusters” (Levshina 2015: 309). The ASW is computed using the combination of (i) `silhouette()` function from the `cluster` package (Maechler et al. 2018) and (ii) the `cutree()` function for cutting the tree into the specified tested cluster.

4. Data and methods

4.1. Data source

This paper uses thirteen corpus files of the *Indonesian Leipzig Corpora* (Biemann et al. 2007; Goldhahn, Eckart & Quasthoff 2012; Quasthoff & Goldhahn 2013). Leipzig Corpora are corpus texts of shuffled sentences without any linguistic tagging (e.g., part-of-speech tagging, syntactic or morphological parsing). The materials for the corpora are predominantly sourced from online newspapers of Indonesian, randomly selected web pages, and Wikipedia dumps. The plain text files of the Leipzig Corpora can be downloaded for free from http://wortschatz.uni-leipzig.de/en/download. Table 4 shows the size for all corpus files combined to generate a vector space model (see §4.3). The size calculation excludes sentence numbers preceding each sentence.

Table 4. Indonesian Leipzig Corpus files used in this study

	Corpus files	Size (in word-tokens)
1	ind_mixed_2012_1M-sentences.txt	15,052,159
2	ind_news_2008_300K-sentences.txt	5,875,376
3	ind_news_2009_300K-sentences.txt	5,868,276
4	ind_news_2010_300K-sentences.txt	5,874,158
5	ind_news_2011_300K-sentences.txt	5,852,211
6	ind_news_2012_300K-sentences.txt	5,873,523
7	ind_newscrawl_2011_1M-sentences.txt	16,376,426
8	ind_newscrawl_2012_1M-sentences.txt	16,916,778
9	ind_web_2011_300K-sentences.txt	4,472,885
10	ind_web_2012_1M-sentences.txt	15,844,629
11	ind_wikipedia_2016_1M-sentences.txt	16,506,714
12	ind-id_web_2013_1M-sentences.txt	16,406,671
13	ind-id_web_2015_3M-sentences.txt	49,849,398

In total, the thirteen corpus files amount to 180,769,204 word-tokens.[10] A *word-token* is defined as any one or more alphanumeric characters, including hyphen (to maintain reduplication, such as *anak-anak* 'children'), separated by whitespace.

We created frequency list from the thirteen corpus files from which the words beginning with *me-* were extracted. We then generated two frequency lists for the *me-* words: (i) the aggregated frequency of each word across all corpus files, and (ii) the frequency of each word per corpus file. The latter is used to keep track of the corpus file in which each *me-*word is attested. This can be useful for further inspection of the full context of certain odd *me-* words.

4.2. Compiling the studied denominal verbs

We used a number of criteria to retrieve the target verbs from the results of parsing the wordlist with the Indonesian morphological parser *MorphInd* (Larasati, Kuboň & Zeman 2011). First, we filtered non-hyphenated words tagged as active verbs (i.e. `_VSA`; see Table 5, `morphind` column), which (i) occur in the three morphological patterns and (ii) show more than twenty tokens across all files. Second, among these filtered words, we took only those with noun-tagged roots. For instance, we take ***men****dasar* 'to be basic', ***men****dasar**i*** 'to underlie sth.', and ***men****dasar**kan*** 'to base sth. on' that are all derived from the nominal root *dasar* 'base; foundation'. Lastly, we excluded tokens with pronominal suffixes identified as their encliticised direct object argument (e.g., *mendasari**nya*** 'to underlie *it*' and *mewakili**ku*** 'to represent *me*; be *my* representative' are excluded). After applying these criteria, we were left with a list of 51 denominal verbs based on 17 root types. Table 5 offers a snippet of the database for the root *dasar* 'base; foundation'.

Table 5. Snippet of the studied denominal verbs with output from *MorphInd*

word	token_freq	root	root_pos	morphind	affix
mendasar	4571	*dasar*	n	meN+dasar<n>_VSA	*me*
mendasari	1365	*dasar*	n	meN+dasar<n>+i_VSA	*me.i*
mendasarkan	781	*dasar*	n	meN+dasar<n>+kan_VSA	*me.kan*

We also checked the retrieved target verbs in the *MALINDO Morph* (Nomoto et al. 2018), a morphological dictionary for Malay/Indonesian language.[11] Unlike *MorphInd*, *MALINDO Morph* currently does not tag the root words with part-of-speech information. There are five words of the 51 possibilities that are not attested in the version of *MALINDO Morph* we used, but in our corpus: *mengakhir* (N = 57 tokens), *membuah* (N = 56), *mengantung* (N = 28), *mewakil* (N = 39), and *mewaris* (N = 24). The database for *MALINDO Morph* is also based on Leipzig Corpora, as is our corpus; the reason for this discrepancy is that the *MALINDO Morph* database includes only the files with 300K sentences and the resulting dictionary only includes words with more than ten tokens in these files (Nomoto et al. 2018: 39). Our per corpus frequency list confirms that there are fewer than ten tokens of each of these words in all of the 300K corpus-files we used.

[10] On the last visit to the download page for the Indonesian corpora (on 14 December 2018 at 12.29 pm), the 3-million-sentence corpus file for ind-id_web_2015_3M-sentences.txt is no longer available; only the 1-million-sentence file is available for download.

[11] *MALINDO Morph* can be accessed at https://github.com/matbahasa/MALINDO_Morph. We used the dictionary file named malindo_dic_20181125.tsv when we wrote the analyses in the paper.

4.3. Generating the vector space models with *word2vec*

This paper uses the recent vector space algorithm called *word2vec* developed by Thomas Mikolov and colleagues at *Google* (Mikolov, Yih & Zweig 2013; Mikolov, Sutskever, et al. 2013; Mikolov, Chen, et al. 2013). We use the R wrapper of the original C codes from the `wordVectors` R package (Schmidt & Li 2017).

Pre-processing on the input corpus included removing punctuations,[12] numbers, one-letter tokens, and lowercasing the word-tokens. To run the model training, we set the parameters[13] as follows:

- Output vector space dimension: 100 dimensions
- Context window: 10-word context-window around a given target word
- Min_count: 10 tokens, which is the minimum token frequency of a word in the whole corpus to be included in the training
- Training algorithm: the default *skip-gram* method

The *skip-gram* method attempts to learn the context words of a given target word and to "maximize classification of a word based on another word in the same sentence" (Mikolov, Chen, et al. 2013: 4). The other training algorithm of *word2vec* is *continuous bag of words* (CBOW), which attempts to learn the target word given its context words.

The output of the training procedure is a matrix table, with all word types occurring at least ten tokens in the input texts as the rows and their vector space representations (up to 100 dimensions) as the columns. The total word types (i.e. rows) in the model are 184,666 words. For further analysis, we extracted the vector matrix of the 51 denominal verbs that we analysed, and we used this vector matrix as the basis for calculating cosine distance measures between the verbs. The distance measures then became the input for *Hierarchical Agglomerative Cluster* (HAC) analysis.

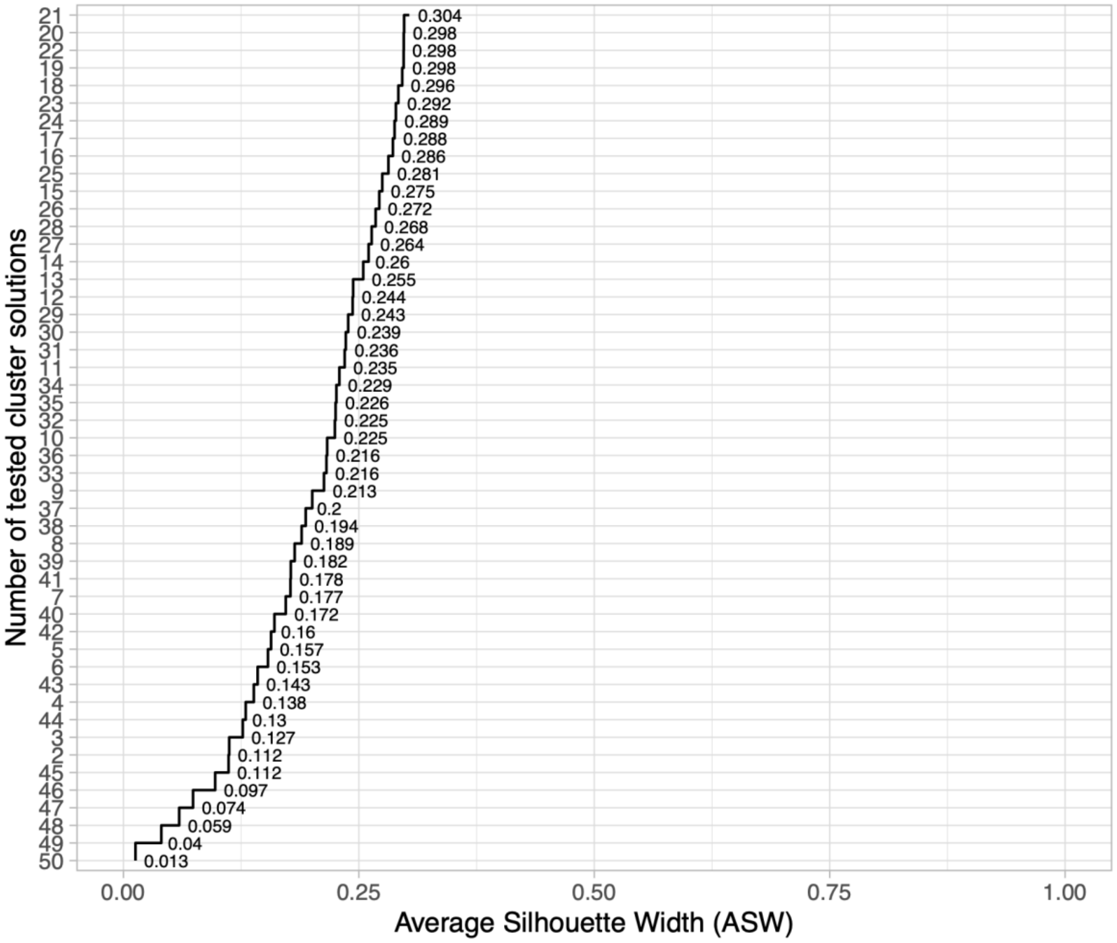
The data used in this paper, including the vector space model, and the R Markdown Notebook containing the R codes for analyses, are available open access (see Rajeg, Denistia & Musgrave 2019a; Rajeg, Denistia & Musgrave 2019b).

5. Results and discussion

Given that we analysed 51 denominal verbs, we tested 2 to 50-clusters solution via *Average Silhouette Width* (ASW). The highest ASW score of 0.304 is for 21-cluster. Figure 3 captures the series of ranked order ASW scores (*x*-axis) over the tested cluster solutions (*y*-axis). The optimal number of cluster solution is shown at the *y*-axis at the top of the plot. The plot is produced with the `ggplot2` R package from the `tidyverse` (Wickham & Grolemund 2017). Figure 3 thus serves similar function as Figure 1.

[12] Each line in the corpus file of the Leipzig Corpora represents a sentence. The results of removing punctuations did not change the structure of the corpus file, in the sense that each line of the cleaned corpus still represents a sentence (with only punctuation removed). Therefore, the input file for VSM training still consists of sentence in each line.

