Article

Karlina Denistia*, Elnaz Shafaei-Bajestan and
R. Harald Baayen

Exploring semantic differences between the Indonesian prefixes *PE-* and *PEN-* using a vector space model

https://doi.org/10.1515/cllt-2020-0023
Received April 17, 2020; accepted March 18, 2021;
published online April 9, 2021

Abstract: Indonesian has two prefixes, *PE-* and *PEN-*, that are similar in form and meaning, but are probably not allomorphs. In this study, we applied a distributional vector space model to clarify whether these prefixes have discriminable semantics. Comparisons of pairs of words within and across morphologically defined sets of words revealed that cosine similarities of pairs consisting of a word with *PE-* and a word with *PEN-* were reduced compared to pairs of only *PE-* words, or of only *PEN-* words. Furthermore, nouns with *PE-* were more similar to their base words than was the case for words with *PEN-*. The specialized use of *PE-* for words denoting agents, and the specialized use of *PEN-* for denoting instruments, was also visible in the semantic vector space. These differences in the semantics of *PE-* and *PEN-* thus provide further quantitative support for the independent status of *PE-* as opposed to *PEN-*.

Keywords: cosine similarity; distributional semantics; Indonesian; morphology; similarity judgements

***Corresponding author: Karlina Denistia**, Faculty of Cultural Sciences, Gadjah Mada University, Jalan Sosio Humaniora, Bulaksumur Jl. Sagan, Sagan, Caturtunggal, Kec. Depok, Kabupaten Sleman, Daerah Istimewa Yogyakarta 55281, Indonesia; and Department of Linguistics, Eberhard Karls Universitaet Tübingen, Tübingen, Germany, E-mail: karlinadenistia@ugm.ac.id
Elnaz Shafaei-Bajestan and R. Harald Baayen, Department of Linguistics, Eberhard Karls Universitaet Tübingen, Tübingen, Germany, E-mail: elnaz.shafaei-bajestan@uni-tuebingen.de (E. Shafaei-Bajestan), harald.baayen@uni-tuebingen.de (R.H. Baayen)

Open Access. © 2021 Karlina Denistia et al., published by De Gruyter. This work is licensed under the Creative Commons Attribution 4.0 International License.

1 Introduction

In Indonesian, there are two nominalisation prefixes: *PE-* and *PEN-*, which derive nouns with a range of similar meanings (agent, instrument, patient, location, causer) from verbs. Qualitative studies mainly describe *PE-* and *PEN-* as independent prefixes (Ramlan 2009; Sneddon et al. 2010), but there are also studies that take them to be allomorphs (Dardjowidjojo 1983; Kridalaksana 2007). It is unclear whether *PE-* is an allomorph of *PEN-* or is actually an independent formative (Denistia 2018).

The first prefix, *PEN-*, is described as having six phonologically-conditioned allomorphs which are in complementary distribution (Ramlan 2009; Sugerman 2016; Sukarno 2017). The *N* in *PEN-* is a mnemonic for the nasal assimilation that characterizes most of its allomorphs. For notational clarity, we write the prefixes in upper case and distinguish between their allomorphs using subscripts: $PEN_{peng\text{-}}$, $PEN_{pen\text{-}}$, $PEN_{pem\text{-}}$, $PEN_{peny\text{-}}$, $PEN_{penge\text{-}}$; and one non-nasalized allomorph $PEN_{pe\text{-}}$. The second prefix, *PE-*, is clearly similar in form, and has been argued to be very similar also in meaning as $PEN_{pe\text{-}}$ (Nomoto 2006).[1]

The reason that *PE-* is taken to be a different prefix is that nouns with *PE-* are derived from verbs with the prefix *BER-*, and nouns with *PEN-* are derived from verbs with *MEN-* (see, e.g., Benjamin 2009; Dardjowidjojo 1983; Ermanto 2016; Nomoto 2006, 2017; Putrayasa 2008; Ramlan 2009; Sneddon et al. 2010), through a process of affix substitution (e.g. *petani* "farmer" - *bertani* "to farm" and *penari* "dancer" - *menari* "to dance"). Similar to *PEN-*, *MEN-* has also six phonologically-conditioned allomorphs: $MEN_{meng\text{-}}$, $MEN_{men\text{-}}$, $MEN_{mem\text{-}}$, $MEN_{meny\text{-}}$, $MEN_{menge\text{-}}$, and $MEN_{me\text{-}}$.

Verbs with *MEN-* can be extended with the suffixes *-i* and *-kan* (Kroeger 2007; Sneddon et al. 2010; Sutanto 2002; Tomasowa 2007). These suffixes add a further argument: a beneficiary, a causer, or a location (e.g. *tulis* "to write" - *menulisi* "to write on something", *menuliskan* "to write for someone") (Arka et al. 2009; Ramli 2006). Verbs with *BER-* are found with *-kan* or *-an* to express possession and reciprocity (e.g. *alamat* "address" - *beralamatkan* "to have an address", *cium* "to kiss", *berciuman* "to kiss each other"). However, derived nouns with *PE-* and *PEN-* do not carry *-i*, *-kan*, or *-an* suffixes, even though they may correspond to verbs with these suffixes (Nomoto 2006). For instance, *pemilik*, "owner", is paradigmatically related to *memiliki* "to own something", with the suffix *-i*. Importantly, the verb *memilik* does not exist.

1 Nomoto (2006) labelled *PE-* as *PER-*, however, Nomoto (2017) refers to the same prefix as *PE-*.

The relation between form and meaning of *PE-* and *PEN-* is elucidated further by Chaer (2008), Benjamin (2009), and Sneddon et al. (2010), who reported that these prefixes are occasionally attested for the same base word with either the same or different a semantic role. For instance, *PEN-* as in *penembak* and *PE-* as in *petembak* are both derived from the base *tembak*, "to shoot", and denote "someone who shoots" and "shooter (athlete) ", respectively. There are also cases in which, having the same base word, the derived form with *PEN-* expresses the agent and the derived form with *PE-* expresses the patient. For instance, *PEN-* as in *penyapa* and *PE-* as in *pesapa* are both derived from the base *sapa*, "to greet/address", and denote "a person who greets/addresser" and "a person who is greeted/addressee" respectively.

Denistia and Baayen (2019) conducted a corpus-based analysis to investigate whether *PE-* is really an allomorph of *PEN-*. Their study also included a quantitative analysis of the paradigmatic relation between *PEN-* and *PE-* with their corresponding verbal prefixes *MEN-* and *BER-*. They argued that *PE-* and *PEN-* actually are two different prefixes, since these prefixes reveal different degrees of productivity and also show semantic specialization: *PEN-* is more productive in forming agents and instruments, whereas *PE-* primarily forms agents and to some extent patients, but not instruments. They also observed that the number of derived words with an allomorph of *PEN-* is correlated with the number of base words with the corresponding allomorph of *MEN-*. *PE-* and its base with *BER-* do not partake in this correlation; it is an exception to the quantitative paradigmatic relations characterizing the allomorphs of *PEN-* and *MEN-*.

In the present study, we used methods from Distributional Semantics Modelling (DSM; Landauer and Dumais 1997) to investigate potential further semantic differences between *PE-* and *PEN-*. In DSM, word meanings are quantified by looking at words' contexts, following the insight of Firth (1957: 11) that "You shall know a word by the company it keeps". DSM builds on the observations that 1) words that have similar meanings usually occur in similar contexts (Rubenstein and Goodenough 1965); and 2) that words appearing in similar contexts tend to have similar meanings (Pantel 2005). To operationalize this, distributional information of words from large language corpora is brought together in high-dimensional vectors (Turney and Pantel 2010). Thanks to this vector representation, geometric methods that quantify vector similarity can be used to measure the semantic similarity between words of interest.

Methods from distributional semantics have proved useful both for natural language processing (e.g., Alfonseca et al. 2009 in information retrieval; McCarthy et al. 2007 in word sense disambiguation; Cheung and Penn 2013 in textual summarization) and for a range of psycholinguistic tasks, including semantic

priming and similarity judgements (e.g., Lowe and McDonald 2000; Lund and Burgess 1996; McDonald and Brew 2004), and studies of morphological processing (Kuperman and Harald 2009; Lazaridou et al. 2013; Marelli and Baroni 2015). Semantic vector spaces also play a central role in a recent computational model of the mental lexicon (Baayen et al. 2019).

DSM was first applied to Indonesian morphology by Fam et al. (2017). They examined the paradigmatic relations for Indonesian derivational affixes (e.g. *beli:dibeli*, "to buy:to be bought", *makan:makanan*, "to eat:food"), and used a vector space model to generate predictions for the meanings of unseen derived words. In the present study, we constructed a semantic vector space from a large Indonesian corpus. If *PE-* and *PEN-* words differ in meaning, they are expected to occur in systematically different contexts, and be distributed differently in the semantic vector space.

