Quantitative studies on the Indonesian prefixes *PE-* and *PEN-*

Dissertation

zur

Erlangung des akademischen Grades

Doktor der Philosophie

in der Philosophischen Fakultät

der Eberhard Karls Universität Tübingen

vorgelegt von

Karlina Denistia

aus Yogyakarta, Indonesia

2020

Gedruckt mit Genehmigung der Philosophischen Fakultät
der Eberhard Karls Universität Tübingen

Dekan	: Prof. Dr. Jürgen Leonhardt
Hauptberichterstatter	: Prof. Dr. R. Harald Baayen
Mitberichterstatter	: Prof. Dr. Detmar Meurers
Mitberichterstatterin	: Prof. Dr. Laura A. Janda
Tag der mündlichen Prüfung	: 26. Oktober 2020

Tübingen, Universitätsbibliothek: TOBIAS-lib

For him, for her, for you, for it, for me,

for them, for you, and for us.

Publications

Four chapters of this dissertation are journal articles that are in part published (chapters 2 and 3), accepted (chapter 4), and submitted (chapter 5). Two published studies are not included in this dissertation as these studies address other topics in Indonesian morphology, and do not focus on the prefixes *PE-* and *PEN-*.

- Denistia, K. (2018). Revisiting the Indonesian prefixes peN-, pe2-, and per-. *Linguistik Indonesia*, 36(2):145–159.

- Denistia, K. and Baayen, H. (2018). PE- and PEN-: A corpus based analysis in allomorphy. In Sukamto, K. E., editor, *Prosiding Kongres Internasional Masyarakat Linguistik Indonesia (KIMLI) 2018*, pages 179–183. Masyarakat linguistik Indonesia..

- Rajeg, G. P. W., Denistia, K., and Rajeg, I. M. (2018). Working with a linguistic corpus using R: An introductory note with Indonesian negating construction. *Linguistik Indonesia*, 36(1):1–36.

- Denistia, K. and Baayen, H. (2019). The Indonesian prefixes PE- and PEN-: A study in productivity and allomorphy. *Morphology*, 29(3):385–407.

- Denistia, K., Shafaei-Bajestan, E., and Baayen, R. H. (2019). Semantic vector model on the Indonesian prefixes PE- and PEN-. In *Proceedings of The 11th International Conference on the Mental Lexicon, 1*.

- Rajeg, G. P. W., Denistia, K., and Musgrave, S. (2019). Vector space models and the usage patterns of Indonesian denominal verbs: A case study of verbs with men-, men-/-kan, and men-/-i affixes. *NUSA: Linguistic studies of languages in and around Indonesia*, 67(1):35–76.

- Denistia, K., Shafaei-Bajestan, E., and Baayen, H. (2020). Exploring semantic differences between the Indonesian prefixes PE- and PEN- using a vector space model. *Corpus Linguistics and Linguistic Theory*. Accepted pending minor revision.

- Denistia, K. and Baayen, H. (2020). Affix substitution in Indonesian: A computational modeling approach. *Linguistics: An Interdisciplinary Journal of the Language Sciences*. Submitted.

Acknowledgements

Above all, I would like to express my greatest gratitude to Allah Subhanahu wa ta'ala for Allah's blessings and everything Allah gave me. My deep gratitudes are also dedicated to my primary financial support, Indonesia Endowment Fund for Education (*Lembaga Pengelola Dana Pendidikan*) under grant number (No. PRJ-1610/LPDP/2015) and to ERC advanced grant 742545 awarded to my supervisor.

This dissertation would not have been possible without a number of individuals who provided assistance. I would like to express my gratitude towards these people, who I hope will continue to be a part of my academic and personal life for many years to come.