[13] See Ben Schmidt's GitHub at https://github.com/bmschmidt/wordVectors for details on `wordVectors`. Moreover, this paper does not aim at comparing results from different vector space models produced via varying the training parameters. Model comparison requires a separate paper. Interested readers are referred to Kiela & Clark (2014) who investigate the impact of varying training parameters on the resulting models.



The tested cluster solution ranges from 2 to 50 (i.e., the length of the analysed words (51) – 1).
The highest ASW of 0.304 is for 21–cluster solution

Figure 3. ***Average Silhouette Widths*** **(ASW) for the tested cluster solutions; the optimal number of cluster solutions is 21 clusters, having the highest ASW score of 0.304**

Next, the 51 denominal verbs were automatically grouped into 21 clusters and visualised via cluster dendrogram (Figure 4), using functions from the `dendextend` package (Galili 2015). The horizontal, *x*-axis in Figure 4 shows the distance of the merger between branches (i.e. between the target denominal verbs). Mergers that are further to the right along that axis (i.e. closer to 0.0) reflect greater similarity between the merged elements; mergers that are further to the left (i.e. away from 0.0) reflect lesser similarity.

In general, the VSM-based cluster analysis captures both the distinctive and similarity hypotheses for *-kan/-i* verb pairs that we adapt from Sneddon et al. (2010) (§1.1). The split between *-kan* and *-i* pairs may reflect the distinctive hypothesis (§5.3) while the clustering of both *-kan* and *-i* pairs may reflect the similarity hypothesis (§5.2). The clustering also reveals several semantically coherent verbs from different roots (§5.1).

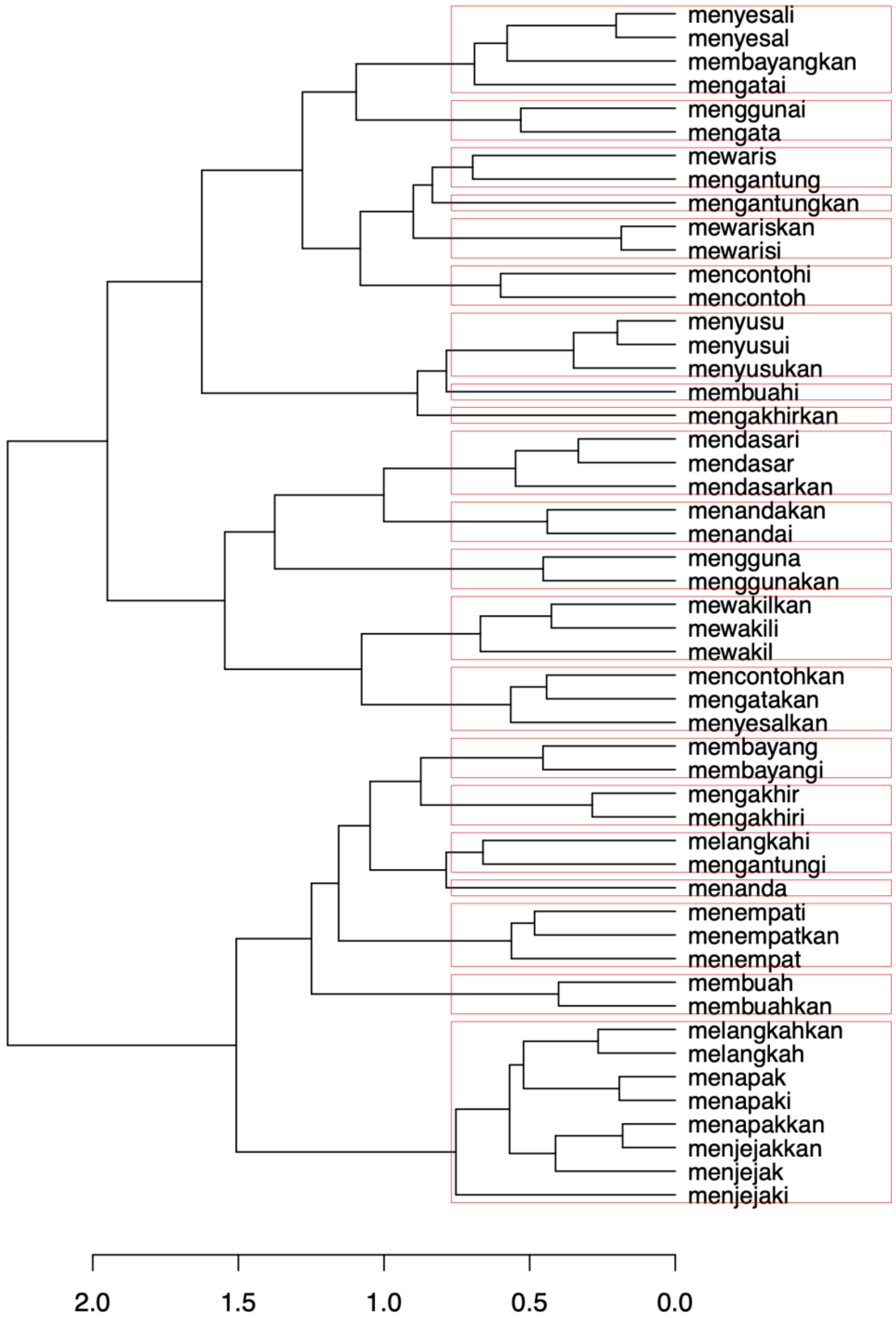


Figure 4. *Hierarchical Agglomerative Clustering* (HAC) dendrogram for the denominal verbs (Distance measure = *Cosine Distance*; Clustering method = *Ward.D2*)

The inclusion of the base *meN-* forms adds another layer of comparison. For most verbs, the base form and the two derivatives with *-kan* and *-i* cluster together. Yet, there are several bases for which one derivative separates from the other two forms. The separated form can be either the *-i* (e.g., *membuahi* 'to fertilise') or the *-kan* derivative (e.g., *mencontohkan* 'to exemplify'). Two cases where the base form is separated from the two derivatives are *menanda* (from *tanda* 'sign') and *mengata* (from *kata* 'word'); §5.4 discusses orthographical issue in our fully unsupervised methods using *menanda* as an example. Overall, different clustering patterns between these affixes indicate differences in the resulting co-occurrence usage patterns of the verbs that the affixes derive.

5.1. Semantic clusters of morphologically heterogenous verbs

At least three semantic clusters can be inferred from Figure 4. These clusters consist of morphologically heterogenous verbs, namely verbs based on formally different noun roots. The most coherent group consists of verbs evoking (physical or metaphorical) MOTION (see the cluster at the bottom of Figure 4 extracted as Figure 5). These MOTION-related verbs are based on FOOT-related roots:

a. *langkah* ‘step’ → *melangkah* ‘to stride; move on; take a step’ (N = 5,193 tokens); *melangkahkan* ‘to move the foot forward’ (N = 339)

b. *tapak* ‘sole of the foot’ → *menapak* ‘to step (the sole of the foot)/thread on X; to walk barefooted’ (N = 303); *menapakkan* ‘to step on the sole of the foot’ (N = 79); *menapaki* ‘to walk in; to set foot on; to enter into’ (N = 582)

c. *jejak* ‘footprint’ → *menjejak* ‘to step (foot) on; to trace/track (down); to reach a phase (e.g., in a competition)’ (N = 165); *menjejakkan* ‘to step sth. on’ (N = 243); *menjejaki* ‘to step (foot) on; to trace/track (down)’ (N = 48)

This cluster has the most members and the three morphological patterns cluster together with the only exception being the *meN-/-i* form of the root *langkah* ‘step’, namely *melangkahi* ‘to step over’ (N = 172), which is quite far away from this cluster. §5.3 takes up in further detail such split cases between *meN-/-i* and *meN-/-kan*.

These MOTION verbs indeed exhibit within-cluster differences, observable through the length of the merge (see the *x*-axis) between the verbs in the cluster. The shorter the merge between a pair of verbs (i.e. the closer it is to 0.0), the more similar they are in their vector space representations; the further the merge to the left, the more dissimilar they are.

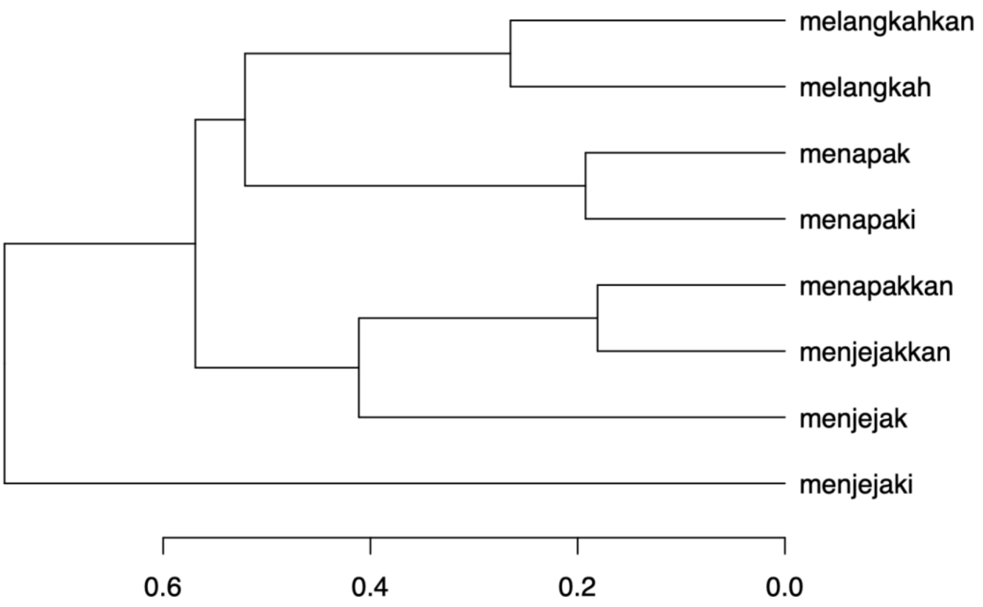


Figure 5. Cluster for the MOTION verbs

The branch for *menapakkan* ‘to step on the sole of the foot (more actively)’[14] is merged first with another *meN-/-kan* verb of different root, namely *menjejakkan* ‘to step the foot’ at the point of 0.181 (at the horizontal, *x*-axis). *Menapakkan* is in separate branch with its base form (i.e. *menapak* ‘to step on’) and its *meN-/-i* derivative (i.e. *menapaki* ‘to walk in’), even though the *meN-/-i* and *meN-/-kan* forms are transitive verbs; their branches are joint at the point of 0.569. We argue that these two transitive verbs with *tapak* ‘sole of the foot’ have different co-occurrence domains. Our VSM-based cluster analysis offers high-level, within-cluster differences through their internal merge and branching.

Should one wish to investigate further the detailed differences between these verbs, one effective way is extracting the *n*-grams for each verb from the corpus. Consider Table 6, showing 3-gram patterns for *menapak*, focusing on its right-side patterns.

Table 6. The ten most frequent 3-grams for *menapak* 'to step on'

	3-grams	gloss	n
1	*menapak_masa_depan*	‘to step onto the future’	10
2	*menapak_ke_babak*	‘to step to the (X) stage (of a competition)’	6
3	*menapak_di_jalan*	‘to step at the street’	4
4	*menapak_di_lantai*	‘to step at the floor’	4
5	*menapak_karir_di*	‘to follow/get into a career at’	4
6	*menapak_di_atas*	‘to step on’	3
7	*menapak_tilas_jejak*	‘to follow the trace (of sth.)’	3
8	*menapak_di_bumi*	‘to step at the Earth’	2
9	*menapak_di_jalanan*	‘to step at the street’	2
10	*menapak_di_permukaan*	‘to step on the surface’	2

Menapak can occur in transitive (item 1, 5, and 7) and intransitive constructions (the remaining items in Table 6). Its transitive usage shares similar right-side patterns with *menapaki* (Table 7), especially their object with *masa depan* ‘future’ and *kari(e)r* ‘career’. Such collocates reflect their metaphorical usages and may partly motivate their merger.

