SEMANTIC VECTOR MODEL ON THE INDONESIAN PREFIXES PE- AND PEN-

Karlina Denistia1, Elnaz Shafaei-Bajestan, and R. Harald Baayen2

Quantitative Linguistics – Eberhard Karls University of Tübingen
karlina.denistia@student.uni-tuebingen.de, elnaz.shafaei-bajestan@uni-tuebingen.de, harald.baayen@uni-tuebingen.de
1 – ORCHID ID http://orcid.org/0000-0002-1060-3548
3 – ORCHID ID http://orcid.org/0000-0003-3178-3944

ABSTRACT

Indonesian has two prefixes which express a range of semantic functions (e.g. agent, instrument, patient). One prefix, *PEN*-, has six allomorphs *(peng-, peny-, pe-, pen-, pem-, penge-)*. A second prefix, *PE*-, is described as having similar form and meaning as *pe*-. In this study, we used computational models of distributional semantics to clarify whether *PE*- and *PEN*- have discriminable semantics. The cosine similarity measure was used to evaluate to what extent the semantic vectors of pairs of words are similar in meaning. We found that the semantic similarities within the *PEN*- words are higher than between *PE*- and *PEN*- words. Additionally, nouns with *PE*- are more similar to their base words compared to nouns with *PEN*-. Furthermore, semantics similarity rating results, on a 5-point Likert scale, gathered from native speakers of Indonesian are in agreement with model predictions.

Keywords: Indonesian morphology, distributional semantics, allomorphy, similarity judgments

Acknowledgements: This study was funded by the Indonesian Endowment Fund for Education (*Lembaga Pengelola Dana Pendidikan*) (No. PRJ-1610/LPDP/2015) to the first author, and an ERC advanced grant (No. 742545) to the last author.

1. INTRODUCTION

Indonesian has two prefixes that create nouns from verbs. These prefixes express different semantic functions (e.g. agent, instrument, patient); a range of meanings similar to that found for the *-er* suffix in English. The first prefix is *PEN*-, in this notation, *N* denotes nasal assimilation. *PEN*- has several allomorphs *(peng-, peny-, pe-, pen-, pem-, penge-)* that are phonologically conditioned and complementary distributed. Interestingly, there is a second prefix, *PE*-, whose form and meaning is similar to *pe*-, but that does not exhibit nasal assimilation.

Several qualitative studies have addressed the formal regularities of these prefixes. However, as pointed out by Denistia [2], there has not been a consensus of whether *PE*- and *PEN*- are allomorphs or independent prefixes. In a recent quantitative study, Denistia and Baayen [3] argue that *PE*- and *PEN*- are independent prefixes with their own semantic specialization; *PE*- is somewhat productive in forming patients, whereas *PEN*- is fully productive creating instruments. In the present study, we used word embeddings from semantics vector space models [13] to further investigate whether *PE*- and *PEN*- have discriminable semantics.

2. MATERIALS

2.1. Indonesian lemmatized corpus

The main corpus used in this study was the Leipzig Corpora Collection [5], compiled from Indonesian on-line written sources dating from 2008 to 2012 [11]. It consists of 2,759,800 sentences, 50,794,093 word tokens, and 112,025 different word types. The MophInd parser [6] was used to obtain morphological analyses for these words. Its precision was 0.98 and its recall was 0.83 for *PE*- and *PEN*- words. The output of the parser was manually checked and corrected when necessary, based on the Indonesian online comprehensive dictionary (fourth edition) [1]. Using the results from MorphInd, we lemmatized the corpus to separate the bound morphs, prolexemes, particles, and number affixes, following Sneddon et al. [14]. Lemmatization with MorphInd involves splittting orthographic words into separate lemmas, for instance, *acaramu* is lemmatized into *acara kamu*, ‘your event’, and *kuajak* into *aku ajak*, ‘I invite’. The databases and the R scripts used to construct these databases are available on-line at http://bit.ly/IndSVM_MenLex. Finally, we excluded numbers, punctuation marks and the 15 highest frequent stop words in the corpus.

2.2. Modeling semantics

We use distributional semantics [17] to quantify semantic similarities between the *PE*- and *PEN*- words based on their distributional properties observed in the lemmatized corpus. According to distributional hypothesis Firth [4], the similarity of lemmas in terms of meaning and similarity of their linguistic contexts are in positive correlation [12, 10]. The representational framework and its computational modeling utilize vector representations from linear algebra, where the meaning of a lemma is denoted as a high-dimensional vector of its linguistic context in a very large corpus [15]. Different vector similarity measures may then be used to assess semantic similarity.