First of all, I am very grateful to Harald Baayen, my supervisor, who has taught, supported, and encouraged me so much in these years. Harald's patience, understanding, enthusiasm to learn Indonesian, knowledge, support, and academic creativity have been a continuous source of inspiration. He has contributed a lot to the content of this dissertation, whether by reading it and offering feedback, editing it, or discussing its issues and problems. I am fortunate to learn quantitative linguistics from one of the experts in this field today. In addition, I would also like to thank Harald for his personal support. His calm and positive attitude has been of great help, particularly in the first year of my doctoral student life. Furthermore, his full support, scientific freedom, and trust inspired me in many ways. I was never expecting that I had such an interesting topic for my PhD, and it now becomes the resource of ideas for my future research.

The thesis has benefited a lot from the comments and suggestions of my fellow linguists. In this connection, my thanks to my former and present colleagues in the Quantitative Linguistics group in Tuebingen: Yu-Ying Chuang, Maja Linke, Tino Sering, Fabian Tomaschek, Jessie Nixon, Elnaz Shafaei Bajestan, Motoki Saito, Ching-Chu Hendrix-Sun, Peter Hendrix, Kun Sun, Karen Beaman, Tian Shen, Petar Millin, Denis Arnold, and Michael Ramscar. The Quantitative Linguistics group provides a research environment that was both extremely inspiring and a lot of fun. I have been fortunate enough to learn what it means to be a linguist from a number of talented and learned individuals. I thank you all for a superb time, for being so generous to allocate your time to discuss and help me in many ways. I also sincerely appreciate Tineke Baayen-Oudshoorn, who was really helpful in taking care of administrative issues and always responded to my requests in cheerful manner. For these, I cannot possibly thank you enough, Tineke.

This work would have been unimaginable without my mom, Ani Indriyanti, for her encouragement from far, her constant love, care, and support in everything in the long last years. This thesis would have been much more difficult without the company of my little sister and her daughter, Karina Adistia and Kaina Shamsa Almayra.

Furthermore, I am indebted to Detmar Meurers for finding time to review my dissertation. Another gratitude is for Laura A. Janda, whose higher-level thoughts helped me obtain a broader perspective on the work presented here. I am also grateful to Melanie J. Bell and Benjamin V. Tucker for their feedback on my work during my defence.

Finally, I owe more than I can say to the various people who, knowingly or unknowingly, provided moral support, hugs, ears, encouragement, books, games, songs, movies, and various other sources of relaxation, diversion, and entertainment: Hatma Cakrakusuma, Bianca Marcia, Amora Emas Gempita, Clara Shinta Misbah, Arif Luqman, Andika Priadiputra, Juwita Metasari, Dara Kurniawan, Novindi Dwi Ratnawati, TT-friends, 4JK-friends, and Indonesian people of Tuebingen community. Thank you, guys. I would also like to specially thank Arum Perwitasari and Gede Primahadi Wijaya Rajeg for inspiring me to be a better linguist.

Abstract

The purpose of this dissertation is to conduct a systematic quantitative approach to analyse the Indonesian prefixes *PE-* and *PEN-*. The questions addressed in this dissertation are (1) whether the two similar prefixes *PE-* and *PEN-* are allomorph or not, (2) to what extent the semantics differences between these two prefixes could be captured by distributional semantics, and (3) how are *PE-* and *PEN-* learn by a computational model, namely Linear Discriminative Learning.

In order to answer the questions, I compiled a database containing 3090 words with *PE-* and *PEN-* from a corpus of written Indonesian. Using the productivity analyses, the data showed that *PEN-* is apparently more productive than *PE-*. The difference productivity also occurs in their semantic roles, inflectional variants, as well as frequency ratio between their corresponding verbal prefixes *BER-* and *MEN-*. This corpus–based research thus suggests that *PE-* and *PEN-* are two independent prefixes.

Furthermore, a distributional vector space model was applied to my data to clarify whether *PE-* and *PEN-* have discriminable semantics. Cosine similarities mean comparisons revealed that pairs consisting of words with *PE-* has a higher similarity than words with *PEN-*. Furthermore, nouns with *PE-* were more similar to their base words than was the case for words with *PEN-*. What drives the higher similarity for *PE-* than for *PEN-* is that *PE-* is specialised to create athletes. These findings provide a further quantitative evidence for treating *PE-* and *PEN-* as two independent prefixes.