[14] The English translation of the Indonesian definition of *menapakkan* in the fifth edition of the online *Kamus Besar Bahasa Indonesia* (KBBI) (https://kbbi.kemdikbud.go.id/entri/menapakkan).

Table 7. The ten most frequent 3-grams for *menapaki* 'to walk in; to set foot on'

	3-grams	gloss	n
1	*menapaki_anak_tangga*	‘to step on the stairs’	11
2	*menapaki_jalan_menuju*	‘to step on the street towards’	11
3	*menapaki_masa_depan*	‘to step onto the future’	11
4	*menapaki_jalan_yang*	‘to step on the street which’	9
5	*menapaki_karir_di*	‘to follow/get into a career at’	9
6	*menapaki_karier_sebagai*	‘to follow/get into a career as’	7
7	*menapaki_karier_di*	‘to follow/get into a career at’	6
8	*menapaki_babak_baru*	‘to step on/enter a new stage/phase’	5
9	*menapaki_dunia_kerja*	‘to step on/get into/enter working world’	5
10	*menapaki_karir_sebagai*	‘to follow/get into a career as’	5

This is different from the transitive usage with *meN-/-kan* affix (Table 8), which predominantly collocates with *kaki* ‘foot’ as its direct object, followed by either locational/directional prepositional phrases or motion verb complements (e.g., *memasuki* ‘to enter’ [item 5] and *maju* ‘to move forward’ [item 8]).

Table 8. The ten most frequent 3-grams for *menapakkan* 'to step on the sole of the foot'

	3-grams	gloss	n
1	*menapakkan_kakinya_di*	‘to step/set/put h(is/er) foot down at’	24
2	*menapakkan_kaki_di*	‘to step/set/put the foot down at’	14
3	*menapakkan_kaki_ke*	‘to step/set/put the foot down to(wards)’	3
4	*menapakkan_dirinya_di*	‘to set(tle) h(im/er)self down at’	2
5	*menapakkan_kaki_memasuki*	‘to step/set/put the foot down (while) entering’	2
6	*menapakkan_kaki_saat*	‘to step/set/put the foot down when’	2
7	*menapakkan_kakinya_ke*	‘to step/set/put h(is/er) foot down to(wards)’	2
8	*menapakkan_kakinya_maju*	‘to step/set/put the foot forwards’	2
9	*menapakkan_bisnis_toko*	‘to set(tle) down a business shop’	1
10	*menapakkan_citra_donnie*	‘to set(tle) down Donnie Yen’s image’	1

The remaining two semantic clusters form a mixture of PSYCH and COMMUNICATION verbs (Levin 1993: 188, 202). For instance, the cluster at the top of the dendrogram consists of four verbs (Figure 6 below), three of which are arguably PSYCH verbs: (i) *menyesal* ‘be/feel sorry/regret’ (N = 2,556), (ii) *menyesali* ‘regret/feel sorry for sth.’ (N = 956), and (iii) *membayangkan* ‘imagine; visualise’ (N = 2,719). The first two verbs with the same root *sesal* ‘regret’ are merged first, thus more similar in their distributional vectors, before later being merged with *membayangkan*. Final merge occurs between these verbs and *mengatai* ‘to rebuke; speak of one’s badness’ (N = 56), which is a COMMUNICATION verb.

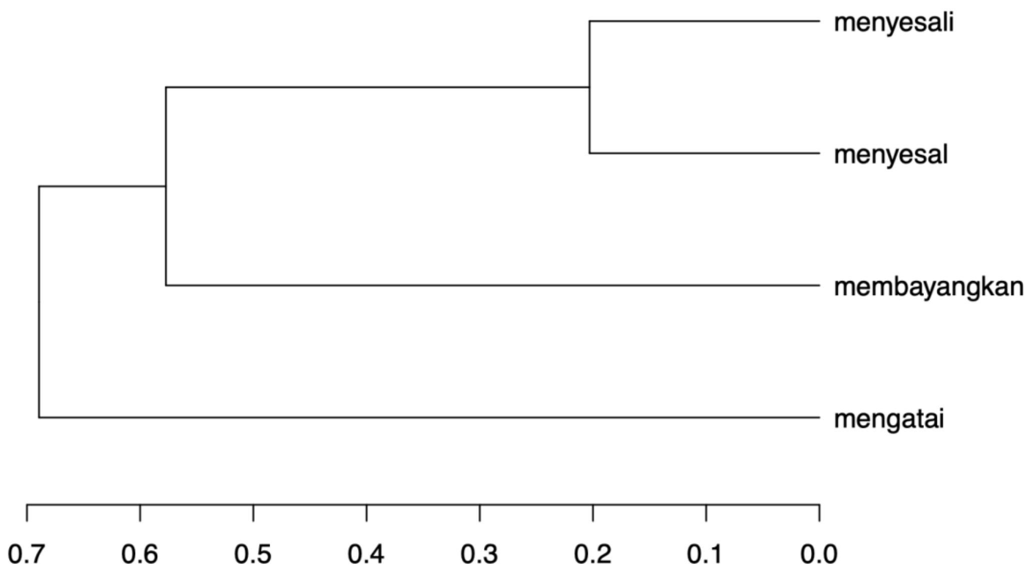


Figure 6. Cluster of PSYCH and COMMUNICATION verbs (a)

Similar semantic cluster can be found in the middle of the dendrogram (see Figure 7). They consist of *mengatakan* 'to say sth.' (N = 265,381), *mencontohkan* 'to exemplify' (N = 4,799), and *menyesalkan* 'regret, repent, resent sth.' (N = 1,976).

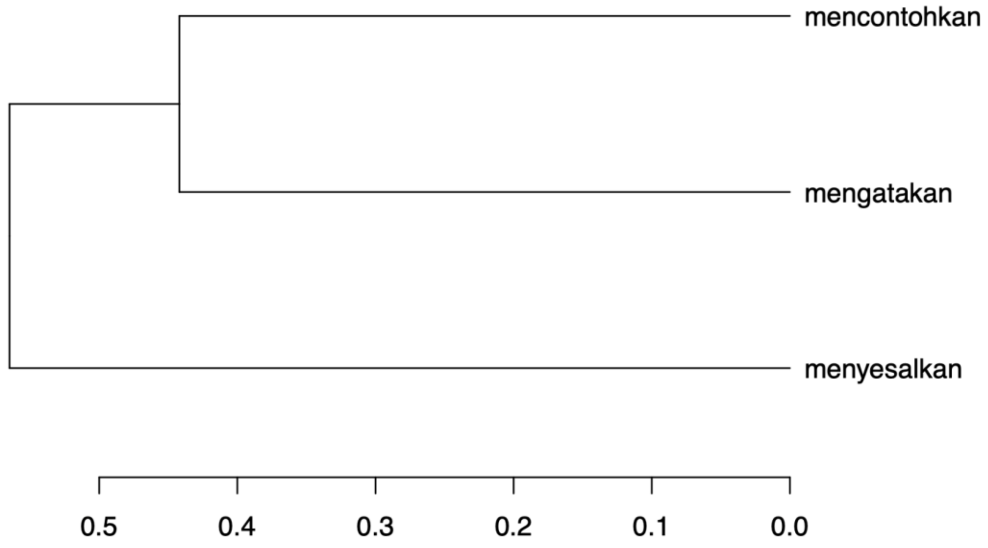


Figure 7. Cluster of PSYCH and COMMUNICATION verbs (b)

What is interesting about verbs in these last two clusters is that the *meN-/-i* (i.e. *mengatai*) and *meN-/-kan* verbs (i.e. *mengatakan*) with the root *kata* 'word' are way apart in the dendrogram, suggesting their distinct co-occurrences. Similar case is apparent between *menyesalkan* separated with *menyesal* and *menyesali,* where the latter two verbs cluster together and are merged first in Figure 6 (see §5.3 for further discussion on such splits).

5.2. Root-based clustering

We have discussed semantic clusters of morphologically heterogenous verbs, but the predominant clustering for most of the verbs is root-based clustering. That is, morphologically homogenous verbs (i.e. derivational family of the same noun roots) occurring in the three morphological patterns form their own clusters. We have seen few examples of these in the MOTION cluster with the root *tapak* ‘sole of the foot’ and *jejak* ‘footprint’, the derived forms of which fall into one cluster but differ in their within-cluster branching (Figure 5). The other examples are as follows:

d. *susu* ‘milk’ → *menyusu* ‘(of a baby/young animal) to suckle’ (N = 458); *menyusui* ‘to breast-feed sb.’ (N = 2,538); *menyusukan* ‘to let sb. suckle; to breast-feed sb.’ (N = 35)

e. *dasar* ‘base’ → *mendasar* ‘basic; foundational’ (N = 4,571); *mendasari* ‘to underlie sth.’ (N = 1,365); *mendasarkan* ‘to base X (on Y)’ (N = 781)

f. *tempat* ‘place; location’ → *menempat* ‘to place/position (sth.) at’ (N = 27); *menempati* ‘to occupy; to reside in’ (N = 11,150); *menempatkan* ‘to put/place/position X at Y; to deploy’ (N = 11,513)

g. *wakil* ‘vice; representative; deputy’ → *mewakil*[15] ‘to (be a) represent(ative of) X’ (N = 39); *mewakili* ‘to (be a) represent(ative of) X’ (N = 12,389); *mewakilkan* ‘to assign sb. as a representative’ (N = 220)

The cluster subsets of these verbs are extracted from Figure 4 into Figure 8 below.

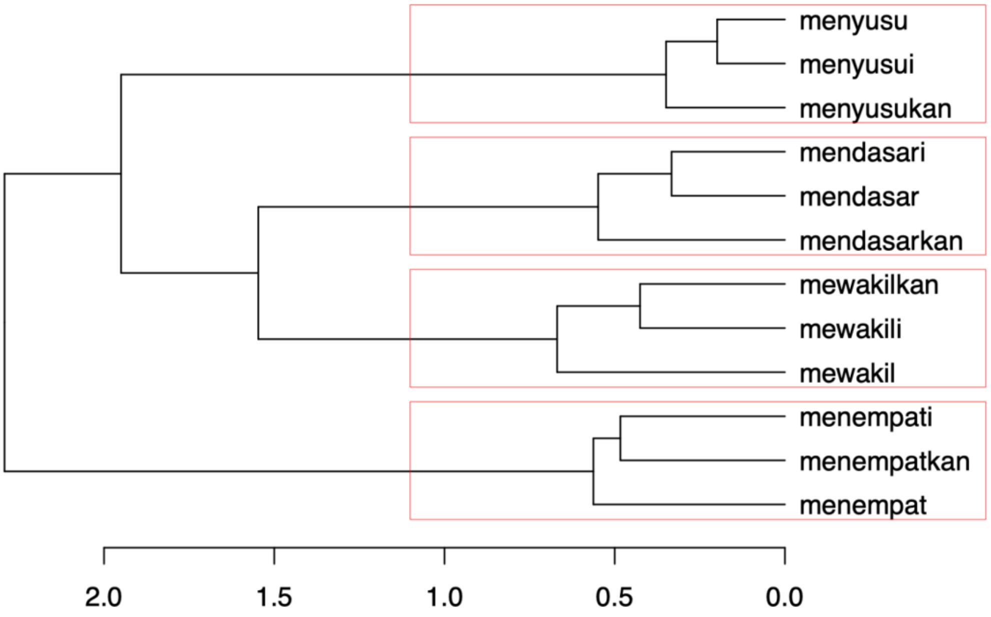


Figure 8. Root-based clusters of the denominal verbs

We assume that verbs in each of these clusters convey similar meanings (i.e. co-occurrence patterns) within the same semantic domain that are distinct from verbs in the other clusters. For instance, verbs with the root *wakil* ‘representative; vice’ convey the ‘delegation/representative’ sense. This is clearly different from the semantic cluster for verbs with the root *susu* ‘milk’, conveying ‘breastfeeding’ sense, or from those with the root *dasar* ‘base’, evoking the sense of ‘to base/underlie’.