We associated each word in the corpus with a semantic vector (known in computational linguistics as a word embedding). The distributional vector representations of *PE*- and *PEN*- target words were extracted using word2vec [8] with the default parameter settings. The similarity of two words was measured with the cosine similarity measure using equation (1), which computes the cosine of the angle between the two corresponding context vectors as follows: let vectors *v* and *w* be two *n* dimensional vectors representing two lemmas. The cosine of the angle θ between *v* and *w* is defined as the inner product of the vectors, after being length-normalized. Thus, similarity is evaluated on the basis of the orientation of the vectors, and not on their lengths.

$$sim(V,W) = \cos(\theta) = \frac{V.W}{||V||\,||W||} \tag{1}$$

2.3. PePeN cossim database

We brought together 79 *PE*- and 877 *PEN*- words in a database and computed the cosine similarity values for all of the 418,034 possible combinations of word pairs, henceforth the PePeNCossim Database, some examples of which are listed in Table 1. The English translation provided in the table is for the reader's convenience only, as the translation is not available in the database. In addition, we also calculated the cosine similarity values between the derived words and their base words.

Table 1: Examples of entries in the PePeNCossim Database.

Lemma1	L1English	Lemma2	L2English	Cosine	PrefixL1	PrefixL2	BaseWordL1	BWL1English	BaseWordL1	BWL1English
petugas	officer	pemerintah	government	0.08	PE	PEN	tugas	task	perintah	command
petugas	officer	pemain	player	0.02	PE	PEN	tugas	task	main	to play
petugas	officer	peanggar	fencing athlete	0.07	PE	PE	tugas	task	anggar	fencing
peserta	participant	peanggar	fencing athlete	0.08	PE	PE	serta	to be together with	anggar	fencing
pemadat	junkies	pemerintah	government	-0.05	PEN	PEN	madat	to use drug	perintah	command
pemadat	junkies	pemain	player	0.05	PEN	PEN	madat	to use drug	main	to play

2.4. Semantic similarity judgments

We also conducted a semantic similarity judgment experiment via an online questionnaire to investigate whether our model predictions regarding semantic similarities between *PE*-, *PEN*- words and their base words are in agreement with human perceived similarity. Eighty-three native Indonesian speakers were asked to rate pairs of words with respect to similarity in meaning on a 5-point Likert scale [7], from 0 representing (no similarity) to 4 (perfect synonymy), following Miller and Charles [9]. We selected 48 noun base words that are attested with either prefix. Across prefixes, we controlled for the frequency of base and derived words, as well as model-predicted cosine similarity. We also provided the participants with an 'I do not know' option and removed those answers from the analysis. The subjects were free to rate and re-rate the pairs before submitting their answers.

3. RESULTS

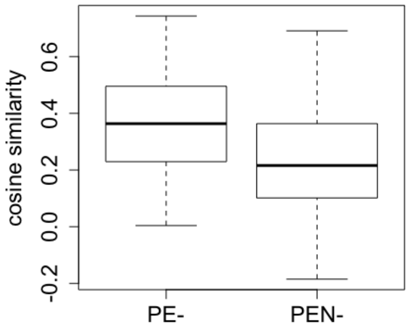
3.1. Comparing cosine similarities

Our first question is whether *PE*- and *PEN*- are distributed differently in terms of semantics. Accordingly, we grouped the database that contains all combinations of *PE*- and *PEN*- words as well as their cosine similarity values into 3 sets: set 1 contains pairs with one *PE*- and one *PEN*- word, set 2 contains pairs with two *PEN*- words and set 3 contains pairs with two *PE*- words.

Wilcoxon tests show that the cosine similarities among sets of groups are significantly different only for between prefix (between set 1 and set 2 (W = 8972200000, p < 2.2e − 16) and between set 1 and set 3 (W = 55310000, p < 2.2e − 16)). There is, however, no significant cosine different for set 2 and set 3 (W = 394120000, p = 0.003345). Within-prefix similarities were greater than between-prefix similarities. The observed high-level difference in semantics between *PE-*, which is less productive and semantically prefers agents (and some patients), and *PEN*-, which is more productive and creates agents or instruments, is thus complemented by a low-level difference as gauged with distributional semantics.

Figure 1 presents the distributions of cosine similarity values between noun base words and their prefixed derivatives, grouped by prefix: *PE*- (left) versus *PEN*- (right). Wilcoxon test shows that *PE*- words are, on average, significantly more similar to their base noun, compared to *PEN*- words (W = 13391, p = 3.103e-06). In addition, there was no significant difference between the similarity of derived words and their verb base words, as well as adjective base words (W = 5452, *p* = 0.5588).