Finally, I made use of a computational model, the 'discriminative lexicon' (DL) model, to investigate the learnability of *PE-* and *PEN-*. As nouns with *PEN-* is corresponding with verbs with *MEN-*, this last study focused on if there is a trade-off between the affix substitution *PEN-* and *MEN-* regularity in comprehension learning. The findings suggest that *PE-* is learned more robustly than *PEN-* for two main reasons. First, *PE-* words tend to be longer and hence have more discriminative triphones. Second, due to cue competition with *MEN-*, the prefixal triphones of *PEN-* are less effective cues than those of *PE-*. A new measure of functional load is also proposed to shed light the relative importance of the triphones in the prefixes.

To sum up, this dissertation clarifies that prefixes, which at first blush look like allomorphs, can have different qualitative and quantitative properties.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The Indonesian language (*Bahasa Indonesia*) is the official and national language of Indonesia. Having more than 17,000 islands, it is not surprising that Indonesia has more than 700 local languages, and is the second most linguistically diverse country (Fearnside, 1997; Nababan, 1991). Thus, the Indonesian language serves as a lingua franca for more than 300 ethnic groups, enabling communication between them.

Indonesian belongs to the Austronesian language family, also known as Malayo-Polynesian languages. These languages are characterised by a rich morphology. Some Austronesian languages use infixation frequently. For instance, in Thao (spoken in central Taiwan), an actor focus infix *-um-* (e.g., *kan* 'eating' - *k-m-an* /k-um-an/ 'to eat' and *lhiur* 'hook' - *lh-m-iur* /lh-um-iur/) has no fewer than eleven phonologically conditioned allomorphs (Blust, 2004). In addition, some other languages have productive reduplication. In Indonesian, reduplication is used to express the plural for nouns, and it realizes a range of semantics function on verbs and adjectives, including intensification and iteration (Sugerman, 2016; Chaer, 2008; Rafferty, 2002; Dalrymple and Mofu, 2012).

Various phenomena in Indonesian Morphology have attracted research. One of these phenomena is agent noun formation, which has been investigated by many researchers (Ramlan, 2009; Sneddon et al., 2010; Dardjowidjojo, 1983; Kridalaksana, 2008). Indonesian has

two nominal prefixes, *PE-* and *PEN-*. The *N* in *PEN-* denotes 'Nasal' as it has five nasalized allomorphs. For notational clarity, I write the prefixes in upper case (*PE-* and *PEN-*) and the allomorphs as subscripts (PEN_{peng}-, PEN_{pen}-, PEN_{pem}-, PEN_{pe}-, PEN_{peny}-, PEN_{penge}-). These prefixes can express agents, instruments, or patients (e.g., *tinju* 'punch' - *petinju* 'boxer' for *PE-* and *buka* 'open' - *pembuka* 'opener' for *PEN-*) (Sneddon et al., 2010).

An unsolved problem in Indonesian theoretical morphology is how to classify these two prefixes. Some have argued that they are allomorphs (Sneddon et al., 2010; Ramlan, 2009). *PE-* occasionally occurs in the same phonological context as the PEN_{pe}-. Both prefixes appear with base words which have /l/, /w/, or /r/ as onset. This similarity in form goes hand in hand with one similarity in meaning, as both prefixes can express agents. However, there are also formal and semantic differences, that have led other researchers to argue against these prefixes being allomorphs (Dardjowidjojo, 1983; Kridalaksana, 2007). An in-depth literature review about the morphological status of these prefixes is provided in Chapter 2.