Despite the clustering of the three morphological patterns for the verbs in (d) to (g), within-cluster differences exist between them, as shown by the branch merger in the

[15] *Mewakil* does not exist in the fifth edition of the online *KBBI*; try searching it via this link: https://kbbi.kemdikbud.go.id/entri/mewakil. *Mewakil* is similar to *mengakhir* and *menapak* in that *mewakil* syntactically behave like the *meN-/-i* forms of the corresponding roots.

cluster. One potential difference[16] is in the argument structures, such as between the *meN-/-i* and *meN-/-kan* verbs within the same cluster. Our VSM-based cluster analysis is an initial step in highlighting which *meN-/-i* and *meN-/-kan* verbs for a given root cluster together given their similar co-occurrence distribution. As in §5.1, one may inspect the right-side *n*-grams for each verb for quick overview of its argument-structure patterns.

Table 9. The ten most frequent 3-grams for *mewakili* 'to (be a) represent(ative of) X'

	3-grams	gloss	n
1	*mewakili_kebijakan_editorial*	‘to represent editorial’s policy’	174
2	*mewakili_indonesia_di*	‘to represent Indonesia at’	142
3	*mewakili_indonesia_dalam*	‘to represent Indonesia in’	129
4	*mewakili_indonesia_pada*	‘to represent Indonesia at’	69
5	*mewakili_lebih_dari*	‘to represent more than’	38
6	*mewakili_iklan_anda*	‘to represent your advertisement’	36
7	*mewakili_kepala_dinas*	‘to represent head of department’	35
8	*mewakili_indonesia_untuk*	‘to represent Indonesia for’	32
9	*mewakili_indonesia_ke*	‘to represent Indonesia to’	26
10	*mewakili_kepala_badan*	‘to represent the head of X institution’	23

The 3-grams for *mewakili* shows that the direct object is linked to the represented entity role while the subject is linked to the representing/representative role (e.g., *dia* ***mewakili*** *saya* ‘(s)he *represents* me’). Table 10. lists the 3-grams for *mewakilkan*.

Table 10. The ten most frequent 3-grams for *mewakilkan* 'to appoint/send X as the representative of Y'

	3-grams	gloss	n
1	*mewakilkan_sebuah_film*	‘to send a movie (as a representative)’	6
2	*mewakilkan_kepada_orang*	‘to delegate to somebody’	5
3	*mewakilkan_orang_lain*	‘to delegate other person (cf. (6))’	4
4	*mewakilkan_benua_asia*	‘to represent the Asian continent’ (cf. (7))	3
5	*mewakilkan_kehadirannya_kepada*	‘to delegate h(is/er) presence to’	3
6	*mewakilkan_kepada_unais*	‘to delegate to UNAIS’	3
7	*mewakilkan_kepada_wakil*	‘to delegate to vice-head’	3
8	*mewakilkan_6_perwakilan*	‘to send six delegations’	2
9	*mewakilkan_bisa_dengan*	‘to delegate can be with’	2
10	*mewakilkan_dirinya_lewat*	‘to make h(im/er)self represented via’	2

Mewakilkan in contrast conveys the sense of an entity (i.e. the agent) makes/sends another entity (i.e. the representing entity) to be the representative of certain (represented) entity

[16] Another potential factor is the frequency of the verbs, which may affect the tokens and types of their contextual co-occurrences. All verbs that are merged last in each cluster in Figure 8 (i.e. *menyusukan, mendasarkan, mewakil,* and *menempat*) are the least frequent compared to other verbs in their cluster.

(5) or events (6). The role for the representing entity is predominantly mapped onto prepositional oblique with *kepada* 'to', as can be seen in Table 10. .

(5)

[*Presiden*]$_{agent}$	*yang*	*berhalangan*	*hadir*	***mewakilkan***
president	REL	be.unable.to	be.present	delegate

[*kehadirannya*]$_{represented\ entity}$	*kepada*	[*para*	*menteri*]$_{representing\ entity}$.
presence.3PL	towards	DEM.PL	minister

'Presidents who are not present *delegate* their presence to their ministers.'
(ind_newscrawl_2012_1M:151843)

(6)

maka	[*dia*]$_{agent}$	*tidak*	*boleh*	***mewakilkan***	[*orang*	*lain*]$_{representing\ entity}$
thus	3SG	NEG	may	delegate	person	other

[*untuk*	*menggantikan*	*hajinya*]$_{represented\ event}$.
for	replace	pilgrim.to.Mecca

'thus (s)he cannot *delegate/send* other person to replace h(er/im) for h(er/im) pilgrim to Mecca'
(ind-id_web_2015_3M:1768782)

However, one pattern for *mewakilkan* (item 4 in Table 10.) appears to be used in analogy to *mewakili* in that the subject maps onto the representing entity for the object:

(7)

Dengan	*keindahan*	*kota*	*dan*	*juga*	[*suasana*]$_{representing\ entity}$	*yang*
with	beauty	city	and	also	condition	REL

benar-benar	***mewakilkan***	[*benua*	*Asia*	*ini*]$_{represented\ entity}$,	*tentu*
truly	represent	continent	Asia	DEM	surely

bisa	*membuat*	*para*	*turis*	*asing*	*betah.*
can	make	DEM.PL	tourist	foreign	feel.at.home

'With the beauty of the city and also the condition that truly *represent* this Asian continent, surely (they) can make the tourists feel at home.'
(ind-id_web_2013_1M:209)

The argument structure for Indonesian denominal verbs so far receives little attention compared to verbs with verbal and adjectival roots (Arka et al. 2009: 1).

The root-based clustering for the three morphological patterns is different from cases where one of the derivatives for a given root gets split from the rest, such as the separation between *melangkahi* 'to step over' and *melangkahkan* 'to move the foot forward' as well as *melangkah* 'to step; to move'. The next sub-section considers such split in more detail.

5.3. Clustering-split between derivational family for a given noun root

The split, especially between *meN-/-kan* and *meN-/-i* verbs, reflects the distinctive hypothesis concerning clear semantic distinction between some pairs of *meN-/-kan* and *meN-/-i* verbs (§1.1). Our VSM approach captures such split through the cluster dendrogram based on the verbs' semantic distances derived from their words co-occurrences patterns. In this section, we will demonstrate the technique called *nearest neighbours* or *closest words* to enrich the characterisation of such difference. The closest words are retrieved from the VSM table in R via the `closest_to()` function from `wordVectors` package (Schmidt & Li 2017).

Recall that the *skip-gram* algorithm of *word2vec* attempts to learn the contextual environments for a target word (§4.3). The idea of *nearest neighbours* is to retrieve a set of words with similar vector distribution (i.e. contextual environments) to a given target word on the basis of their cosine similarities. The higher their cosine similarities, the more similar they are, thus the *closer* they are in the *semantic space*.

To illustrate, consider the split between the transitive *melangkahkan* 'to move the foot forward' (*meN-/-kan* pattern) and *melangkahi* 'to step over' (*meN-/-i*). Table 11 shows the nearest neighbours to *melangkahi* 'to step over'.

Table 11. The ten closest words to *melangkahi* 'to step over'

	word	gloss	similarity to "*melangkahi*"
1	*mengangkangi*	'to straddle'	0.5508479
2	*berkeras*	'to be obstinate; persist; insist'	0.5435145
3	*memperhitungkannya*	'to take X into account'	0.5337096
4	*mengacuhkan*	'to heed, care about'	0.5174865
5	*memagari*	'to fence in; protect; demarcate'	0.5163426
6	*memegang*	'to grab hold of'	0.5040103
7	*membelakangi*	'to turn one's back on; disregard'	0.5036757
8	*mematuhi*	'to obey'	0.4959915
9	*bersikeras*	'to be obstinate; persist; insist'	0.4949589
10	*berbenturan*	'to collide; in conflict (with X)'	0.4929108

The closest words may not necessarily similar in meaning (e.g., near-synonyms), but may exhibit different kind of relationships, such as antonyms or even morphosyntactic (see further below). Words conveying more or less antonymous sense to *melangkahi* 'to step over; to disregard' include *mengacuhkan* 'to care about/heed sth.', *mematuhi* 'to obey', and arguably *memperhitungkannya* 'to take it into account'. *Mengangkangi* 'to straddle sth.' is the closest one in meaning with *melangkahi* as it can be extended into 'disregarding' sense from its physical, posture sense. Informal inspection on the 2-gram data for *mengangkangi* across the entire corpus shows that it can occur with rules-related objects, such as *hukum* 'law' (3 tokens), *peraturan* 'regulation' (3), *kebenaran* 'the truth' (2), *prinsip* 'principles' (2), *undang-undang* 'constitution' (2), and *aturan* 'rules' (1), among others. The potential reason for why one may disregard certain rules or principles could be due to not finding common ground, hitting a dead-end in negotiation. This idea might motivate the presence of *berbenturan* 'collide (with each other)' (item 10 in Table 11), which is commonly used in the context of disagreement.

Should one inspect the *n*-gram data for *melangkahi* (Table 12), its predominant sense of 'disregarding; violating' is reflected by the high frequency of the right-side collocates evoking (i) rules/protocols (i.e. *aturan* 'rules', *batas-batas*/*batasan* 'limits; restriction', *mekanisme* 'mechanism'), (ii) foundation (*dasar-dasar*), or (iii) authority (*kewenangan*). Yet, the nearest neighbours as in Table 11 scale-up nuances offered by VSM for data-driven lexical semantics, as they capture, for instance, antonymous and synonymous concepts for the semantics of *melangkahi*.

Table 12. The ten most frequent 2-grams for *melangkahi* 'to step over'

	2-grams	gloss	n
1	*melangkahi_kewenangan*	‘to transgress authority’	7
2	*melangkahi_aturan*	‘to transgress rules’	6
3	*melangkahi_batas-batas*	‘to transgress limits’	4
4	*melangkahi_apa*	‘to transgress what’	3
5	*melangkahi_beberapa*	‘to transgress few’	3
6	*melangkahi_mekanisme*	‘to transgress mechanism’	3
7	*melangkahi_pundak*	‘to step over a shoulder’	3
8	*melangkahi_tlundak*	‘to step over a door step’ (see (11) below)	3
9	*melangkahi_batasan*	‘to transgress boundary’	2
10	*melangkahi_dasar-dasar*	‘to transgress principles’	2

In contrast, *melangkahkan* ‘to move the foot forward’ appears to have similar distribution mostly with other MOTION verbs (Table 13 below), especially those with which *melangkahkan* clusters together in Figure 5, namely *menjejakkan*, *melangkah*, and *menapakkan*. This fact further supports their clustering and separation with *melangkahi*.