The reminder of this paper is structured as follows. We first introduce the corpus used for this study and the databases that we derived from this corpus. In Section 3, we then describe how we constructed the semantic vector space, derived model-based similarity measures, and obtained human judgements on word similarities. We also present the analyses of the model-predicted similarity values, and a comparison of model predictions with human judgements. Finally, we discuss the results obtained and conclude the study in Section 4.

2 Materials

The main corpus used in this study was the Leipzig Corpora Collection (henceforth, LCC) available at http://corpora2.informatik.uni-leipzig.de/download.html. This corpus was compiled from different sources such as the web, newspapers, and the Wikipedia pages dating from 2008 to 2012 (Goldhahn et al. 2012). It consists of 2, 759,800 sentences, 50,794,093 word tokens, and 112,025 different word types. We obtained the morphological structure of the non-compound words using the MorphInd parser (Larasati et al. 2011) and checked the results manually against the online version of *Kamus Besar Bahasa Indonesia*, a comprehensive dictionary of Indonesian (Alwi 2012). The precision of the parser was at 0.98 with a recall of 0.8 in parsing all the *PE-* and *PEN-* words of the corpus. Overall, we obtained 560,633 Indonesian word types, 47,217,467 tokens, and 314,448 hapax legomena. We processed the data using the R version 3.4.3 programming language (R Core Team 2017). The databases and the R scripts are available online at http://bit.ly/PePeNSemVector.

2.1 Indonesian lemmatized database

Using the morphological analyses provided by MorphInd, we lemmatized the LCC corpus. In a preliminary processing step preceding lemmatization, we lower-cased all words and excluded numbers, punctuation marks, and the 15 highest frequency stop words.[2] During lemmatization, the bound morphemes (*ku-* "I", *-ku* "my", *kau-* "you", *-mu* "your", *-nya* "his/her/its"), prolexemes (e.g. *non-*, *anti-*, *pra-*, *pasca-*), particles (e.g. *-lah* and *-pun* to express emphasis, *-kah* to ask a question), and numeric affixes (e.g. *se-* "one", *per-* "per") were separated from their base word as suggested by Sneddon et al. (2010). We also marked *-nya*, when its function is to emphasize a question word, by *nya-WH* (Pastika 2012). Besides, although MorphInd identifies *antar* as a prolexeme, we did not separate the prolexeme and the base into two tokens as *antar* has a different meaning when it occurs as a simple word (e.g. *antaragama* "among religions" - *antar* "to pick up").

Hyphenated words were dealt with as a special case in the lemmatization process since the hyphen can indicate various morphological word formation patterns such as full reduplication, partial reduplication, imitative reduplication, affixed reduplication, or compounding. Hyphens may also appear in proper names and when an affix is attached to a loan word (Sunendar 2016). The hyphens for *-Nya*, *-Ku*, and *-Mu* (note the capital *N*, *K* and *M*) were lemmatized to *Tuhan* "God" (e.g. *kepada-Mu*, *kepada Tuhan* "to God'). We did not parse reduplicated forms as this word formation process is used to convey different meanings (e.g. plurality, intensification, or iteration; Chaer 2008; Dalrymple and Mofu 2012; Rafferty 2002; Sugerman 2016). Several examples illustrating the output of the lemmatization process are shown in Table 1.

An excerpt from the LLC corpus is presented here, before and after lemmatization. Without lemmatization:

> *Terimakasih karena kau selalu memperhatikanku saat di Korea, saat aku rindu ibuku kau yang menyuruhku untuk menelponnya, bahkan kau juga mengajakku bertemu dengan ibumu untuk mencairkan kerinduanku saat aku benar-benar merindukan ibuku.*

With lemmatization:

> *Terimakasih kau selalu memperhatikan aku saat Korea saat aku rindu ibu ku kau menyuruh aku menelpon dia bahkan kau mengajak aku bertemu ibu mu mencairkan kerinduan ku saat aku benar-benar merindukan ibu ku.*

2 The complete list of the removed stop words comprises *yang* "which", *dan* "and", *di* "in", *itu* "that", *dengan* "with", *untuk* "to/for", *ini* "this", *dari* "from", *tidak* "not", *dalam* "inside", *pada* "of", *akan* "will", *juga* "also", *ke* "to", and *karena* "because".

Table 1: Examples of the lemmatization.

Word	Lemma	English translation
kuajak	*aku ajak*	I invite
acaraku	*acara ku*	My event
mengajarkanku	*mengajarkan aku*	Teach me
bilaku	*bila aku*	If I
kauajar	*kamu ajar*	You teach
acaramu	*acara mu*	Your event
bersamamu	*bersama kamu*	Together with you
acaranya	*acara nya*	His/her event
mengajaknya	*mengajak dia*	Invite him/her
kapannya	*kapan nya-WH*	When
abadilah	*abadi lah*	Eternal-lah
antiagama	*anti agama*	Anti religion
antigennya	*anti gen nya*	His/her anti gen
nonagama	*non agama*	Non religion
pascaacara	*pasca acara*	After event
perempatnya	*per empat nya*	One fourth
praanggapan	*pra anggapan*	Hypothesis
seabad	*satu abad*	One century
hiruk-pikuk	*hiruk-pikuk*	Hustle and bustle
berhari-hari	*berhari-hari*	For days
al-quran	*al-quran*	The Quran
kepada-mu	*kepada tuhan*	To God
rahmat-nya	*rahmat tuhan*	God's blessing
kera-jinan	*kerajinan*	Craft
menying-gung	*menyinggung*	To offend
tetangga-tetangga	*tetangga-tetangga*	Neighbours

"Thank you for always paying attention to me while in Korea, when I missed my mom you told me to call her, even you also invited me to meet your mother to attenuate my longing when I really miss my mother."

2.2 Modelling semantics

The distributional vector representations of *PE-* and *PEN-* target words were extracted from the LLC corpus using word2vec (Mikolov et al. 2013) with the default parameter settings[3] (see also Altszyler et al. 2017 for other methods). Cosine similarity was employed to measure the degree of semantic similarity of two lemmas.

3 We used a skip-gram model, with a window size of five, a vector size of two hundred, and no hierarchical softmax. Items occurring less than five times in the corpus were not included.

Let vectors v and w be two n dimensional vectors representing two lemmas. The cosine similarity of v and w is the cosine of the angle θ between $\vec{v}$ and $\vec{w}$, and is equal to the inner product of the vectors, after being length-normalized (see Equation (1)). Thus, similarity judgement is based on the orientation, and not the magnitude, of the vectors.

Equation 1: Calculation of cosine similarity value between two vectors.

$$\sin(v, w) = \cos(\theta) = \frac{v.w}{\|v\| \, \|w\|} = \frac{\sum_{i=1}^{n} v_i w_i}{\sqrt{\sum_{i=1}^{n} v_i^2} \sqrt{\sum_{i=1}^{n} w_i^2}}$$

2.3 Data sets

Using the cosine similarity, we constructed two datasets, henceforth the CosSim database and the PePeNCos database.[4] The CosSim database contains the cosine similarity values for all possible combinations of pairs of words from the set of *PE-*, *PEN-*, *BER-*, and *MEN-* words. This database also includes the cosine values for *PE-*, *PEN-*, *BER-*, and *MEN-* words with their respective base words. For each of its 37,003,784 entries, the CosSim database provides the following information: Lemma1; Lemma2; Cosine similarity of Lemma 1 and Lemma 2; Prefix (the prefix which the lemma contains, either *PE-*, *PEN-*, *BER-*, or *MEN-*); Base word; Semantic role of the nominalization with *PE-* or *PEN-*: agent, instrument, causer, patient, location; Derived-base cosine similarity, i.e., the cosine similarity of the derived word and its base word; and the word category of the base word. For agent nouns formed with *PE-*, we also specified whether the word refers to an athlete or a non athlete. Example entries of this database are listed in Table 2.

The semantic roles assigned to the nominalizations with *PE-* and *PEN-* are based on manual annotation carried out by the first author, based on words' occurrences in the corpus. For each type, at least one token was sampled from the corpus, and checked against the *Kamus Besar Bahasa Indonesia*. Nominalizations that may express multiple semantic roles, cf. "opener" in English, *pembuka* in Indonesian, are linked with an "agent-instrument" semantic role. Manual inspection of all of the 579,695 *PE-* and *PEN-* word tokens in the corpus was not feasible. Thus, the manual annotation of semantic roles is necessarily incomplete.

4 In this database, we did not distinguish between instrument and impersonal agent (see Booij 1986 for impersonal agent as in the Dutch word *zender* "radio station" which has both an agentive reading, "one who sends", and an instrumental reading, "transmitter").