In the course of my research on *PE-* and *PEN-*, I became more and more convinced that treating these prefixes as allomorphs is unhelpful. For instance, Baayen et al. (2013) studied two Russian prefixes *pere-* and *pre-* that look like allomorphs. However, the Russian prefixes have subtly different semantic uses, which these authors took as evidence against allomorphy. A systematic investigation of the range of meanings of these prefixes was an obvious first step, but in the existing literature, nearly the same limited set of examples appears time and again. I therefore decided to adopt the methods of corpus linguistics and quantitative linguistics to study *PE-* and *PEN-*.

Chapter 3 introduces the database of nouns with *PE-* and *PEN-* and their base words that I compiled from a corpus of written Indonesian, available as part of the Leipzig collection of corpora (Goldhahn et al., 2012). A quantitative survey clarified that, similar to the Russian prefixes *pere-* and *pre-*, the Indonesian prefixes have different semantic uses. Even though both express agents, *PE-* can also realize patients, whereas *PEN-* is often used for instruments. Furthermore, a quantitative analysis of the productivity of the Indonesian derived nouns and their (morphologically complex) base words revealed that, compared to *PEN-* and its allomorphs,

PE- is an outlier. In this chapter, I therefore concluded that *PE-* is not an allomorph of *PEN-*.

When compiling the database of words with *PE-* and *PEN-*, I assigned to each word its semantic role (agent, patient, instrument, ...) using a dictionary of Indonesian, (Alwi, 2012), and if necessary my own intuitions given that I am a native speaker. However, this classical philological approach can be complemented with quantitative methods from computational linguistics. Specifically, is it possible to discern semantic differences in the use of *PE-* and *PEN-* using word embeddings (also known as semantic vectors), using methods from distributional semantics, (Firth, 1957; Landauer and Dumais, 1997; Rubenstein and Goodenough, 1965; Pantel, 2005). Earlier work by Jalaluddin and Syah (2009) had already pointed to the importance of words in the context for understanding the different meanings of *PE-* and *PEN-*. Since distributional semantics gauges semantic similarity through similarity in contexts of use, Chapter 4 presents one of the very first explorations of distributional semantics for Indonesian morphology, using word embeddings that I derived, with the help of my colleague Elnaz Shafaei-Bajestan, from the 36 million word corpus of written Indonesian. Compared to the size of corpora used to train English word embeddings, the size of the Indonesian corpus is rather small. Chapter 4 shows that nevertheless some progress has been possible.

An interesting property of the *PE-* and *PEN-* prefixes is that they are paradigmatically related to base verbs with the prefixes *BER-* and *MEN-* through a process of affix substitution, a characteristic trait of the morphology of many Austronesian languages (Sneddon et al., 2010; Blust, 2004). Thus, we find word pairs such as *pedagang* 'seller' and *berdagang* 'to sell' for *PE-*, and for *PEN-*, we have *pengirim* 'sender' and *mengirim* 'to send something'. Since *PEN-* is more productive than *PE-*, and since *PEN-* is more similar to *MEN-* than *PE-* is similar to *BER-*, the question arises of whether this high similarity in form for *PEN-* and *MEN-* is helpful for learning. To address this question, in Chapter 5, I made use of the computational model of the discriminative lexicon (Baayen et al., 2019), and specifically its mappings between form and meaning based on linear discriminative learning. Contrary to my expectations, this model of subliminal error-driven learning clarified that the greater form similarity of *PEN-* and *MEN-* makes *PEN-* somewhat more difficult to learn. By means of a novel measure for the functional load of (tri)phones, I was able to show how *PEN-, PE-*, and *MEN-* shift higher functional loads

to different (tri)phones. The insights obtained with linear discriminative learning show that this quantitative tool can be useful for linguistic analysis, complementing other existing tools such as Analogical Modeling of Language (AML, Skousen, 1989) and Memory-Based Learning (TiMBL, Daelemans and Van den Bosch, 2005).

The final chapter of this dissertation presents the main findings and the conclusions that I have drawn on the basis of my research.

Bibliography

Alwi, H. (2012). *Kamus Besar Bahasa Indonesia*. Gramedia Pustaka Utama, Jakarta, fourth edition.