Table 13. The ten closest words to *melangkahkan* 'to move the foot forward'

	word	gloss	similarity to “*melangkahkan*”
1	*menjejakkan*	‘to step sth. on’	0.7487976
2	*melangkah*	‘to stride; move on; take a step’	0.7351338
3	*dilangkahkan*	‘(one’s foot) to be moved forward’	0.7294263
4	*berlari*	‘to run’	0.7258124
5	*kakiku*	‘my foot’	0.7195594
6	*menghunjamkan*	‘to make sth. dive down; to stab into’	0.7150664
7	*menapakkan*	‘to step on the foot’	0.7126413
8	*berjingkat*	‘to stand on tiptoe’	0.7079603
9	*kakinya*	‘h(is/er) foot’	0.7068803
10	*langkahkan*	‘move the foot forward’	0.7066612

Note that nearest neighbours may also capture words with syntagmatic and morphological relationships. The former is indicated by the word *kakiku* ‘my foot’ and *kakinya* ‘h(is/er) foot’ in Table 13. *Kaki* ‘foot’ with its possessive inflection is in the top-3 most frequent direct-object collocate in 2-gram patterns for *melangkahkan*: (i) *melangkahkan_kaki* ‘to move the foot’ (N = 162), (ii) *melangkahkan_kakinya* ‘to move h(is/er) foot’ (N = 123), and (iii) *melangkahkan_kakiku* ‘to move my foot’ (N = 13). Morphological relationship can be seen from the *di-* passive of *melangkahkan* and its bare form *langkahkan* (particularly used in imperative clause). This suggests that the active *meN-*, passive *di-*, and the bare (imperative) forms with the root *langkah* have similar co-occurrence distribution. We have started here with cluster analysis based on one set of the morphological patterns with a given root. Extending that analysis via nearest neighbours has brought us to another set of morphologically related patterns with the same root, suggesting that these patterns could be the input for another cluster analysis.

Despite the frequent co-occurrence of *melangkahkan* with *kaki*, this pattern can appear in the context of concrete physical motion (8) and figurative senses (see (9) and (10)):

(8)	*Perlahan-lahan*	*ia*	***melangkahkan***	*kakinya*	*di*	*atas*	*jalan*
	slowly	3SG	step; move	foot.3SG.POSS	LOC	on	street

yang	*berbatu-batu.*
REL	rocky

‘(S)he slowly *moves/steps* (on h[is/er] foot) onto the rocky street.’ (ind_web_2011_300K:50903)

(9)	*pelantun*	*Matahariku*	*itu ...*	*sudah*	***melangkahkan***	*kakinya*
	singer	sun.1SG.POSS	DEM	already	step; move	foot.3SG.POSS

di	*blantika*	*musik*	*Internasional.*
LOC	industry (of entertainment)	music	international

‘The singer of the song titled *Matahariku* has *entered/set* (h[is/er] foot) *into* international music industry.’ (ind-id_web_2015_3M:2570936)

(10)	*Ia*	*tidak*	*pernah*	*lagi*	*berdoa*	*atau*
	3SG	NEG	ever	again	pray	or

melangkahkan	*kakinya*	*ke*	*dalam*	*gereja.*
step; move	foot.3SG.POSS	to	inside	church

‘(S)he never again prays or *set* h(is/er) foot at the church (i.e. it metonymically refers to ‘come to’ the church)’ (ind-id_web_2015_3M:972360)

Melangkahi also appears with both senses. Despite its predominant metaphorical sense of ‘disregarding; violating’, its concrete translational motion does not bleach (11), such as when co-occurring with *tlundak* ‘stepping stone’ as its object (item 8 in Table 12).

(11)	*kedua*	*orang*	*itu*	*bersama-sama*	***melangkahi***	*tlundak*	*pintu*	*masuk*
	both	person	DEM	together	step over	steps	door	enter

ke	*dalam*	*gubug*	*itu*	*pula.*
to	inside	hut	DEM	also

‘The two/both of the persons together *step over* the door steps entering the hut as well.’ (ind_web_2011_300K:71549)

The next interesting split in Figure 4 is between *mengatai* ‘to rebuke; insult; speak of one’s badness’ (N = 56) and *mengatakan* ‘to say sth.’ (N = 265,381) based on *kata* ‘word’. The closest words for *mengatai* (Table 14) and *mengatakan* (Table 15) clearly show that these verbs capture different facets of verbal activity, corroborating their large distance in Figure 4. *Mengatai* is associated with abusive and emotional (verbal) behaviour.

Table 14. The ten closest words to *mengatai* 'to rebuke; speak of one's badness'

	word	gloss	similarity to "*mengatai*"
1	*memaki*	'to use abusive language to s.o.'	0.7476359
2	*marah-marah*	'to keep on being angry'	0.6818458
3	*cerewet*	'nagging; to talk too much'	0.6806308
4	*mengejek*	'to mock; ridicule'	0.6795740
5	*memaki-maki*	'to use abusive language to s.o.'	0.6713904
6	*jengkel*	'annoyed; irritated'	0.6675079
7	*diejek*	'to be mocked; ridiculed'	0.6645705
8	*diolok-olok*	'to be mocked; derided; jeered at'	0.6641099
9	*meledek*	'to make fun of; nag; mock'	0.6628723
10	*berbohong*	'to lie'	0.6597248

In contrast, *mengatakan* mostly appears as communication verb with similar vector distribution to other reported speech verbs, mostly in the *meN-/-kan* affix.

Table 15. The ten closest words to *mengatakan* 'to say sth.'

	word	gloss	similarity to "mengatakan"
1	*menegaskan*	'to assert; affirm'	0.8376837
2	*menyatakan*	'to state; indicate'	0.8318030
3	*mengungkapkan*	'to reveal; disclose'	0.8079668
4	*mengemukakan*	'to suggest; offer; utter'	0.7967164
5	*menuturkan*	'to say; narrate; tell about'	0.7925858
6	*menjelaskan*	'to explain; clarify'	0.7808896
7	*menyebutkan*	'to mention'	0.7640667
8	*menerangkan*	'to explain; clarify'	0.7568715
9	*mengakui*	'to admit'	0.7495804
10	*mengatkan*[17]	'to say sth.'	0.7467988

Another clear case of split between *meN-/-kan* and *meN-/-i* verbs of the same noun root is based on *buah* 'fruit' in *membuahi* 'to put seed at sth.; to fertilise' (N = 116) and *membuahkan* 'to produce fruit; to cause/bring forth' (N = 3,338). Their semantic difference is robustly captured by their closest words (see Table 16 and Table 17).

[17] This is a misspelling for *mengatakan* 'to say sth.'.

Table 16. The ten closest words to *membuahi* 'to fertilise'

	word	gloss	similarity to “*membuahi*”
1	*dibuahi*	‘to be fertilised’	0.8372309
2	*ovum*	‘ovum’	0.8330801
3	*sperma*	‘sperm’	0.7965994
4	*gamet*	‘gamete; reproductive cells’	0.7326975
5	*pembuahan*	‘fertilisation; ovulation’	0.7300101
6	*terbuahi*	‘to be fertilised’	0.7016097
7	*spermatozoid*	‘spermatozoid’	0.6901719
8	*spermatozoa*	‘spermatozoa’	0.6794663
9	*parthenogenesis*	‘parthenogenesis’	0.6775866
10	*zigot*	‘zygote’	0.6712285

The list in Table 16 is interesting. First, most of the closest words include biological terminologies, especially in the domain of fertilisation, such as *ovum* ‘ovum’, *sperma* ‘sperm’, *spermatozoid* ‘spermatozoa’. These words strongly support the idiosyncratic meaning of *membuahi* in the domain of fertilisation. This idiosyncrasy is evidenced in the entry of *membuahi* in *Kamus Besar Bahasa Indonesia* (KBBI) that is exclusively defined in the biology domain.

Secondly, both the dynamic and static passive forms (Arka 2010) of *membuahi*, namely *dibuahi* ‘to be fertilised’ and *terbuahi* ‘to be fertilised’ respectively, appear to share similar co-occurrence with *membuahi*. This is also the case for the nominalisation with *pe- -an* affix (i.e. *pembuahan*) that is lexicalised in the domain of fertilisation (item 5). *Pembuahan* is not the nominalised form for the meaning conveyed by *membuahkan* ‘to bear a fruit; to bring forth’ and VSM allows us to determine this through cosine similarity. This argument is further supported by Table 17 in which *pembuahan* is absent. Our analysis of *membuahi* corroborates our finding on *melangkahkan* ‘to move foot forward’ (Table 13) that retrieving a word’s nearest neighbours in semantic space is an efficient way to test if the other set of morphological patterns of the given word conveying a certain semantic concept (e.g., fertilisation) has similar semantic vectors, suggesting lexical diversity of the semantic concept.

Table 17. The ten closest words to *membuahkan* 'to bear a fruit; to bring forth'

	word	gloss	similarity to "*membuahkan*"
1	*berbuah*	'to bear a fruit; produce; result in'	0.6716316
2	*mem-buahkan*	'to bear a fruit; produce; bring forth'	0.6296921
3	*tercipta*	'to be created'	0.6214626
4	*membuah*	'to bear a fruit; produce; bring forth'	0.5991543
5	*menuai*	'to harvest; reap'	0.5729809
6	*tendangannya*	'h(is/er) shot/kick'	0.5533927
7	*ditepis*	'to be parried; warded off'	0.5530693
8	*kerasnya*	'the strength'	0.5528045
9	*pinalti*	'penalty (kick)'	0.5482891
10	*dimentahkan*	'to be foiled'	0.5478375

For *membuahkan*, a number of interesting observations can be made. First, *membuah* 'to bear fruit; to result in; to bring forth' appears amongst the top-10 closest words for *membuahkan*. The form *membuah* appears in total of 56 citations across the thirteen corpus files. Manual inspection on these citations reveals that 24 occurrences[18] are used in transitive construction in analogy to *membuahkan*, evoking a causative sense 'to bring forth' (compare (12) and (13)). This partly explains why in the dendrogram in Figure 4 *membuah* clusters together with *membuahkan*, but not with *membuahi*.