Table 2: Examples of entries in the CosSim database.

Lemma1	Lemma2	Cos	PrefixL1	PrefixL2	SemRoleL1	SemRoleL2
menjadi "to become"	*dalam* "inside"	−0.07	PEN			
bekerja "to work"	*abadi* "eternal"	−0.07	BER			
mengatakan "to say"	*menjadi* "to become"	0.09	MEN	MEN		
melakukan "to do"	*bekerja* "to work"	0.19	MEN	BER		
pemerintah "government"	*dalam* "inside"	−0.08	MEN		agent-instrument	
petugas "officer"	*pemerintah* "government"	0.08	PE	PEN	agent	agent-instrument

The PePeNCos database is a subset of the CosSim database and contains 81 derived words with *PE-* and 910 derived words with *PEN-*. The database specifies the cosine similarity of the derived word and the corresponding base word, the word class of the base word, and the semantic role of the derived word. From this database, we excluded *PE-* and *PEN-* words that do not have a verbal base that co-occurs with the prefix *MEN-* or *BER-* (Dardjowidjojo 1983; Kridalaksana 2007; Nomoto 2017; Ramlan 2009; Sneddon et al. 2010). Table 3 presents some examples of entries in this database.

2.4 Semantic similarity rating

Eighty-three Indonesian native speakers were asked, by means of an online questionnaire, to rate pairs of words with respect to their similarity in meaning on a

Table 3: Examples of entries in the PePeNCos database.

DerivedWord	BaseWord	Cos	Prefix	BaseWordClass	SemRole
peanggar "fencing athlete"	*anggar* "fencing"	0.05	PE-	n	agent
pebasket "basketball player"	*basket* "basketball"	0.35	PE-	n	agent
pebisnis "businessman"	*bisnis* "business"	0.56	PE-	n	agent
pemain "player"	*main* "to play"	0.22	PEN-	v	agent-instrument
pemerintah "government"	*perintah* "order"	0.08	PEN-	n	agent-instrument
penulis "writer"	*tulis* "to write"	0.45	PEN-	v	agent

5-point Likert scale (Likert 1932), following Miller and Charles (1991). Participants were first presented with a set of instructions that illustrated and exemplified the task. Subsequently, they were requested to judge the similarity between 48 noun base words and the corresponding derived words with *PE-* and *PEN-* on a scale from 0 (no similarity in meaning) to 4 (very similar in meaning). An "I don't know" option was provided to the participants just in case some low frequency words would not be recognized. These responses were removed from our analyses. Participants were free to re-rate any pairs before submitting their final judgements.

Our word materials consisted of 24 *PE-* words and 24 *PEN-* words and their base words. Out of the set of 48 *PE-* and *PEN-* words, 47 have unique base words; two *PEN-* words share the same base word. Across prefixes, we controlled for the frequency of base and derived words, in which both of them displayed a comparable wide range of cosine similarity values. The words were selected pseudorandomly, while ensuring that different base word frequencies (High and Low), different derived noun frequencies (High and Low), and different cosine values (see Figure 1) were present in the dataset. A word's frequency was classified as High or Low when present in the list of the top 20% or the bottom 20% most frequent words, respectively. This data set, which contains the human ratings as well as the cosine similarity values, is available in the supplementary materials.[5] Example entries are listed in Table 4.

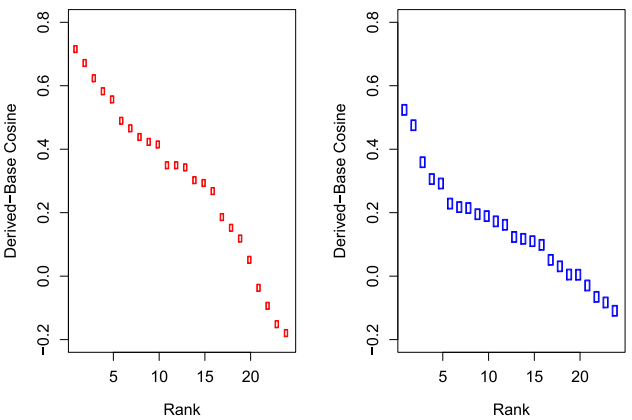

Figure 1: Rank distribution of cosine similarities of words with *PE-* (left panel) and words with *PEN-* (right panel) with their respective base words, as used in the semantic similarity judgement task.

5 The supplementary materials are accessible at http://bit.ly/PePeNSemVector.

Table 4: Examples of entries of the database with human similarity ratings. Part: participant.

NounBase	DerivedNoun	Part.	SimScore	PE	BaseFreq	DerivedFreq	Cos
jalan "street"	*pejalan* "walker"	1	1	T	40	30	0.47
jalan "street"	*pejalan* "walker"	80	4	T	40	30	0.47
obat "medicine"	*pengobat* "who/which cures"	1	1	F	30	13	0.18
obat "medicine"	*pengobat* "who/which cures"	80	2	F	30	13	0.18
runding "discussion"	*perunding* "who discuss"	1	2	T	11	20	0.44
runding "discussion"	*perunding* "who discuss"	80	4	T	11	20	0.44
rintis "pioneer"	*perintis* "pioneer"	1	2	F	9	29	0.03
rintis "pioneer"	*perintis* "pioneer"	80	4	F	9	29	0.03
tenis "tennis"	*petenis* "tennis player"	1	3	T	23	38	0.49
tenis "tennis"	*petenis* "tennis player"	80	4	T	23	38	0.49
waris "inheritance"	*pewaris* "heir"	1	2	F	16	26	0.31
waris "inheritance"	*pewaris* "heir"	80	4	F	16	26	0.31
anggar "fencing"	*peanggar* "fencer"	1	4	T	12	9	0.05
anggar "fencing"	*peanggar* "fencer"	80	1	T	12	9	0.05
saksi "witness"	*penyaksi* "who witness"	1	1	F	33	4	0.05
saksi "witness"	*penyaksi* "who witness"	80	2	F	33	4	0.05

3 Analysis

In what follows, we first compare the semantic similarities within and between the sets of words with *PE-* and *PEN-* (Section 3.1). In Section 3.2, we address the semantic similarities of the base words of these prefixes. Following this, we address the different semantic roles that are realized by words with *PE-* and *PEN-* again using the cosine similarity measure (Section 3.3). Section 3.4 investigates semantic similarity for base words and their prefixed derivatives, and Section 3.5 concludes with comparing the corpus-based semantic similarities with human ratings of semantic similarity.

3.1 Cosine similarity of *PE-* and *PEN-*

We made use of linear discriminant analysis (LDA) to clarify whether the *PE-* and *PEN-* words are separable in semantic space. The LDA was able to reach 95% classification accuracy for 81 *PE-* (27 athlete, 54 non athlete) and 910 *PEN-* words (all of which have a minimal token frequency of 5). As shown in Table 5 (left), the model assigned nearly half of the *PE-* words correctly. A second LDA was given the task to discriminate between *PEN-*, *PE- athletes, and PE- non-athletes*. Interestingly, as shown in the right subtable of Table 5, the nine *PEN-* words that were misclassified as *PE-* were assigned to the $PE_{non\text{-}athelete}$- group. *PEN-* is never confused with $PE_{athlete}$-. The athlete subset is clearly less confusable with *PEN-* than the non-athlete subset.

We complemented the LDA analysis with visualization using Principal-Components. Figure 2, left panel, shows the locations of *PE-* and *PEN-* words in the space spanned by the first two principal components. Independent-samples t-tests were conducted to compare the mean of *PE-* and *PEN-* vectors for each dimension. For the first dimension, the mean of *PE-* is –1.18, whereas *PEN-* is 0.11 ($p < 0.0001$). For the second dimension, the mean of *PE-* is –0.36, while *PEN-* is 0.03 ($p = 0.03473$). Further independent-samples t-tests for the first and the second dimension showed different means for $PE\text{-}_{athlete}$ (–1.9 and –1.03) and $PE\text{-}_{non\text{-}athlete}$ (–0.82 and –0.02; $p = 0.026$ for the first comparison and $p = 0.001$ for the second comparison).

Figure 3, left panel, presents boxplots summarizing the distributions of cosine similarities for three sets of word pairs: *PE-/PEN-* pairs (set 1), *PEN-/PEN-* pairs (set 2), and *PE-/PE-* pairs (set 3); see examples in Table 6. Although the distributions show considerable overlap, differences in mean cosine similarity do reach significance for the between prefix comparisons (*PE-/PEN-*) and within-prefix

Table 5: The confusion table of model prediction between *PE-* and *PEN-* (left) using linear discriminant analysis, and between *PE-* and *PEN-* prediction when *PE-* is split into athlete and non-athlete (right). Columns: observed, rows: predicted.