Baayen, R., Janda, L. A., Nesset, T., Dickey, S., Endresen, A., and Makarova, A. (2013). Making choices in Russian: Pros and cons of statistical methods for rival forms. *Russian Linguistics*, 37:253–291.

Baayen, R. H., Chuang, Y.-Y., Shafaei-Bajestan, E., and Blevins, J. (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*.

Blust, R. (2004). Austronesian nasal substitution: A survey. *Oceanic Linguist*, 43(1):73–148.

Chaer, A. (2008). *Morfologi Bahasa Indonesia (Pendekatan Proses)*. PT Rineka Cipta, Jakarta.

Daelemans, W. and Van den Bosch, A. (2005). *Memory-based language processing*. Cambridge University Press, Cambridge.

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

Dardjowidjojo, S. (1983). *Some Aspects of Indonesian Linguistics*. Djambatan, Jakarta.

Fearnside, P. M. (1997). Transmigration in Indonesia: Lessons from its environmental and social impacts. *Environmental Management*, 21(4):553–570.

Firth, J. R. (1957). *Studies in Linguistic Analysis*, chapter A synopsis of linguistic theory 1930–1955, pages 1–32. Basil Blackwell, Oxford.

Goldhahn, D., Eckart, T., and Quasthoff, U. (2012). Building large monolingual dictionaries at the Leipzig Corpora Collection: From 100 to 200 languages. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation*, pages 1799–1802.

Jalaluddin, N. H. and Syah, A. H. (2009). Penelitian makna imbuhan pen- dalam Bahasa Melayu: Satu kajian rangka rujuk silang. *GEMA Online Journal of Language Studies*, 9(2):57–72.

Kridalaksana, H. (2007). *Kelas Kata dalam Bahasa Indonesia*. Gramedia Pustaka Utama, Jakarta, second edition.

Kridalaksana, H. (2008). *Kamus Linguistik*. PT Gramedia Pustaka Utama, Jakarta, 4th edition.

Landauer, T. and Dumais, S. (1997). A solution to plato's problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*, 104(2):211–240.

Nababan, P. W. J. (1991). Language in education: The case of Indonesia. *International Review of Education*, 37(1):115–131.

Pantel, P. (2005). Inducing ontological co-occurrence vectors. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 125–132. Association for Computational Linguistics.

Rafferty, E. (2002). Reduplication of nouns and adjectives in Indonesian. *Papers from the Tenth Annual Meeting of the Southeast Asian Linguistics Society*, pages 317–332.

Ramlan, M. (2009). *Morfologi: Suatu Tinjauan Deskriptif*. CV Karyono, Yogyakarta.

Rubenstein, H. and Goodenough, J. B. (1965). Contextual correlates of synonymy. *Communications of the ACM*, 8(10):627–633.

Skousen, R. (1989). *Analogical Modeling of Language*. Kluwer, Dordrecht.

Sneddon, J. N., Adelaar, A., Djenar, D. N., and Ewing, M. C. (2010). *Indonesian: A Comprehensive Grammar*. Routledge, New York, second edition.

Sugerman (2016). *Morfologi Bahasa Indonesia: Kajian ke Arah Linguistik Deskriptif*. Penerbit Ombak, Yogyakarta.

Chapter 2

Revisiting the Indonesian prefixes *PEN-*, *PE-*, and *PER-*

This chapter has been published as Denistia, K. (2018). Revisiting the Indonesian prefixes peN-, pe2-, and per-. *Linguistik Indonesia*, 36(2):145-159. The notation for *PE-, PEN-* and *PER-* in this chapter is modified from the original notation in the published paper: *peN-*, *pe2*, and *per-*, for the sake of notation consistency in this dissertation.