(12) *Peluang Irak lewat tendangan Younis Khalef di dalam*
chance Iraq pass kick NAME LOC inside

kotak penalti juga masih belum ***membuah*** *hasil.*
box penalty also still not.yet bear.fruit results

'The chance for Iraq through Younis Khalef's shot inside the penalty box still has not *produced* a result.' (ind_mixed_2012_1M:912835)

(13) *menekan pertahanan Persija ... sehingga* ***membuahkan***
press defence NAME so.that bear.fruit

tendangan pojok beberapa kali.
kick corner several times

'putting pressure to Persija's defence … so that *producing/bringing forth* a number of corner kicks.' (ind_news_2009_300K:591)

Secondly, similarity between *membuah* and *membuahkan* leads us to hypothesise that *membuah* can be a case of backformation in analogy to *membuahkan.* This is evident from (i) the vastly lower frequency of *membuah* than *membuahkan* and (ii) the similar

[18] The remaining citations of *membuah* include (i) mispelling for *membuat* 'to make' (25 citations) as used in periphrastic causative construction; (ii) mispelling for *membuang* 'to throw' (5 citations), especially to throw trash; (iii) split part of *membuahkan* (1 citation); and (iv) one unclear case in context as to whether it refers to the transitive *membuah* or *membuat* 'to make'.

syntactic behaviour of *membuah* in transitive construction. *Membuah* also appears to have similar usage pattern as the intransitive *berbuah*, which is the closest word to *membuahkan* and has much higher token frequency than *membuah* (i.e. $N_{berbuah}$ = 2,356 vs. $N_{membuah}$ = 56). *Berbuah* can take nominal (14) and adjectival (15) complements associated with the result role of the causation event (as in the object of *membuahkan* and *membuah*) while the subjects are mapped onto the cause role:

(14)	*Hal*	*ini*	***berbuah***	[*skandal*]$_{\text{nom comp}}$	*dalam*	*keluarganya.*
	matter	DEM	bear.fruit	scandal	inside	family.3SG

‘This matter *results in* scandal in h(is/er) family.’
(ind_mixed_2012_1M:847228)

(15)	*Keseriusan*	*Dewi*	*dalam*	*kajian*	*sosiologis ...*	***berbuah***	[*manis*]$_{\text{adj comp}}$.
	seriousness	NAME	inside	study	sociology	bear.fruit	sweet

‘Dewi’s studiousness in the study of sociology … *produces* great results.’
(ind_mixed_2012_1M:908048)

This result further illustrates that finding nearest neighbours via VSM allows us to capture morphosyntactic variation of a certain semantic concept. That is, *membuahkan*, *berbuah*, and *membuah* convey the same generic meaning of causation via different morphosyntax; similar relationship exists for the concept of fertilisation manifested in *membuahi, dibuahi, terbuahi,* and *pembuahan*. Furthermore, manual inspection on all citations for such rare cases as *membuah* reveals to what extent it reflects orthographical relics as well as insights on new phenomenon, such as potential backformation based on analogy.

Another interesting point to make is the topical domains closest to *membuahkan* include terms common in Indonesian football/soccer commentary, such as *tendangannya* ‘kick; shot’, *ditepis* ‘to be parried; warded off’, *dimentahkan* ‘to be foiled’, *pinalti* ‘penalty kick’. The causative semantics and topical domain of *membuahkan* are set off from the other derived forms, namely *membuahi* (and its passive variants) and *pembuahan*, having specific semantics in the domain of fertilisation (Table 16).

The last example of affix split to be discussed is between *mengakhiri* ‘to end sth.’ (N = 8,512) and *mengakhirkan* ‘to put sth. at the end/back’ (N = 116) based on the root *akhir* ‘end’. The latter forms a one-member cluster (as for *membuahi* ‘to fertilise’) in Figure 4.

Table 18 reiterates the fact that extended, nearest-neighbours analysis of a morphological pattern with a given root (*mengakhiri*) may lead to another set of morphologically related forms with the same root (i.e. *mengakhir* and *berakhir*), in addition to other near-synonyms (*memupus, menyudahi*) and antonym (*memperpanjang*) of *mengakhiri*. The case of backformation can be assumed for *mengakhir* (N = 57) (cf. (17) below), the closest word to *mengakhiri*. In Figure 4, these verbs cluster together and separated from the *meN-/-kan* form. Manual inspection for all usage sentences for *mengakhir* reveals that it occurs in full form, that is, none of them represents orthographical split from *mengakhiri* or *mengakhirkan*. What is more interesting is that all tokens of *mengakhir* are in transitive constructions analogous to *mengakhiri*. Notice for instance the similar reference of direct objects of the verbs in (16) and (17).

Table 18. The ten closest words to *mengakhiri* 'to end sth.'

	word	Gloss	similarity to “*mengakhiri*”
1	*mengakhir*	‘to put sth. to an end’	0.7149269
2	*menyudahi*	‘to end; terminate; conclude’	0.6486199
3	*mengahiri*[19]	‘to put sth. to an end; to end’	0.6346930
4	*akhiri*	‘to put sth. to an end; to end’	0.6329971
5	*berakhir*	‘sth. ends’	0.6263411
6	*mengakhirinya*	‘to end it’	0.6060989
7	*mengakiri*	‘to put sth. to an end; to end’	0.5811454
8	*memimpin*	‘to lead’	0.5763634
9	*memperpanjang*	‘to extend’	0.5681369
10	*memupus*	‘to wipe out; obliterate; exterminate’	0.5669336

(16)

Presiden	*Yudhoyono*	*berharap*	*bisa*	***mengakhiri***		
president	NAME	hope	can	put.to.an.end		
masa	*jabatannya*	*hingga*	*tahun*	*2014*	*dengan*	*baik*
period	position	until	year	YEAR	with	good

‘President Yudhoyono hopes that he can *end* his term of office well until 2014’ (ind_news_2011_300K:20587)

(17)

Gubernur	*mengatakan,*	*dirinya*	*akan*	*segera*	***mengakhir***
governor	say	self.3SG	FUT	soon	put.to.an.end
masa	*tugas*	*bersama*	*wakil*	*gubernur*	
period	assignment	together	vice	governor	

‘The Governor said that s(he) h(im/er)self will soon *end* h(is/er) term of office together with the Vice Governor’ (ind_news_2011_300K:117663)

Similar usages between *mengakhiri* and *mengakhir*, and the much lower frequency of the latter, lead us to assume the backformation status of the latter. Note that KBBI has entry for *mengakhir* with similar meaning to *mengakhiri.* Our corpus-based study with VSM and manual inspection captures not only their similar usages but also the asymmetry in their frequency. The clear split of the previous two words with *mengakhirkan* may be due to the highly restricted thematic domain of *mengakhirkan* for Islam (see Table 19).

[19] *Mengahiri* (item 3) and *mengakiri* (item 7) are misspelling for *mengakhiri* ‘to end’.

Table 19. The ten closest words to *mengakhirkan* 'to put sth. at the end; to postpone sth.'

	word	gloss	similarity to "*mengakhirkan*"
1	*shalat*	'Moslem's ritual prayer'	0.8595279
2	*menjamak*	'to combine two prayers into a single prayer'	0.8582150
3	*shubuh*	'dawn (prayer)'	0.8510190
4	*dijamak*	'two prayers to be combined into a single prayer'	0.8504550
5	*zhuhur*	'noon (prayer)'	0.8434480
6	*qashar*	'Moslem's prayer whose *rakaat* is shortened'	0.8383760
7	*isya*	'early evening (prayer)'	0.8300283
8	*disunnahkan*	'to be recommended (according to Islamic law)'	0.8250431
9	*menjama*	'to combine two prayers into a single prayer'	0.8227094
10	*qabliyah*	'a kind of Moslem's prayer/*shalat*'	0.8215906

This section provides evidence for the distinctive hypothesis (§1.1) that some pair of verbs with the same root but of two different morphological patterns differentiate each other. Our VSM approach leverages large corpus data to capture the distributional differences through closest words technique and visualisation as in Figure 4.

It is worth noting that in the dendrogram for all 21 clusters (Figure 4), there are four clusters that contain only a single verb and that three of these consist of a suffixed verb: *mengantungkan*, *membuahi* and *mengakhirkan*. The fourth single word cluster consists of *menanda*. Both *mengantung* and *menanda* are problematic for reasons to do with orthography (see discussion in §5.4), but it is notable that the other two examples (*membuahi* and *mengakhirkan*) are cases where one of the three member derivational family is separated from the other two forms in the cluster analysis.

5.4. Issues and challenges

In this section, we discuss problems in the input data which can affect the results obtained in parsing with *MorphInd* and, subsequently might affect our analysis. We will examine two closely related problems here: spelling variation (or error if one prefers) and orthographical relics.

Regarding spelling variation, we noticed it in the words *mengantung*, *mengantungi*, and *mengantungkan*. In Figure 4, these three words are all split up. *MorphInd* parsed them as based on the root *kantung* 'pocket', which is a pronunciation variant of *kantong* 'pocket'. Given this parsing, and from the native speaker intuitions of the first and the second authors, we assume that these words might conflate forms based on two different roots: the verb *gantung* 'to hang' and the noun *kantung* 'pocket'. The former is reflected in *mengantung* and *mengantungkan*, which should be spelled with double *g*, thus *menggantung* 'to hang' and *menggantungkan* 'to hang sth. (onto sth.)'. The latter is in correct form as *mengantungi* 'to pocket sth.; to gain sth.', with the nasal of the prefix being assimilated to the velar sound at the start of the base. We can test our assumption by pulling out the nearest neighbours for each of these problematic words. Table 20 reveals that that the nearest neighbour of *mengantung* is *men**gg**antung* with double *g*.

Table 20. The ten closest words to *mengantung* 'to hang'

	word	gloss	similarity to “*mengantung*”
1	*menggantung*	‘to hang’	0.6580294
2	*blandar*	‘a kind of wood as part of a house construction’	0.6541215
3	*memaku*	‘to nail’	0.6125143
4	*ditelungkupkan*	‘to be turned over (faced down)’	0.5888941
5	*dibaringkan*	‘to lay sth. down’	0.5863362
6	*loso*[20]	‘a person’s name’	0.5808274
7	*digergaji*	‘to be sawed/sawn’	0.5766285
8	*tertelungkup*	‘to be turned over (faced down)’	0.5721740
9	*tersekap*	‘to be trapped; imprisoned’	0.5694923
10	*ditindih*	‘to be pressed down’	0.5677198

This may confirm our intuition that *mengantung* (N = 28) could be misspelling, or perhaps spelling variation, for the considerably more frequent form *menggantung* (N = 1,100). Presumably, such spelling variation may be influenced from the way *menggantung* is pronounced in (fast) speech. Similar evidence is found for *mengantungkan* (N = 44), showing *menggantungkan* (N = 1,264) as the closest word (Table 21).

Table 21. The ten closest words to *mengantungkan* 'to hang sth. (onto sth.)'

	word	gloss	similarity to “*mengantungkan*”
1	*menggantungkan*	‘to hang sth. (onto sth.)’	0.7764547
2	*penghidupannya*	‘h(is/er) sustenance; livelihood’	0.6517582
3	*peladang*	‘farmer; cultivator’	0.5890552
4	*matapencaharian*	‘livelihood’	0.5649209
5	*mengentaskannya*	‘to relieve (usually poverty)’	0.5516661
6	*bersawah*	‘to have/till rice fields’	0.5515007
7	*petani-petani*	‘farmers’	0.5469401
8	*pengais*	‘s.o. who scratches around for sth.’	0.5458413
9	*bertani*	‘to farm (for a living)’	0.5404030
10	*upahan*	‘wage earner; hired man’	0.5370267

The assumption is also confirmed for *mengantungi* (N = 327) as a variant of the more common form *mengantongi* (N = 3,574) from the root *kantong* ‘pocket’ (Table 22). Note that no form with the double *g* appears in this list.

[20] Most usage tokens of the case-insensitive form *loso* (with whitespace in either side of it) in the thirteen Leipzig corpus files turn out to be a person’s name. Two tokens get split from the word *filosofis* (in the file *ind-id_web_2013_1M* sentence no. 172313 and 662198).