	PE-	**PEN-**
PE-	39	9
PEN-	42	901

	$PE_{athlete}$-	**$PE_{non\text{-}athelete}$-**	**PEN-**
$PE_{athlete}$-	16	0	0
$PE_{non\text{-}athelete}$-	0	25	8
PEN-	11	29	902

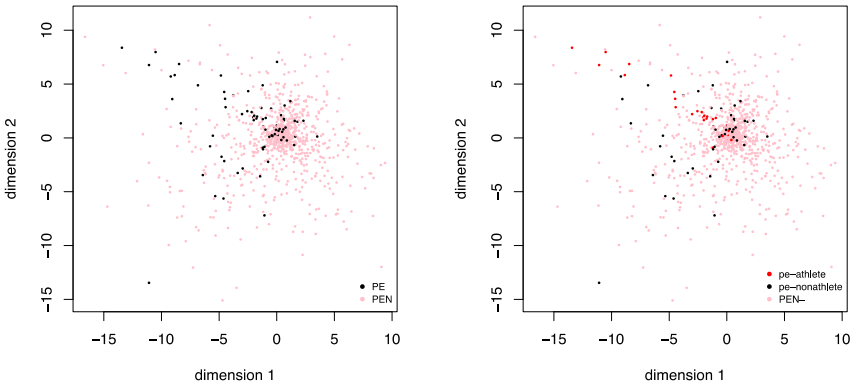

Figure 2: *PEN-* words (red) and *PE-* words (black) in the plane spanned by the first two principal components of PCA analysis of the semantic vectors of these words. Left panel: *PE-* and *PEN-*. *PE-* is clustered more on the central to left part, whereas *PEN-* is more to the central-right part. Right panel: *PE-* (broken down by athlete and non-athlete) and *PEN-*. *PE-athlete* and *PE-non-athlete* are reasonably well separated.

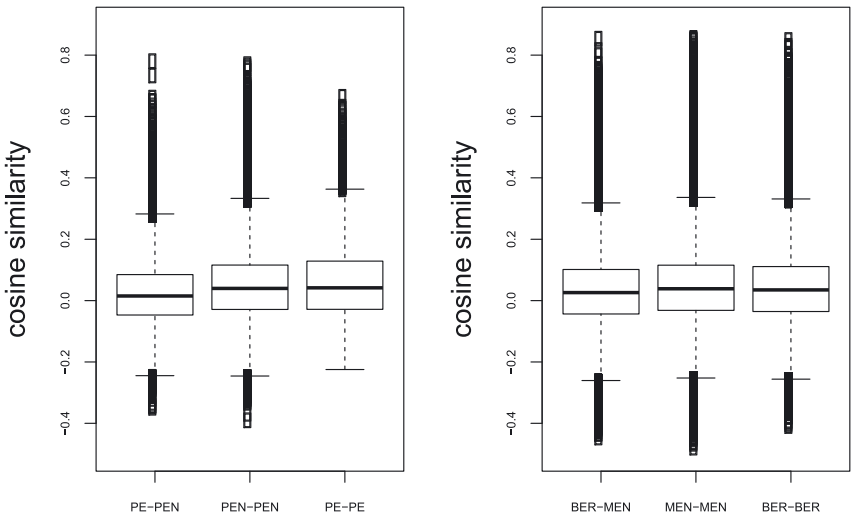

Figure 3: Boxplots for the distributions of cosine similarities. Left panel: cosine similarities for between *PE-* and *PEN-*, within *PEN-*, and within *PE-* words. Within and between prefix cosine similarities, group means are significantly different only for between prefix comparisons. Right panel: cosine similarities between *MEN-* and *BER-*, within *MEN-* and within *BER-*. For these base words, all pairs of group means are significantly different.

comparisons (either *PEN-PEN* or *PE-PE*). A Kruskal-Wallis rank sum test confirmed the presence of at least one significant difference ($\chi^2_{(2)} = 2535.1, p < 0.0001$; mean cosine similarities: 0.024 for set 1, 0.049 for set 2, and 0.07 for set 3). Post-hoc pairwise multiple comparisons using the Nemenyi test and p-value adjustment using the Bonferroni correction confirmed that mean cosine similarity for the *PE-/PEN-* group is indeed significantly lower than that for the *PEN-/PEN-* and the *PE-/PE-* groups ($p < 0.0001$ for both comparisons). The between-prefix cosine similarities indicate that *PE-* and *PEN-* formations form relatively cohesive clusters within their own class in semantic space, and that these classes are not fully overlapping in semantic space. The mean cosine similarity for word pairs within the *PEN-* group, however, is not convincingly different from the cosine similarity of pairs within the *PE-* group ($p = 0.049$).

3.2 Cosine similarity and paradigmatic relations

Since *PE-* and *PEN-* are paradigmatically related with the verbal prefixes *MEN-* and *BER-*, respectively, that occur in the nominalization's base words (see Benjamin

Table 6: Examples of entries for each prefix and semantics role set. BCL1: word class of the base of lemma 1, BCL2: word class of the base of lemma 2.

Lemma1	Lemma2	Cos	PrefixTag	SemRoleTag	BCL1	BCL2
pelari "runner"	*peanggar* "fencing athlete"	0.06	PE-PE	agent-agent	v	n
pelari "runner"	*pejuang* "fighter"	0.04	PE-PE	agent-agent	v	v
pembisik "whisperer"	*pecandu* "drug addict"	0.07	PEN-PEN	agent-agent	n	n
pengabdi "devoter"	*pecandu* "drug addict"	0.09	PEN-PEN	agent-agent	n	n
pelacak "detector"	*pelindung* "protector"	0.18	PEN-PEN	agent-instrument-agent-instrument	v	v
pelacak "detector"	*pemandu* "guide"	0.14	PEN-PEN	agent-instrument-agent-instrument	v	n
pelangsing "slimming pill"	*peledak* "exploder"	0.12	PEN-PEN	instrument-instrument	adj	v
pelangsing "slimming pill"	*pelembap* "moisturizer"	0.46	PEN-PEN	instrument-instrument	adj	adj
pejuang "fighter"	*pecandu* "drug addict"	−0.02	PE-PEN	agent-agent	v	n
petenis "tennis player"	*pecandu* "drug addict"	−0.03	PE-PEN	agent-agent	n	n

2009; Dardjowidjojo 1983; Ermanto 2016; Nomoto 2017; Putrayasa 2008; Ramlan 2009; Sneddon et al. 2010), we investigated whether verbs with *MEN-* and verbs with *BER-* show a similar trend as the corresponding nouns, such that within-prefix similarities (*MEN-/MEN-*; *BER-/BER-*) are greater than between prefix similarities *MEN-/BER-*. For this comparison, we selected all verbs with *MEN-* and *BER-*, regardless of whether they correspond to *PEN-* and *PE-* or not. Table 7 shows how often *MEN-*, *BER-*, *PE-*, and *PEN-* prefixes attach to monomorphemic base words, as well as the prevalence of verb-noun affix substitution pairs.

Figure 3, right panel, presents boxplots summarizing the distributions of cosine similarities for *BER-/MEN-*, *MEN-/MEN-*, and *BER-/BER-* pairs. The Kruskal-Wallis rank sum test ($hskip2pt\chi^2_{(2)} = 34699, p < 0.0001$) and Bonferroni-corrected pairwise tests clarified that the mean for *BER-/MEN-* pairs (0.032) is significantly smaller than those for the within-prefix pairs ($p < 0.0001$ for both comparisons). In addition, the mean cosine similarity for word pairs within the *BER-* set (0.042) is significantly lower than the mean of the pairs within the *MEN-* set (0.046; $p < 0.0001$). Although the differences for the base verbs are smaller than for the nominalizations, it is the case that for both nouns and verbs the comparisons between prefixes yield somewhat lower mean similarities than those within prefixes. We can therefore conclude that the paradigmatic system of *PE-/PEN-* and *BER-/MEN-* shows coherence not only at the level of form, but also to some extent at the level of semantics.

3.3 Cosine similarity and semantic roles

We observed that within-prefix word pairs are more similar in their semantics than between-prefix pairs. Since Denistia and Baayen (2019) have shown that *PE-* can realize the patient semantic role, and that *PEN-* can realize the instrument semantic role, and that both may realize the agent semantic role, the question arises whether the present semantic vectors are sufficiently sensitive to reflect these differences in what

Table 7: Counts of tokens and types for *MEN-*, *BER-*, *PEN-*, and *PE-*. The noun-verb correspondence is calculated based on how often the same base word occurs with the prefixes of interest.