Abstract

This paper presents a literature review on three nominalising prefixes in Indonesian: *PEN-*, *PE–*, and *PER-* whose function is to create agent, instrument, or patient (e.g. *tulis* 'to write' - *penulis* 'writer', *wisata* 'travel' - *pewisata* 'traveller' and *tapa* 'ascetic' - *pertapa* 'hermit'). The 'N-' in *PEN-* stands for 'nasal' due to its five nasalised allomorphs (e.g., PEN_{pen}-, PEN_{pem}-, PEN_{peng}-, PEN_{peny}-, PEN_{penge}-). However, there is one *PEN-* allomorph which is not nasalised, henceforth called PEN_{pe}-. *PE–*, the other prefix, is described as having similar in form and meaning as PEN_{pe}-. *PER-*, is described as the archaic nominalisation prefix. Some theorists believe that Indonesian nominalisation is derived from *PEN-* and *PER-* in which *PE-* belongs to *PER-*, some argued that it is formed from *PEN-* in which *PE-* is one of *PEN-* or *PER-* variant,

and some stated that nouns are derived from *PEN*-, *PE*- or *PER*-. *PEN*- is described as the most productive of the three prefixes and is believed to correlate with the verbal prefix *MEN*- (e.g. *menulis* 'to write' - *penulis* 'writer') with the process of affix substitution, whereas *PE*- is described as corresponding with the verbal prefix *BER*- (e.g. *berwisata* 'to travel' - *pewisata* 'traveller'). Thus far, there has been no consensus addressing whether *PE*- is the allomorph of *PEN*- or *PER*- or none of them. This paper will examine existing theories and research relevant to this issue.

Keywords: prefix, allomorphs, affix substitution

2.1 Introduction

Similar to the English -er nominalisations, Indonesian has *PEN*-, *PE*- and *PER*- as nominalising[1] prefixes to form agents, instruments or patients (e.g., *buka* 'open' - *pembuka* 'opener', *tinju* 'punch' - *petinju* 'boxer', *tapa* 'ascetic' - *pertapa* 'hermit'). *PEN*- has five nasalised allomorphs (e.g., PEN_{pen}-, PEN_{pem}-, PEN_{peng}-, PEN_{peny}-, PEN_{penge}-) and one non nasalised variant (PEN_{pe}-). The latter allomorph does not follow the nasalisation rule. Furthermore, PEN_{pe}- has similar phonological condition to the invariant *PE*-.

In most cases, a noun with *PEN*- expresses agents, causers, or instruments whereas form with *PE*- expresses patients or agents. However, when both *PEN*- and *PE*- attach to the same base, both prefixes create either similar or different semantics as listed in Table 2.1 Sneddon et al. (2010). Chaer (2008) added that *PE*- has a specific meaning that relates to a profession or athlete. *PER*-, in addition, is considered as an unproductive prefix (Dardjowidjojo, 1983; Benjamin, 2009).

[1] *PEN*- can function as an adjectival prefix, as in diam 'silent' - pendiam 'silent person' and malu 'shy' - pemalu 'shy person'. In this paper, I will focus more on the nominalisation to facilitate equal comparison with *PE*- and *PER*-.

Base Word	Base Translation	*PEN-*	*PEN-* Translation	*PE-*	*PE-* Translation	*PEN-* and *PE-* Semantic Role
sapa	to address	penyapa	addressor	pesapa	addressee	agent - patient
kasih	love	pengasih	lover	pekasih	love poison	agent - instrument
sakit	sick	penyakit	disease	pesakit	a person with a disease	causer - patient
tinju	punch	peninju	puncher	petinju	boxer	agent - athlete
selam	to dive	penyelam	someone who dives	peselam	diver	agent - athlete