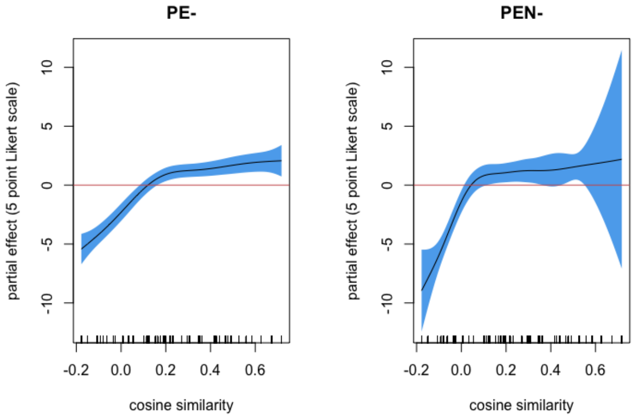
Figure 1: Cosine similarity for noun base and derived words is higher for *PE*- compared to *PEN*-.



3.2. Human judgments

Figure 2 shows the partial effect of the cosine similarity between derived words and their base words as predictor for the corresponding human similarity judgements, elicited on a 5-point Likert scale. We used a GAM (Generalized Additive Model, MGCV package version 1.8-17, Wood [16]) to model the human judgment rating as a function of prefix *PE*- or *PEN*-. Interestingly, perceived similarity increases with cosine similarity across the full range of values of the distributional measure in the case of *PE*-, whereas for *PEN*-, the cosine similarities are predictive only for the first quarter of its range. Apparently, *PE*- words are, on average, somewhat more distributionally related to their noun bases than is the case for *PEN*-.

Figure 2: Partial effect of cosine similarity on ratings for *PE*- (left) and *PEN*- (right). Model uncertainty is less for *PE*-.(Generalized additive mixed model fitted with mgcv using the ocat family for ordinal responses.)



4. CONCLUSIONS

In this paper, we addressed the question of whether Indonesian nominal prefixes *PE*- and *PEN*- have discriminable semantics. We used semantic vector space models to investigate the similarity of these two prefixes based on their distributional properties in a corpus of Indonesian.

Our results show that *PE*- and *PEN*- are somewhat differently positioned in Indonesian distributional space. The difference in mean similarity is small, but is supported statistically. Moreover, our model shows that *PE*- is somewhat more similar to its noun base, compared to *PEN*-. Furthermore, human judgments were found to be correlated more with cosine similarity for *PE*- as compared to *PEN*- . These results provide further evidence that *PE*- and *PEN*-, which in the literature on word formation in Indonesian have been described as being basically the same in meaning, express subtly different semantics and even semantically relate to their base words in slightly different ways. As argued by Denistia and Baayen [3], *PE*- and *PEN*- appear to be phonologically similar but functionally distinct prefixes.

REFERENCES

1. Alwi, H. (2012). *Kamus Besar Bahasa Indonesia*. Gramedia Pustaka Utama, Jakarta, fourth edition.
2. Denistia, K. (2018). Revisiting the Indonesian prefixes PEN-, PE2-, and PER-. *Linguistik Indonesia*, 36(2):145–159.
3. Denistia, K. and Baayen, R. H. (2019). The Indonesian prefixes PE- and PEN-: A study in productivity and allomorphy. *Morphology*, pages 1–23.
4. Firth, J. R. (1957). A synopsis of linguistic theory, 1930-1955. *Studies in linguistic analysis*.
5. Goldhahn, D., Eckart, T., and Quasthoff, U. (2012). Building large monolingual dictionaries at the leipzig corpora collection: From 100 to 200 languages. In *LREC*, volume 29, pages 31–43.
6. Larasati, S., Kubonˇ, V., and Zeman, D. (2011). Indonesian morphology tool (morphind): Towards an Indonesian corpus. *Systems and Frameworks for Computational Morphology*, pages 119–129.
7. Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, (140):1–55.
8. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In Burges, C. J. C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Sys- tems 26*, pages 3111–3119. Curran Associates, Inc.
9. Miller, G. A. and Charles, W. G. (1991). Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 1(28):1–28.
10. Pantel, P. (2005). Inducing ontological cooccurrence vectors. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 125–132. Association for Computational Linguistics.
11. Quasthoff, U., M.Richter, and C.Biemann (2006). Corpus portal for search in monolingual corpora. pages 1799–1802, Genoa. The Fifth Inter- national Conference on Language Resources and Evaluation.
12. Rubenstein, H. and Goodenough, J. B. (1965). Contextual correlates of synonymy. *Communications of the ACM*, 8(10):627–633.
13. Salton, G., Wong, A., and Yang, C.-S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620.
14. Sneddon, J. N., Adelaar, A., Djenar, D. N., and Ewing, M. C. (2010). *Indonesian: A Comprehensive Grammar*. Routledge, New York, second edition.
15. Turney, P. D. and Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, 37:141–188.
16. Wood, S. (2006). *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC.
17. Zellig, H. (1954). Distributional structure. *Word*, 10(23):146–162.