Prefix	Tokens	Types
MEN-	2,912,664	3,131
BER-	840,025	1,453
PE-	80,996	81
PEN-	497,207	910
Corresponding PEN- and MEN- formations	471,250	822
Corresponding PE- and BER- formations	75,491	53

semantic roles the different prefixes may realize. The most frequent semantic roles for each prefix, agent for *PE-* and agent and instrument for *PEN-*, were selected for further analysis. Patient *PE-* observations were too few to be included. *PEN-* words were further distinguished by whether they realized multiple semantic roles (both agent and instrument) depending on the context (Jalaluddin and Syah 2009). Of specific interest are five groups of word pairs: (1) *PE-* and *PEN-* words expressing agent, (2) *PE-* words expressing agent, (3) *PEN-* words expressing agent, (4) *PEN-* words expressing instrument, and (5) *PEN-* words expressing both agent and instrument.

Figure 4, left panel, shows that the distribution of cosine similarities for *PE-/PEN-* pairs is shifted down compared to the distributions for the pairs of words with *PE-* and pairs of words with *PEN-*. A Kruskal-Wallis rank sum test ($hskip2ptpt\chi^2_{(2)} = 362.41, p < 0.0001$) and Bonferroni-corrected pairwise tests clarified that the means for within-prefix agent pairs, *PE-* as agents (0.082) and *PEN-* as agents (0.044), are significantly higher than the mean for between-prefix agent pairs *PEN-/PE-* (0.033). Furthermore, the tests also clarified that agents with the less productive *PE-* prefix are significantly more similar than those with the more productive *PEN-* prefix ($p < 0.0001$).

In our data, *PEN-* expresses agent, instrument, or sometimes both agent and instrument, and has a productivity index V1/N (Baayen 2009) of 0.00085 for agents

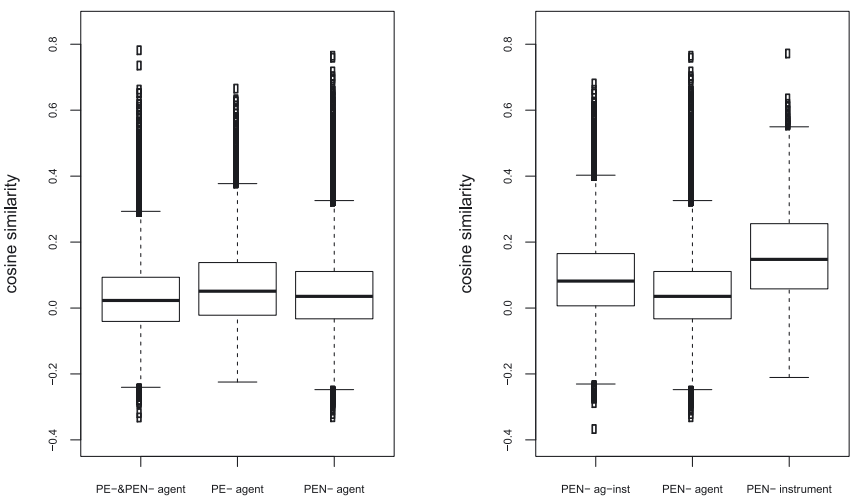

Figure 4: Boxplots for the distributions of cosine similarities for cross-prefix pairs of words with *PE-* and *PEN-* expressing agents, as well as for within-prefix pairs expressing agents (left panel). The right panel compares the distributions of cosine similarities for words with *PEN-*, comparing pairs of words that can realize both agent and instrument, and those realizing either agent or instrument. All pairs of group means are significantly different for both the left and right panels.

that is greater than the productivity index for instruments (0.00035) and that for the mixed cases (0.00001). Within the set of words with *PEN-*, see the right panel of Figure 4, we observe differences in mean cosine similarity between the mixed group and agents (lowest similarities) on the one hand, and the mixed group and instruments (highest similarities) on the other hand. The mixed group is positioned in between the two extreme groups, as expected. A Kruskal-Wallis rank sum test ($hskip2pt\chi^2_{(2)} = 6895.1$, $p < 0.0001$) and Bonferroni-corrected pairwise tests clarified that the mean cosine similarity for *PEN-* words in the mixed set (0.091) was significantly different from the mean for words realizing only the agent (0.044) or only the instrument (0.161, $p < 0.0001$). Interestingly, the mean cosine similarity for *PEN-* agents is lower than that for *PEN-* instruments. In other words, the set of words with *PEN-* realizing instruments is internally more similar. This may be due to more consistent contextual collocations for instruments. For instance, instruments are often used with specific prepositions such as *dengan* "with" or with verbs such as *menggunakan* and *memakai* "to use something" in their context.

Returning to *PE-*, Chaer (2008) observed that *PE-* is the prefix of choice for agents that are athletes (e.g., *petinju* "boxer" and *pecatur* "chess player"). Accordingly, one might suspect that observing a higher cosine similarity for *PE-* as agent compared to *PEN-* as agent in Figure 4 is due to the specific use of *PE-* for athletes. In order to investigate this possibility, we split the set of *PE-* words expressing agents into two subsets, with one subset ($PE\text{-}_{athletes}$) comprising the athletes and the other ($PE\text{-}_{non\text{-}athletes}$) the non-athletes.

As shown in Figure 5, cosine similarities within the $PE\text{-}_{athletes}$ set are quite high (mean 0.255) compared to both non-athletes realized with *PE-* and between-prefix comparisons with (non-athlete) nouns with *PEN-*. A Kruskal-Wallis rank sum test ($hskip2pt\chi^2_{(3)} = 525.99$, $p < 0.0001$) and Bonferroni-corrected pairwise tests clarified that the mean cosine similarities of pairs within the $PE\text{-}_{athletes}$ set are significantly higher than those for the pairs of words in the other sets of agent nouns ($p < 0.0001$). When both $PE\text{-}_{athletes}$ and $PE\text{-}_{non\text{-}athletes}$ are merged into one set, the mean cosine similarity decreases to 0.049; see the left panel of Figure 4. Apparently, the high cosine similarities within the *PE-* agents group are due mainly to the subset of agent nouns that refer to athletes. As we can see in Figure 5, pairs of words are much less similar semantically when only one, or none, refer to an athlete, irrespective of whether they are formed with *PE-* or *PEN-*. However, the small differences in the mean between these three distributions do receive statistical support (all $p < 0.0001$).

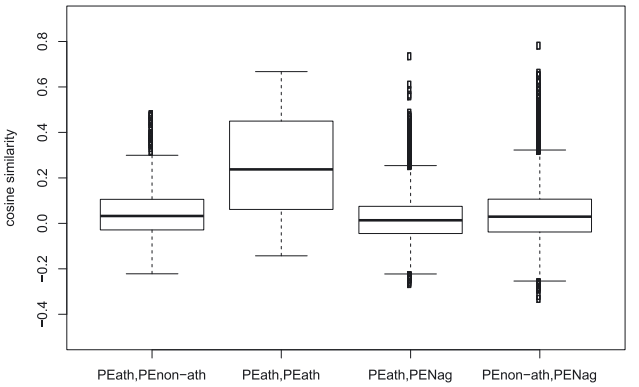

Figure 5: Boxplots for the cosine similarity for *PE-* partition into nouns for athletes and nouns for non-athletes, and agent nouns with *PEN-*.

3.4 Cosine similarity for base-derived pairs

As observed by Chaer (2008), *PE-* is used specifically to coin words for athletes; 34% of types in our data. We therefore expected that base-derived word pairs with *PE-* have a greater mean cosine similarity compared to base-derived word pairs with *PEN-*.

The left panel of Figure 6 presents boxplots for the distributions of cosine similarities for word pairs consisting of a base word and the corresponding nominalization, once for *PE-* and once for *PEN-*. A Wilcoxon test (W = 44,626, $p < 0.0001$) clarified that the mean cosine similarity for *PE-*/BASE word pairs (0.315) is significantly higher than the mean cosine similarity for *PEN-*/BASE word pairs (0.211), as expected. Subsequent analyses that focused on the word category of the base word clarified that the overall pattern is driven entirely by pairs with nouns as base word (W = 2,488, p = 0.648 for verbs; W = 790, p = 0.1329 for adjectives; but W = 5,932, $p < 0.0001$ for nouns). The right panel of Figure 6 shows the distributions for base-derived pairs with noun bases. Since most formations with *PE-* denoting athletes have a nominal base, the larger cosine similarities for *PE-* are again driven primarily by this particular semantic field.