Table 2.1: Examples of *PEN-* and *PE-* attaching to the same base word

Several theories have discussed these prefixes and classified them according to form, meaning, and corresponding verbal prefix with a process of an affix substitution. What it is meant by a corresponding noun-verb prefix with the affix substitution process is that prior to creating nouns with *PEN-* or *PE-*, a verb with *MEN-* or *BER-* should be formulated. For example, *bungkus* 'a wrap' could be derived into *pembungkus* 'wrapper' because the verb *membungkus* 'to wrap' exists. However, it would not be possible to derive *kotak* 'a square' into **pengotak* 'squarer' as the verb **mengotak* 'to square' does not exist. Both *PEN-* and *MEN-* has six allomorphs (PEN_{pen}-, PEN_{pem}-, PEN_{peng}-, PEN_{peny}-, PEN_{penge}-, PEN_{pe}- and MEN_{men}-, MEN_{mem}-, MEN_{meng}-, MEN_{meny}-, MEN_{menge}-, MEN_{me}-). Meanwhile, *PE-* has either *BER-* or *di-* as its corresponding verbal prefix (e.g., *petani* 'rice farmer' - *bertani* 'to farm' and *pesapa* 'addressee' - *disapa* 'to be addressed') (Sneddon et al., 2010; Ramlan, 2009; Chaer, 2008; Putrayasa, 2008; Dardjowidjojo, 1983; Benjamin, 2009; Ermanto, 2016; Subroto, 2012; Sugerman, 2016).

Non-native Indonesians may find it difficult to differentiate PEN_{pe}- and *PE-* because they occur in the same phonological environment. The only way to distinguish them is by relating the base word to their verbal affix substitution. For example, PEN_{pe}- and *PE-* both appear before /l/ initial phoneme in base words *lari* 'to run' and *lukis* 'to paint'. From those base words, native speaker of Indonesia can tell that *pelari* 'runner' is *PE-* because it corresponds to the verb *berlari* 'to run', whereas *pelukis* 'painter' is PEN_{pe}- as it corresponds to *MEN-* verb, *melukis* 'to paint' (Chaer, 2008).

Interestingly, Indonesian works of literature have different consensus to classify the nominalising prefixes from *PEN-*, *PE-* and *PER-*. Firstly, (Dardjowidjojo, 1983) and (Kridalaksana, 2007) do not distinguish *PE-* and *PEN-* as they group both prefixes as *PE-*. They argued that there are two prefixes creating nouns in Indonesian and those are *PE-* (with *PEN-* included

in it) and *PER*-. Secondly, (Chaer, 2008; Subroto, 2012; Putrayasa, 2008) and (Ermanto, 2016) stated that *PE*- is the variant of *PER*- and that they are related (e.g., *PE*- as in *petapa* is derived from the deleted /r/ in *PER*- as in *pertapa*, both of which mean 'hermit'). Accordingly, they believed that *PE*- and *PER*- should not be treated as one prefix and thus the Indonesian nominalisation is formed by *PEN*- and *PE*-. Thirdly, (Benjamin, 2009) claimed that the nominalisation is formulated by prefixes *PEN*- and *PER*- in which *PE*- belongs to *PER*- due to its transformation from the archaic to the more common form. Fourthly, (Sneddon et al., 2010) and (Ramlan, 2009) believed that Indonesian nominalising prefixes consist of *PEN*-, *PE*- and *PER*-, all of which are invariants on the basis that *PER*- is the unproductive nominalising prefix.

Regarding this unclear classification, I compiled previous research on nominalisation with *PEN*-, *PE*- and *PER*-, including their meaning and corresponding verbal prefix. The purposes of this paper are therefore to examine the theories related to the classification of *PEN*-, *PE*- and *PER*-. In the following sections, I will cover nominalisation with *PEN*-, *PE*-, the overlapping *PEN*- and *PE*-, nominalisation with *PER*-, relevant discussion, possible further research, and some concluding comments.

2.2 Nominalisation with *PEN*-

PEN- is one of the most productive nominalising prefixes that can be attached to a noun, adjective or verb to express agent, causer or instrument (Sneddon et al., 2010; Ramlan, 2009; Chaer, 2008; Rajeg, 2013; Benjamin, 2009; Putrayasa, 2008; Dardjowidjojo, 1983; Ermanto, 2016; Subroto, 2012; Sugerman, 2016). Table 2.1 lists some examples for *PEN*- that are derived from adjective, noun or verb with a different semantic role for the nouns.