Table 22. The ten closest words to *mengantungi* 'to pocket sth.'

	word	gloss	similarity to “*mengantungi*”
1	*mengantongi*	‘to pocket; win (a medal)’	0.7021752
2	*mengoleksi*	‘to collect (stamps, unique goods, etc.)’	0.6954547
3	*memuncaki*	‘to reach a high/top point/position’	0.6904767
4	*raihan*	‘(what is) gotten; achieved’	0.6513341
5	*terpaut*	‘(X number) apart; separated by’	0.6435489
6	*mengungguli*	‘to surpass; do better than’	0.6411234
7	*torehan*	‘incision’[21]	0.6217388
8	*pemuncak*	‘the (top) leader; champion’	0.6164198
9	*mengemas*	‘to pack up; gather (a medal, point, etc.)’	0.6040049
10	*diposisi*[22]	‘at the position’	0.6004999

Finally, issue concerning orthographical relics is illustrated with *menanda*. The form occurs in total of 121 citations across the corpus files and 113 of these are split cases of *menandatangani* ‘to sign’ based on the compund root *tanda tangan* ‘signature; lit. hand sign’.[23] Nearest neighbour’s data show that *menanda* is indeed similar to *menandatangani* and its near-synonym, namely *meneken* ‘to sign’ (item 10). Other items in Table 23 represent (i) misspelling (e.g., items 4 and 5), (ii) passive *di-* form (e.g., item 7), or (iii) split case (i.e. item 6) from *menandatangani*, thus no glosses are given.

Table 23. The ten closest words to *menanda*

	word	similarity to “*menanda*”
1	*ditanda*	0.7982222
2	*menandatangani*	0.6721747
3	*menanda-tangani*	0.6597414
4	*menandatangai*	0.6295324
5	*menandatangi*	0.6148956
6	*tanganinya*	0.6112830
7	*ditandatangani*	0.6063122
8	*ditanda-tangani*	0.6057313
9	*ditandatanganinya*	0.5950392
10	*meneken*	0.5891363

Table 23 further indicates variation exists in the spelling of derived verbs with compound root in Indonesian (i.e. *menanda tangani*, *menandatangani* and *menanda-tangani*).

[21] Commonly used in the sense of ‘what is gotten’ in a tournament (e.g., Olympics, etc.).

[22] Misspelling for *di posisi*.

[23] The remaining citations of *menanda* include (i) five intransitive usages followed by prepositional oblique (e.g., *menanda pada X* ‘to leave a mark/sign at X’), (ii) two transitive usages (one in analogy to *menandai* ‘to mark sth.’ and one in analogy to *menandakan* ‘to indicate’), and (iii) one ambiguous as to whether it is intransitive or transitive usages.

The case of inconsistent splitting of compounds is tricky for computational tokenisation that considers a word-token to consist of one or more alphabetic characters (that may include hyphen) separated by whitespace. Thus, *menanda* and *tangani* in *menanda tangani* will be considered as two different tokens separated by whitespace. This result influences further computational processing on the tokens such as morphological parsing, that identifies *menanda* as based on the root *tanda* ‘sign’. This parsing is partly true because there are five out of 121 tokens for *menanda* used in intransitive construction (see examples below) that convey the sense of ‘to leave a mark/sign at sth.’.

(18)	*Beberapa*	*kerutan*	***menanda***	*pada*	*wajah.*
	several	wrinkles	mark	at	face

‘Several wrinkles *mark* (lit. leave their traces/marks) at the face.’ (ind_mixed_2012_1M:138307)

(19)	*Tulisan*	*itu*	***menanda***	*pada*	*sebuah*	*rumah*	*adat*
	writing	DEM	mark	at	ART	house	traditional

‘That writing *marks* at a traditional house’ (ind_mixed_2012_1M: 278193)

Intuitively, *menanda* sounds odd at the first blush. However, checking all its usages reveal its certain meaning in an intransitive construction. This case for *menanda* also points to another complementary challenge for our unsupervised approach, namely manual scrutiny on all, or a sample of, citations for a given verb to identify its syntactico-semantic patterns (e.g., transitivity and argument-structures), a desideratum for further research.

6. Conclusion

This paper set out to provide usage-based, quantitative corpus-based analyses on the semantics of derivational family of Indonesian denominal verbs with *meN-*, *meN-/-kan*, and *meN-/-i* affixes. The study is framed within, and aims to test, the two hypotheses by Sneddon et al. (2010) concerning the semantic (dis)similarity between *-i/-kan* verb pairs, the hypotheses that we call the *distinctive* and *similarity hypotheses* (cf. §1.1 and §2). Our analyses use both the original corpus data as well as a vector space model (VSM) trained on that data (§3, §4.3).

The VSM-based *Hierarchical Agglomerative Clustering* analysis captures both the distinctive and similarity hypotheses for *-kan/-i* verb pairs (§1.1) (see Figure 4). The split between *-kan* and *-i* pairs may reflect the distinctive hypothesis (§5.3) while the clustering of both *-kan* and *-i* pairs may reflect the similarity hypothesis (§5.2). The inclusion of the base *meN-* forms adds another layer of comparison. We showed that while most groups of morphologically homogenous verbs (i.e. of the same noun roots) clustered together, there were cases where one member of the derivational family was separated from the other two. These cases were examined further by looking at the words closest to them (i.e. *nearest neighbours* (§5.3)) in semantic space as represented by the VSM, and by looking at *n*-grams. Interestingly, the same method (i.e. nearest neighbours) also proved useful in tracking orthographic anomalies in the corpora (§5.4). Moreover, the clustering revealed several semantically coherent verbs from different roots (§5.1) (i.e. MOTION-related (Figure 5) and PSYCH/COMMUNICATION verbs (Figure 6 and Figure 7)).

Our investigation gives some explanation as to how forms with these three morphological affixes differ in their semantic distribution and therefore gives at least some support to the traditional view (as set out by for example Sneddon et al. (2010:100)) that *-i* and *-kan* derivatives have different semantics (§5.3). However, our investigation using 3-gram data focusing on the right-side patterns shows that differences in transitivity may be the key factor supporting difference in meaning (e.g., *melangkahi* vs. *melangkah* and *melangkahkan*; *menapakkan* vs. *menapak* and *menapaki*) (§5.1). Moreover, this difference in transitivity also appears with other noun bases (e.g., *susu* 'milk', *dasar* 'base', *tempat* 'place; location', *wakil* 'representative') (§5.2). We have used nearest neighbours to identify near-synonyms of the target words (e.g., *melangkahi* and *mengangkangi* vs. *melangkahkan* and *menapakkan*; *mengatai* and *memaki* vs. *mengatakan* and *menyatakan*) (§5.3), and this helped in some cases to discriminate between metaphorical and literal usages. We have also shown that in some cases one member of the morphological patterns does not cluster with the forms sharing the root but has developed a specialised meaning (§5.3). Thus, we found that in the case of *mengakhiri* and *mengakhirkan*; *mengakhirkan* is restricted to the thematic domain of Islam (Table 19), and that in the case of *membuahi* and *membuahkan*, *membuahi* is restricted to the thematic domain of biology (Table 16). It is well-known that words created by the processes of derivational morphology can develop idiosyncratic meanings (cf. Booij 2007: 57–58, 61) and these seem typical examples of such effects.

Our exploration of sets of words sharing one set of morphological relations also brought up evidence that other sets of Indonesian words derived from a single root look like promising areas for an analysis similar to that which we have presented here (cf. e.g., Table 16). We also hope to have offered a new perspective on investigating a decades-long issue in Indonesian linguistics, leveraging the availability of large corpora and the advance of computational, quantitative corpus linguistics. More generally, we suggest that combining VSM-based approaches and central corpus linguistic approaches, such as *n*-grams, can be a powerful research strategy for the usage-based study of language.

Abbreviations

1	first person	2	second person
3	third person	ART	article
ASW	Average Silhouette Width	AV	actor voice
CBOW	Continuous Bag of Words	DEM	demonstrative
DH	Distributional Hypothesis	FUT	future marker
HAC	Hierarchical Agglomerative Clustering	KBBI	Kamus Besar Bahasa Indonesia
LOC	locative	NEG	negator
PL	plural	POSS	possessive
REL	relativiser	SG	singular
VSM	Vector Space Model		

Appendix: The analysed denominal verbs and their senses

Word	Frequency	Senses	Notes
mengakhir	57	'to end sth.'	Analogous to *mengakhiri*
mengakhiri	8512	'to end sth.'	
mengakhirkan	116	'to put sth. at the end, i.e. to postpone'	This is restricted in the domain of Islam
membayang	199	'to shade, overshadow'; 'to shadow'	
membayangi	464	'to cast a shadow over'; 'to trail, follow s.o.'	
membayangkan	2719	'to imagine, visualise sth.'	

Word	Frequency	Senses	Notes
membuah	56	‘to cause/bring forth sth.’	
membuahi	116	‘to fertilise’	
membuahkan	3388	‘to cause/bring forth sth.’	
mencontoh	730	‘to imitate, follow’	
mencontohi	25	‘to imitate, follow’	21 out of 25 tokens have the same sense as *mencontoh* ‘to imitate, follow the direct object’; 1 token means ‘subject acts as the example *for* the direct object; to exemplify’; the remaining tokens are unclear.
mencontohkan	4799	‘to exemplify’	
mendasar	4571	‘basic; fundamental’	
mendasari	1365	‘to form the basis of sth.’	
mendasarkan	781	‘to base sth. on’	
mengguna	37	‘to use sth’ (17 tokens); usage analogy of *menggunakan*	The other 17 tokens are suffix-split from *menggunakan*; the remaining 3 tokens are misspelling for the deverbal noun *pengguna* ‘user’
menggunai	28	‘to put a curse/black magin on	
menggunakan	160721	‘to use, utilise sth.’	
menjejak	165	‘to step one’s foot on sth.; to trace/track (down) sth’	
menjejaki	48	‘to step (foot) on; to trace/track (down)’	
menjejakkan	243	‘to step sth. (usually foot) on’	
mengantung	28	‘to hang sth.’	Misspelling from *mengggantung* ‘to hang sth.’ (cf. §5.4)
mengantungi	327	‘to pocket sth; to win sth. (e.g., a medal)’	Spelling variation from *mengantongi* (cf. §5.4)
mengantungkan	44	‘to hang (up) sth. (onto sth.); to have sth. depend on’	Misspelling from *mengggantungkan* (cf. §5.4)
mengata	48	‘to say’	19 of the 48 tokens were split cases with *-kan* & 1 token is ambiguous; ‘to say’ sense is inferred from the 28 tokens of the the full form.

Word	Frequency	Senses	Notes
mengatai	56	'to rebuke; speak of one's badness'	
mengatakan	265381	'to say'	
melangkah	5193	'to move'; 'to progress (metaphorically)'	
melangkahi	172	'to disregard'; 'to step over'	
melangkahkan	339	'to move/step the foot forward'	
menyesal	2556	'to be/feel sorry/regret'	
menyesali	956	'to regret/feel sorry for sth.; to deplore'	
menyesalkan	1976	'regret, repent, resent sth.'	
menyusu	458	'(of a baby/young animal) to suckle'	
menyusui	2538	'to breast-feed (sb.)'	
menyusukan	35	'to let sb. suckle; to breast-feed sb.'	
menanda	121	'to sign' (113 tokens); 'to mark at sth.' (5 tokens)	*Menanda* was part of the orthographical issue discussed in §5.4.
menandai	3392	'to mark, label sth.'	
menandakan	3327	'to indicate, mean, signify'	
menapak	303	'to step/thread on'	
menapaki	582	'to walk in; to set foot on'	
menapakkan	79	'to step on the sole of the foot'	
menempat	27	'to occupy; to reside in'; cf. the NOTE column	3 tokens are split of *menempatkan*; 1 token is misspelling of *menempa* 'to forge'; 9 tokens are analogous to *menempatkan*; the remaining tokens are analogous to *menempati* 'to occupy; reside in'
menempati	11150	'to occupy; to reside in'	
menempatkan	11513	'to place X at Y; to deploy'	
mewakil	39	'to (be the) represent(ative of) X'	
mewakili	12389	'to (be the) represent(ative of) X'	

Word	Frequency	Senses	Notes
mewakilkan	220	‘to assign X as a representative’	
mewaris	24	‘to become heir’	
mewarisi	1216	‘to inherit’	
mewariskan	573	‘to bequeath, will, pass down’	

References

Arka, I Wayan. 2010. Dynamic and stative passives in Indonesian & their computational implementation. Presented at the MALINDO Workshop, Jakarta.