3.5 Modelling human judgement for base-derived pairs

To further validate the corpus-based semantic vectors and the cosine similarity measure, we carried out a rating task in which participants were requested to evaluate the semantic similarity between 48 nominal base words and their

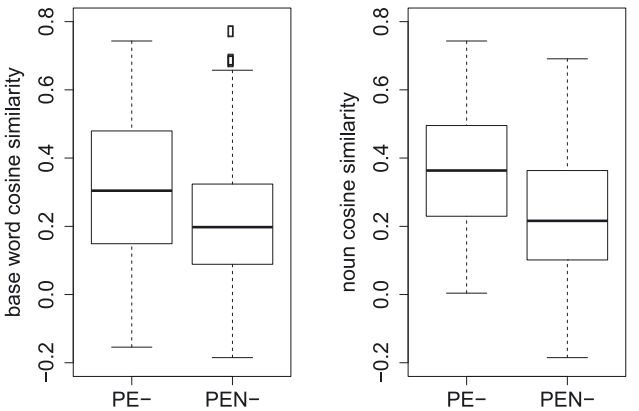

Figure 6: Boxplots for the distributions of cosine similarities for word pairs consisting of the base and the derived word (left panel) and the noun base and the derived word (right panel). Mean cosine similarity is higher for *PE-* compared to *PEN-* in both comparisons.

nominalizations with *PE-* and *PEN-*. Given the results reported in the previous section, we expected the ratings to be lower for the 24 pairs involving *PEN-* than for the 24 pairs involving *PE-*.

Participants were asked to provide ratings on a five-point Likert scale (1–5), for each of the 48 derived/base pairs. Participants were requested to use the full scale. The set of items comprised two subsets of pairs, depending on whether or not the affix of the derived word is *PE-* or *PEN-* (Affix). We selected the items in such a way that there was no strong difference in mean cosine similarity between the *PE-* and *PEN-* groups ($W = 401, p = 0.01937$). For both the derived and the base word, we included their frequency of occurrence as covariates (FrequencyDerived, FrequencyBase).

Out of 83 participants, 13 never used more than three options of the five options available on the rating scale (see Figure 7). These participants were removed prior to analysis. We used a GAMM (Generalized Additive Model, MGCV package version 1.8-17 (Wood 2006, 2011)), for statistical evaluation to investigate whether the cosine similarities and human judgements are correlated. Table 8 presents the summary of a model with a smooth for *PE-* and a difference smooth for *PEN-*. These curves are shown in the left and right panels of Figure 7. A thin plate regression spline was used to model the non-linear interaction of base frequency and derived frequency, and by-participant random intercepts were included as well. Random intercepts for item were not included because an analysis of concurvity indicated item was too strongly confounded with the other item-bound predictors.

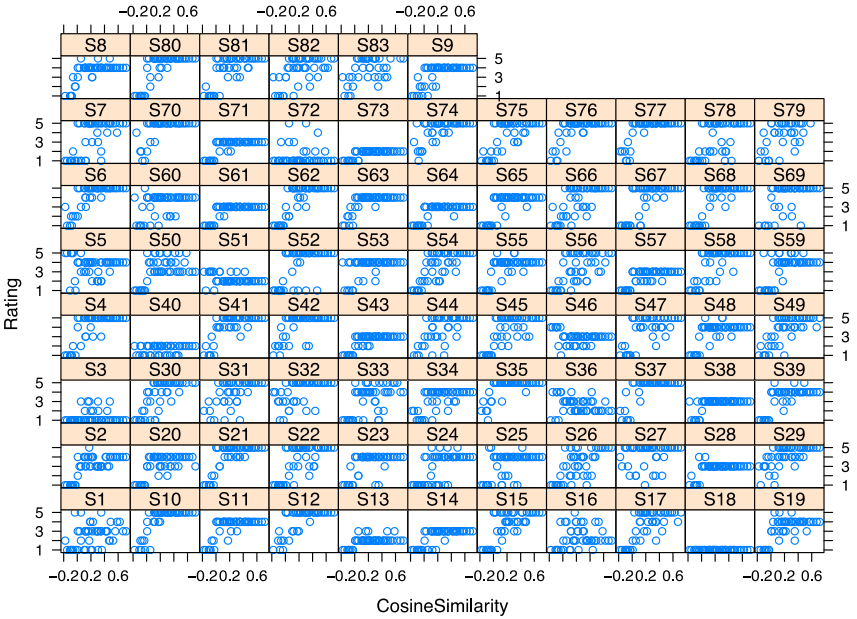

Figure 7: Scatter plot matrix for ratings by cosine similarity for the 83 participants in the human similarity judgement experiment. Participants 3, 13, 14, 18, 38, 40, 43, 51, 57, 61, 64, 71, 73 were removed from the model because of their too restricted use of the rating scale.

Apparently, the way in which human ratings can be predicted from the cosine similarity is different for the two prefixes. As can be seen by comparing the left and centre panels of Figure 8, the effect of cosine similarity is limited to the first two-thirds of the range of its values; the effect levels off for the highest cosine similarity values. This indicates that a large part of the range of cosine similarities is indeed predictive for human intuitions about the semantic similarity between *PE-* and

Table 8: GAMM fitted to the ratings elicited for 48 pairs of *PE-* and *PEN-* nominalizations and their base words.

A. parametric coefficients	**Estimate**	**Std. Error**	**t-value**	**p-value**
Intercept (PE-)	3.63522	0.07466	48.69	<0.0001
B. smooth terms	edf	Ref.df	F-value	p-value
s(CosineSimilarity) [PE-]	2.830	3.521	18.517	<0.0001
s(CosineSimilarity) difference curve PEN-	8.354	9.206	8.786	<0.0001
s(FrequencyBase,FrequencyDerived)	14.457	14.917	29.774	<0.0001
Random intercepts participant	64.879	69.000	15.958	<0.0001

PEN- words and their base words. Furthermore, the upward slope of the regression curve in the predictive range of cosine is steeper for *PE-* than that for *PEN-*, suggesting a greater sensitivity of the cosine of the angle of two semantic vectors as a similarity measure for the prefix *PE-*. The difference curve in the right panel shows that we indeed have a significant difference: around a cosine similarity of 0, the predicted partial effect of *PE-* is significantly lower, and around a cosine similarity of 0.2, it is significantly higher.

4 General discussion

Studies in Indonesian allomorphy have generally focused on words' internal structure. Denistia and Baayen (2019) is the first corpus-based study systematically investigating how complex words are used in written Indonesian. In the present study, we extend their investigation using methods of distributional semantics to study the prefixes *PE-* and *PEN-*, which have been described as having similar form and meaning (Rajeg 2013; Sneddon et al. 2010), have their own quantitative semantic profiles; if so, this would provide further support for *PE-* and *PEN-* being

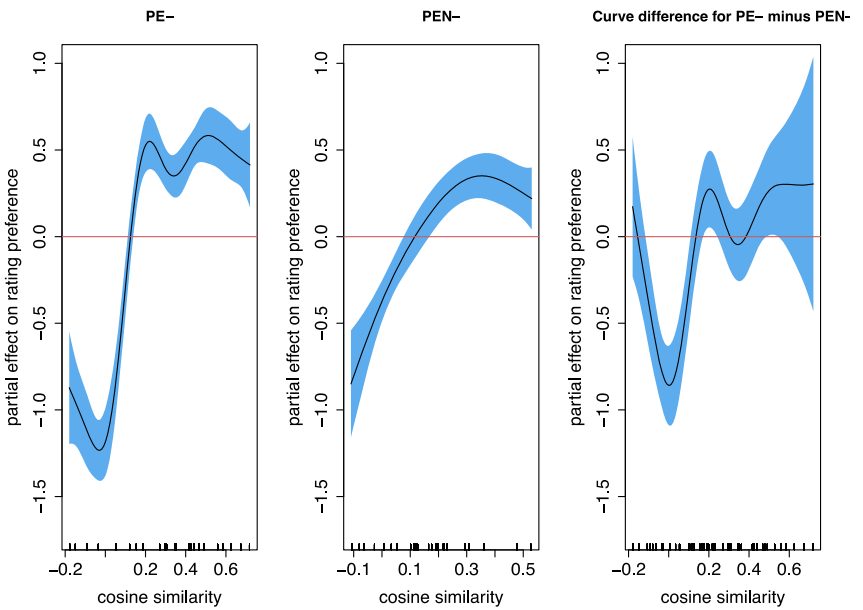

Figure 8: Partial effects for cosine similarity as a predictor of human ratings for *PE-* (left panel) and *PEN-* (middle panel). Right panel: the difference curve which, when added to the curve of *PEN-*, yields the curve of *PE-*.

separate affixes rather than allomorphs (Denistia and Baayen 2019; Nomoto 2017; Ramlan 2009; Sneddon et al. 2010). We used methods from distributional semantics to obtain semantics vectors (also known as word embeddings) for all words with *PE-* and *PEN-*, as well as for their base words and their paradigmatically related verbs with *BER-* and *MEN-*. In addition, we investigated whether the corpus-based cosine similarity measure was predictive for human similarity judgements.

There are subtle but statistically significant differences in the distributions of cosine similarities between *PE-* and *PEN-*. The finding that *PE-* words are less similar to *PEN-* words than to other *PE-* words, and likewise that *PEN-* words are less similar to *PE-* words compared to *PEN-* words, dovetails well with the hypothesis that *PE-* and *PEN-* are different prefixes, rather than allomorphs.

The semantic analyses using embeddings provides further support for paradigmatic consistency between *PE-/PEN-* and *BER-/MEN-* (Benjamin 2009; Dardjowidjojo 1983; Denistia and Baayen 2019; Ermanto 2016; Nomoto 2017; Putrayasa 2008; Ramlan 2009; Sneddon et al. 2010). Cosine similarities calculated between formations with *PE-* and formations with *PEN-* tend to be somewhat smaller than cosine similarities calculated for pairs of words with *PE-* and likewise for pairs of words with *PEN-*. A similar pattern is found for the corresponding base words with *BER-* and *MEN-*. This difference is likely to be due to well described differences in the semantic functions of these prefixes (Arka et al. 2009; Chaer 2008; Kroeger 2007; Putrayasa 2008; Sneddon et al. 2010; Sutanto 2002; Tomasowa 2007). *MEN-* typically renders a verb explicitly active either, transitive or intransitive, and can carry the suffixes *-i* and *-kan*. These suffixes express intensification or iteration (in addition to adding a further argument, either a beneficiary, a location, or a causer). *BER-*, by contrast, is described as a prefix which typically forms intransitive verbs and expresses reciprocals, reflectives, or possessives.

PE- and *PEN-* differ also in that nouns with *PE-* are more similar to their base word compared to nouns with *PEN-*. This finding was supported by a rating experiment, which also suggested that the semantic vectors are indeed predictive of intuitive human judgements of semantic similarity.

Finally, a closer investigation of the semantic roles realized by nominalizations with *PE-* and *PEN-* reveals that the mean cosine similarity for pairs of *PE-* words expressing agents is higher than the mean for pairs of *PEN-* words expressing agents. Furthermore, words with *PEN-* as instruments have a higher mean cosine similarity compared to pairs of words with *PEN-* that express agents.

We have seen that the semantic similarities of pairs of agents realized with *PE-* is slightly greater in the mean than the semantic similarities of pairs of agents realized with *PEN-* (see Figure 4). Furthermore, the semantic similarities of pairs of base and derived words are greater for *PE-* than for *PEN-* (Figure 6). These results

are perhaps surprising given that of the two prefixes, it is *PE-* that is the least productive (Denistia and Baayen 2019). Typically, one would expect greater semantic transparency between base and derived word for more productive affixes.

The somewhat greater transparency of agents with *PE-* is likely to be due to the specific use of *PE-* to express athletes (e.g., *petinju* "boxer" and *perenang* "swimmer"). The overall less productive prefix has found a small semantic niche in which it is strongly established. By way of comparison, irregular verbs in English, German, and Dutch have found a semantic niche comprising actions and positions involving the body (Baayen and Moscoso del Prado Martin 2005). Likewise in Dutch, the less productive suffix *-te* (compare *-th* in English) typically expresses measures (e.g., *lengte*, English *length*), whereas the more productive rival suffix *-heid* is also used for character traits and anaphoric reference (Baayen and Neijt 1997).

In summary, using distributional semantics as analytical tool, we have been able to provide corpus-based evidence for subtle differences in the semantics of the Indonesian prefixes *PE-* and *PEN-*. The present results provide further support for *PE-* and *PEN-* being different prefixes, supplementing earlier studies pointing to differences in their phonological conditioning (Ramlan 2009; Sneddon et al. 2010), differences in their paradigmatic relations with the verbal prefixes of their base words (Nomoto 2017), and differences in their productivity (Denistia and Baayen 2019).

The semantic effects that we have documented in the present study are small. This is likely to be due not only to the enormous differences in words' meanings, but also to the small size of the corpus from which we derived our embeddings. Whereas in natural language processing applications, corpora of several billions of words are favoured, our corpus comprises only 47 million words. As a consequence, our vectors are noisy, especially for lower-frequency words. Further replication studies based on larger corpora will be essential for consolidating the present exploratory results. At the same time, our embeddings have turned out to be surprisingly useful. Several of our observations are predated in the qualitative literature, but it is difficult to evaluate the importance of these observations for the language system. Embeddings have allowed us to provide quantitative corpus-based support for several aspects of the semantics of Indonesian prefixal morphology, and thus provide novel external support and enhanced predictive precision for previous qualitative research.

Acknowledgments: This study was funded by Indonesian Endowment Fund for Education (*Lembaga Pengelola Dana Pendidikan*) (No. PRJ-1610/LPDP/2015) and ERC advanced grant 742545.

References

Alfonseca, Enrique, Keith Hall & Hartmann Silvana. 2009. Large-scale computation of distributional similarities for queries. In *Proceedings of human language technologies: The 2009 annual conference of the North American chapter of the association for computational linguistics, companion volume: Short papers, association for computational linguistics, 29–32*. Boulder: Association for Computational Linguistics.

Altszyler, Edgar, Sidarta Ribeiro, Mariano Sigman & Diego Fernández Slezak. 2017. The interpretation of dream meaning: Resolving ambiguity using latent semantic analysis in a small corpus of text. *Consciousness and Cognition* 56. 178–187.

Alwi, Hasan. 2012. *Kamus besar bahasa Indonesia* [A Comprehensive dictionary of Indonesian], 4th edn. Jakarta: Gramedia Pustaka Utama.

Arka, I. Wayan, Mary Dalrymple, Meladel Mistica, Suriel Mofu, Avery Andrews & Jane Simpson. 2009. A linguistic and computational morphosyntactic analysis for the applicative -I in Indonesian. In Miriam Butt & Tracy Holloway King (eds.), *International lexical functional grammar conference (Lfg)*, 85–105. Cambridge: CSLI Publications.

Baayen, R. Harald. 2009. Corpus linguistics in morphology: Morphological productivity. In Anke Lüdeling & Merja Kyto (eds.), *Corpus linguistics. An international handbook*, 900–919. Berlin: Mouton De Gruyter.

Baayen, R. Harald & Fermin Moscoso del Prado Martin. 2005. Semantic density and past-tense formation in three Germanic languages. *Language* 81(3). 666–698.

Baayen, R. Harald & Anneke Neijt. 1997. Productivity in context: A case study of a Dutch suffix. *Linguistics* 35. 565–587.

Baayen, R. Harald, Yu-Ying Chuang, Elnaz Shafaei-Bajestan & James P. Blevins. 2019. The discriminative lexicon: A unified computational model for the lexicon and lexical processing in comprehension and production grounded not in (de)composition but in linear discriminative learning. *Complexity* 2019. 1–39.

Benjamin, Geoffrey. 2009. Affixes, Austronesian and iconicity in Malay. *Bijdragen Tot de Taal-, Land- En Volkenkunde* 165(2–3). 291–323.

Booij, Geert. 1986. Form and meaning in morphology: The case of Dutch agent nouns. *Linguistics* 24. 503–517.

Chaer, Abdul. 2008. *Morfologi bahasa Indonesia (Pendekatan Proses)* [Indonesian morphology: A processing approach]. Jakarta: PT Rineka Cipta.

Cheung, Jackie Chi Kit & Gerald Penn. 2013. Probabilistic domain modelling with contextualized distributional semantic vectors. *Association for Computational Linguistics (ACL)* 1. 392–401.

Dalrymple, Mary & Suriel Mofu. 2012. Plural semantics, reduplication, and numeral modification in Indonesian. *Journal of Semantics* 29(2). 229–260.

Dardjowidjojo, Soenjono. 1983. *Some aspects of Indonesian linguistics*. Jakarta: Djambatan.

Denistia, Karlina. 2018. Revisiting the Indonesian prefixes PEN-, Pe2-, and PER-. *Linguistik Indonesia* 36(2). 146–159.

Denistia, Karlina & R. Harald Baayen. 2019. The Indonesian prefixes PE- and PEN-: A study in productivity and allomorphy. *Morphology* 29. 385–407.

Ermanto. 2016. *Morfologi afiksasi bahasa Indonesia masa kini: Tinjauan dari Morfologi Derivasi dan Infleksi* [The current Indonesian morphological affiation: A study of derivational and inflectional morphology]. Jakarta: Kencana.

Fam, Rashel, Yves Lepage, Susanti Gojali & Ayu Purwarianti. 2017. Indonesian unseen words explained by form, morphology and distributional semantics at the same time. In 23rd annual meeting of the Japanese association for natural language processing, 178–181.

Firth, John Rupert. 1957. A synopsis of linguistic theory. *Studies in Linguistic Analysis* 147. 1930–1955.