Arka, I Wayan. 2012. Developing a deep grammar of Indonesian within the ParGram framework: Theoretical and implementational challenges. *Proceedings of the 26th Pacific Asia Conference on Language, Information and Computation*, 19–38. Faculty of Computer Science, Universitas Indonesia. http://aclweb.org/anthology/Y12-1002 (7 November, 2018).

Arka, I Wayan, Mary Dalrymple, Meladel Mistica, Suriel Mofu, Avery D. Andrews & Jane Simpson. 2009. A linguistic and computational morphosyntactic analysis for the applicative *-i* in Indonesian. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG09 Conference*. CSLI Publications.

Arka, I Wayan & Nurenzia Yannuar. 2016. On the morphosyntax and pragmatics of *-in* in Colloquial Jakartan Indonesian. *Indonesia and the Malay World* 1–23. doi:10.1080/13639811.2016.1215129.

Biemann, Chris, Gerhard Heyer, Uwe Quasthoff & Matthias Richter. 2007. The Leipzig Corpora Collection: Monolingual corpora of standard size. In Matthew Davies, Paul Rayson, Susan Hunston & Pernilla Danielsson (eds.), *Proceedings of the Corpus Linguistics Conference*. University of Birmingham, UK. http://ucrel.lancs.ac.uk/publications/CL2007/paper/190_Paper.pdf (6 March, 2014).

Booij, Geert. 2007. *The Grammar of Words: An Introduction to Linguistic Morphology*. 2nd edn. Oxford: Oxford University Press.

Church, Kenneth Ward & Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics* 16(1). 22–29.

Cole, Peter & Min-Jeong Son. 2004. The argument structure of verbs with the suffix *-kan* in Indonesian. *Oceanic Linguistics* 43(2). 339–364.

Denistia, Karlina, Elnaz Shafei-Bajestan & R. Harald Baayen. 2019. Semantic vector model on the Indonesian prefixes *PE-* and *PEN-*. *Proceedings of the 11th International Conference on the Mental Lexicon*, vol. 1. Edmonton, Alberta, Canada. doi:10.7939/r3-s6b9-cm04.

Desagulier, Guillaume. 2017. *Corpus Linguistics and Statistics with R*. Cham: Springer International Publishing. doi:10.1007/978-3-319-64572-8 (27 January, 2018).

Erk, Katrin. 2012. Vector space models of word meaning and phrase meaning: A survey. *Language & Linguistics Compass* 6(10). 635–653. doi:10.1002/lnco.362.

Firth, John Rupert. 1957. A synopsis of linguistic theory 1930–55. In *Special Volume of the Philosophical Society*, 1–32. Oxford: Oxford University Press.

Galili, Tal. 2015. dendextend: an R package for visualizing, adjusting, and comparing trees of hierarchical clustering. *Bioinformatics*. doi:10.1093/bioinformatics/btv428. https://academic.oup.com/bioinformatics/article/31/22/3718/240978/dendextend-an-R-package-for-visualizing-adjusting.

Goldhahn, Dirk, Thomas Eckart & Uwe Quasthoff. 2012. Building large monolingual dictionaries at the Leipzig Corpora Collection: From 100 to 200 languages. *Proceedings of the 8th Language Resources and Evaluation Conference (LREC) 2012*, 759–765. Istanbul. http://www.lrec-conf.org/proceedings/lrec2012/pdf/327_Paper.pdf (6 March, 2014).

Gries, Stefan Th. 2013. *Statistics for linguistics with R: A practical introduction*. 2nd edn. Berlin: Mouton de Gruyter.

Gries, Stefan Th. & Anatol Stefanowitsch. 2004. Extending collostructional analysis: A corpus-based perspective on "alternations." *International Journal of Corpus Linguistics* 9(1). 97–129.

Harris, Zellig S. 1954. Distributional structure. *WORD* 10(2–3). 146–162. doi:10.1080/00437956.1954.11659520.

Heylen, Kris, Thomas Wielfaert, Dirk Speelman & Dirk Geeraerts. 2015. Monitoring polysemy: Word space models as a tool for large-scale lexical semantic analysis. *Lingua* (Polysemy: Current Perspectives and Approaches) 157. 153–172. doi:10.1016/j.lingua.2014.12.001.

Hilpert, Martin & Florent Perek. 2015. Meaning change in a petri dish: constructions, semantic vector spaces, and motion charts. *Linguistics Vanguard* 1(1). doi:10.1515/lingvan-2015-0013 (13 February, 2016).

Hilpert, Martin & David Correia Saavedra. 2017. Using token-based semantic vector spaces for corpus-linguistic analyses: From practical applications to tests of theoretical claims. *Corpus Linguistics and Linguistic Theory* 0(0). doi:10.1515/cllt-2017-0009 (30 October, 2017).

Kiela, Douwe & Stephen Clark. 2014. A systematic study of semantic vector space model parameters. *Proceedings of the 2nd Workshop on Continuous Vector Space Models and their Compositionality (CVSC)*, 21–30. Gothenburg, Sweden: Association for Computational Linguistics. http://www.aclweb.org/anthology/W14-1503 (29 December, 2018).

Kroeger, Paul. 2007. Morphosyntactic vs. morphosemantic functions of Indonesian *-kan*. In Annie Zaenen (ed.), *Architectures, Rules, and Preferences: Variations on Themes by Joan W. Bresnan*, 229–251. 1st edn. Stanford, Calif: Center for the Study of Language and Information.

Larasati, Septina Dian, Vladislav Kuboň & Daniel Zeman. 2011. Indonesian morphology tool (MorphInd): Towards an Indonesian corpus. *Systems and Frameworks for Computational Morphology*, 119–129. Springer, Berlin, Heidelberg. doi:10.1007/978-3-642-23138-4_8 (12 August, 2017).

Levin, Beth. 1993. *English verb classes and alternations: A preliminary investigation*. Chicago: University of Chicago Press.

Levshina, Natalia. 2014. Geographic variation of quite + ADJ in twenty national varieties of English: A pilot study. *Yearbook of the German Cognitive Linguistics Association* 2(1). 109–126. doi:10.1515/gcla-2014-0008.

Levshina, Natalia. 2015. *How to Do Linguistics with R: Data Exploration and Statistical Analysis*. John Benjamins Publishing Company. doi:10.1075/z.195.

Maechler, Martin, Peter Rousseeuw, Anja Struyf, Mia Hubert & Kurt Hornik. 2018. *cluster: Cluster Analysis Basics and Extensions*. R package version 2.0.7-1.

Mikolov, Tomas, Kai Chen, Greg Corrado & Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv:1301.3781 [cs]*. http://arxiv.org/abs/1301.3781 (14 December, 2018).

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg Corrado & Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. *arXiv:1310.4546 [cs, stat]*. http://arxiv.org/abs/1310.4546 (14 December, 2018).

Mikolov, Tomas, Wen-tau Yih & Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 746–751. Atlanta, Georgia: Association for Computational Linguistics. http://www.aclweb.org/anthology/N13-1090 (15 December, 2018).

Musgrave, Simon. 2001. *Non-Subject Arguments in Indonesian*. Melbourne, Australia: The University of Melbourne PhD thesis.

Nomoto, Hiroki, Hannah Choi, David Moeljadi & Francis Bond. 2018. MALINDO Morph: Morphological dictionary and analyser for Malay/Indonesian. *Proceedings of the LREC 2018 Workshop "The 13th Workshop on Asian Language Resources,"* 36–43. http://lrec-conf.org/workshops/lrec2018/W29/pdf/8_W29.pdf.

Pennington, Jeffrey, Richard Socher & Christopher D. Manning. 2014. GloVe: Global vectors for word representation. *Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543. http://www.aclweb.org/anthology/D14-1162.

Quasthoff, Uwe & Dirk Goldhahn. 2013. *Indonesian corpora*. Leipzig, Germany: Abteilung Automatische Sprachverarbeitung, Institut für Informatik, Universität Leipzig. http://asvdoku.informatik.uni-leipzig.de/corpora/data/uploads/corpus-building-vol7-ind.pdf (27 July, 2015).

Rajeg, Gede Primahadi Wijaya, Karlina Denistia & Simon Musgrave. 2019a. Dataset for *Vector space model and the usage patterns of Indonesian denominal verbs*. figshare. doi:10.6084/m9.figshare.8187155.

Rajeg, Gede Primahadi Wijaya, Karlina Denistia & Simon Musgrave. 2019b. R Markdown Notebook for *Vector space model and the usage patterns of Indonesian denominal verbs*. figshare. doi:10.6084/m9.figshare.9970205.

Rajeg, Gede Primahadi Wijaya & I Made Rajeg. 2019. Analisis Koleksem Khas dan potensinya untuk kajian kemiripan makna konstruksional dalam Bahasa Indonesia. In I Nengah Sudipa (ed.), *ETIKA BAHASA Buku Persembahan Menapaki Usia Pensiun: I Ketut Tika*, vol. 1, 65–83. Denpasar, Bali, Indonesia: Swasta Nulus. doi:10.26180/5bf4e49ea1582. https://osf.io/preprints/inarxiv/uwzts/ (30 January, 2019).

Schmidt, Ben & Jian Li. 2017. *WordVectors: Tools for Creating and Analyzing Vector-Space Models of Texts*. R package. http://github.com/bmschmidt/wordVectors.

Shiohara, Asako. 2012. Applicatives in Standard Indonesian. In Wataru Nakamura & Ritsuko Kikusawa (eds.), *Objectivization and Subjectivization: A Typology of Voice Systems* (Senri Ethnological Studies 77), 59–76. Osaka: National Museum of Ethnology.

Sneddon, James Neil, Alexander Adelaar, Dwi Noverini Djenar & Michael C. Ewing. 2010. *Indonesian Reference Grammar*. 2nd edn. Crows Nest, New South Wales, Australia: Allen & Unwin.

Stefanowitsch, Anatol. 2013. Collostructional analysis. In Thomas Hoffmann & Graeme Trousdale (eds.), *The Oxford Handbook of Construction Grammar* (Oxford Handbooks Online). Oxford: Oxford University Press. doi:10.1093/oxfordhb/9780195396683.013.0016.

Turney, Peter D. & Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research* 37(1). 141–188. doi:10.1613/jair.2934.

Wickham, Hadley & Garrett Grolemund. 2017. *R for Data Science*. Canada: O'Reilly. http://r4ds.had.co.nz/ (7 March, 2017).

Wittgenstein, Ludwig. 1953. *Philosophical investigations*. New York: Macmillan.