Goldhahn, Dirk, Thomas Eckart & Uwe Quasthoff. 2012. Building large monolingual dictionaries at the leipzig corpora collection: From 100 to 200 languages. In Proceedings of the 8th international language ressources and evaluation (LREC'12), 759–765.

Jalaluddin, Nor Hashimah & Harith Syah Ahmad. 2009. Penelitian makna imbuhan Pen- dalam Bahasa Melayu: Satu kajian rangka rujuk silang [A Research of meaning on Pen- affix in Malay: A *rangka rujuk silang* study]. *GEMA Online Journal of Language Studies* 9(2). 57–72.

Kridalaksana, Harimurti. 2007. *Kelas kata dalam bahasa Indonesia* [Word class in Indonesian], 2nd edn. Jakarta: Gramedia Pustaka Utama.

Kroeger, Paul. 2007. Morphosyntactic versus morphosemantic functions. In Annie Zaenen, Jane Simpson, Tracy Holloway King, Jane Grimshaw, Joan Maling & Chris Manning (eds.), *Architectures, rules, and preferences: Variations on themes of Joan Bresnan*, 229–251. California: CSLI Publications.

Kuperman, Victor & R. Baayen Harald. 2009. Semantic transparency revisited. In *Paper presented at the 6th international morphological processing conference*. Finland: University of Turku.

Landauer, Thomas K. & Susan T. Dumais. 1997. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review* 104. 211–240.

Larasati, Septina Dian, Vladislav Kuboň & Daniel Zeman. 2011. Indonesian morphology tool (morphind): towards an Indonesian corpus. *Systems and Frameworks for Computational Morphology* 100. 119–129.

Lazaridou, Angeliki, Marco Marelli, Roberto Zamparelli & Marco Baroni. 2013. Compositionally derived representations of morphologically complex words in distributional semantics. In *Proceedings of the 51st annual meeting of the association for computational linguistics, 1517–1526*. Sofia: Association for Computational Linguistics.

Likert, Rensis. 1932. A technique for the measurement of attitudes. *Archives of Psychology* 140. 1–55.

Lowe, Will & McDonald Scott. 2000. The direct route: Mediated priming in semantic space. In Lila R. Gleitman & Aravind K. Joshi (eds.), *Proceedings of the twenty-second annual conference of the cognitive science society*, 806–811.

Lund, Kevin & Curt Burgess. 1996. Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers* 28(2). 203–8.

Marelli, Marco & Marco Baroni. 2015. Affixation in semantic space: Modeling morpheme meanings with compositional distributional semantics. *Psychological Review* 122(3). 485–515.

McCarthy, Diana, Rob Koeling, Julie Weeds & John Carroll. 2007. Unsupervised acquisition of predominant word senses. *Computational Linguistics* 33(4). 553–590.

McDonald, Scott & Chris Brew. 2004. A distributional model of semantic context effects in lexical processing. In *Proceedings of the 42nd annual meeting on association for computational linguistics, 17–24,* Barcelona.

Mikolov, Thomas, Kai Chen, Greg Corrado & Jeffrey Dean. 2013. Efficient estimation of word representations in vector space, *arXiv, Preprint arXiv:1301.3781*.

Miller, George A. & Walter G. Charles. 1991. Contextual correlates of semantic similarity. *Language and Cognitive Processes* 1(28). 1–28.

Nomoto, Hiroki. 2006. *A study on complex existential sentences in Malay*. Tokyo: Universiti Bahasa Asing Tokyo MA thesis.
Nomoto, Hiroki. 2017. The syntax of Malay nominalization. In Rogayah Abd. Razak & Radiah Yusoff (eds.), *Aspek Teori Sintaksis Bahasa Melayu*, 71–117. Kuala Lumpur: Dewan Bahasa dan Pustaka.
Pantel, Patrick. 2005. Inducing ontological co-occurrence vectors. In *Proceedings of the 43rd annual meeting on association for computational linguistics*, 125–132. Ann Arbor: Association for Computational Linguistics.
Pastika, I. Wayan. 2012. Klitik -Nya dalam bahasa Indonesia [-Nya clitic in Indonesian]. *Adabiyat* 11(1). 122–142.
Putrayasa, Ida Bagus. 2008. *Kajian Morfologi: Bentuk Derivasional dan Infleksional* [Morphological study: Derivation and inflection]. Bandung: PT Refika Aditama.
R Core Team. 2017. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org (accessed 10 October 2019).
Rafferty, Ellen. 2002. Reduplication of nouns and adjectives in Indonesian. In Papers from the tenth annual meeting of the southeast Asian linguistics society, 317–332.
Rajeg, Gede & Primahadi Wijaya. 2013. Metonymy in Indonesian prefixal word-formation. *Lingual: Journal of Language and Culture* 1(2). 64–81.
Ramlan, Muhammad. 2009. *Morfologi: Suatu tinjuan deskriptif* [Morphology: A descriptive approach]. Yogyakarta: CV Karyono.
Ramli, Md. Salleh. 2006. Imbuhan dan penandaan tematik dalam bahasa Melayu [Affix and thematic roles in Malay]. *Jurnal Melayu* 2. 47–54.
Rubenstein, Herbert & John B. Goodenough. 1965. Contextual correlates of synonymy. *Communications of the ACM* 8(10). 627–633.
Sneddon, James Neil, Adelaar Alexander, Dwi Noverini Djenar & Michael C. Ewing. 2010. *Indonesian: A comprehensive grammar*, 2nd edn. New York: Routledge.
Sugerman. 2016. *Morfologi bahasa Indonesia: Kajian ke arah linguistik deskriptif* [Indonesian morphology: A descriptive linguistics study]. Yogyakarta: Penerbit Ombak.
Sukarno, Sukarno. 2017. The behaviours of the general nasal /N/ in Indonesian active prefixed verbs. *International Journal of Language and Linguistics* 4(2). 48–52.
Sunendar, Dadang. 2016. *Pedoman umum ejaan bahasa Indonesia [General guidelines for Indonesian spelling]*, 4th edn. Jakarta: Badan Pengembangan dan Pembinaan Bahasa Kementerian Pendidikan dan Kebudayaan.
Sutanto, Irzanti. 2002. Verba berkata dasar sama dengan gabungan afiks meN-i atau meN-kan [MeN-i or meN-kan verbs with similar stem]. *Makara, Sosial-Humaniora* 6(2). 82–87.
Tomasowa, Francien Herlen. 2007. The reflective experiential aspect of meaning of the affix -i in Indonesian. *Linguistik Indonesia* 25(2). 83–96.
Turney, Peter D. & Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research* 37. 141–188.
Wood, Simon N. 2006. *Generalized additive models: An introduction with R*. Chapman: Hall/CRC.
Wood, Simon N. 2011. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society (B)* 1(73). 3–36.

Bionotes

Karlina Denistia
Faculty of Cultural Sciences, Gadjah Mada University, Yogyakarta, Indonesia
Department of Linguistics, Eberhard Karls Universitaet Tübingen, Tübingen, Germany
karlinadenistia@ugm.ac.id

Karlina Denistia obtained her PhD in 2020 from Quantitative Linguistics Department, Eberhard Karls Universitæt Tübingen, under a grant from Indonesian Endowment Fund for Education (LPDP). Her researches focus in using corpora to provide quantitative analyses of Indonesian Morphology and Semantics (e.g., the study on productivity and distributional semantics). She is currently working together with Professor Harald Baayen to apply a computational model, a Linear Discriminative Learning, in Indonesian data. She is now joining Faculty of Cultural Sciences, Gadjah Mada University, as a lecturer.

Elnaz Shafaei-Bajestan
Department of Linguistics, Eberhard Karls Universitaet Tübingen, Tübingen, Germany
elnaz.shafaei-bajestan@uni-tuebingen.de

Elnaz Shafaei-Bajestan studied software engineering and computational linguistics and obtained her Master's degree from University of Stuttgart in 2017. She is currently pursuing a doctoral degree at the Quantitative Linguistics group, Eberhard Karls University of Tübingen under the supervision of Professor Harald Baayen. Her work focuses mainly on psycholinguistics computational modeling of spoken and written language, and its implications for theories of speech perception and theories of morphology.

R. Harald Baayen
Department of Linguistics, Eberhard Karls Universitaet Tübingen, Tübingen, Germany
harald.baayen@uni-tuebingen.de

R. Harald Baayen studied linguistics and obtained his PhD 1989 with a quantitative study on morphological productivity. In 1990 he joined the Max Planck Institute for Psycholinguistics in Nijmegen. In 2007 he took up a professorship at the University of Alberta in Edmonton, Canada. In 2011 he received an Alexander von Humboldt research award, which brought him to the University of Tübingen. An ERC advanced grant is supporting his current research programme on discriminative learning.