As shown in Table 2.2, *PEN*- transforms into allomorphs such as PEN_{pem}- as in *pemalsu*, PEN_{peny}- as in *penyakit*, and PEN_{peng}- as in *penguap*. Sneddon et al. (2010); Sugerman (2016); Ramlan (2009); Chaer (2008); Putrayasa (2008) and (Ermanto, 2016) characterised the occurrences of *PEN*- allomorphs as follows:

1. *-N* becomes *-ng* before vowels a,i,u,e,o and with the initial g, k, h, kh

   peN- + *olah* ‘to cultivate’ → *pengolah* ‘cultivator’

   peN- + *urus* ‘to look after’ → *pengurus* ‘committee’

   peN- + *goda* ‘to flirt’ → *penggoda* ‘who flirts’

   peN- + *hancur* ‘to destroy’ → *penghancur* ‘destroyer’

   peN- + *khianat* ‘to betray’ → *pengkhianat* ‘traitor’

2. *-N* becomes *-m* with initial b, p, f

   peN- + *beli* ‘to buy’ → *pembeli* ‘buyer’

   peN- + *fitnah* ‘to salnder’ → *pemfitnah* ‘who slanders’

3. *-N* becomes *-n* with initial d, t, c, j, sy, z

   peN- + *dengar* ‘to listen’ → *pendengar* ‘listener’

   peN- + *cari* ‘to seek ’ → *pencari* ‘seeker’

   peN- + *tolak* ‘to reject’ → *penolak* ‘who rejects’

   peN- + *jajah* ‘to colonialize’ → *penjajah* ‘colonizer’

4. *-N* becomes *-ny* with initial s

   peN- + *sewa* ‘to rent’ → *penyewa* ‘who rents’

5. *-N* is lost before initial l, r, m, n, ng, ny, w, y

   peN- + *lamar* ‘to propose’ → *pelamar* ‘who proposes’

   peN- + *ramal* ‘to forecast’ → *peramal* ‘fortune teller’

   peN- + *warna* ‘to color’ → *pewarna* ‘which gives color’

peN- + *masak* ‘to cook’ → *pemasak* ‘chef’

peN- + *nyanyi* ‘to sing’ → *penyanyi* ‘singer’

6. *penge-* occurs in a single syllable base

peN- + *bom* ‘bomb’ → *pengebom* ‘bomber’

Base Word	Base Translation	Noun Word	Noun Translation	Base Word Class	Semantic Role
palsu	fake	pemalsu	counterfeiter	adj	agent
panas	hot	memanaskan	to heat	pemanas	heater
adj	instrument				
sakit	sick			penyakit	disease
adj	causer				
pancing	fishing rod	pemancing	fisherman	n	agent
uap	steam	penguap	steamer	n	instrument
pantau	to observe	pemantau	observer	v	agent
baca	to read	pembaca	reader	v	instrument

Table 2.2: Examples of *PEN-* attached to a different base word class to express a different semantic role

However, Sneddon et al. (2010) list some exceptions, stating that bases with the initial /p/, /t/, /s/, or /k/ are not assimilated if the stem is loaned from other languages. When the loaned word is more widely accepted as Indonesian, two formations can be found; for example, the Arabic-borrowed stem *terjemah* ‘to translate’, has *penerjemah* and *penterjemah* ‘translator’ as its derived nouns. In few cases, *PEN-* nouns occur in two different ortographical realisations with same meaning (e.g., *pesaing - penyaing* ‘competitor’, *pecinta - pencinta* ‘lover’, *pengrajin - perajin* ‘crafter’).

Nouns with *PEN-* are described as having a corresponding verbal prefix with *MEN-* (e.g., *pembuka* ‘opener’ is derived from *membuka* ‘to open’) (Benjamin, 2009; Tjia, 2015). Table 2.3 shows that one of *PEN-* allomorphs is characterised by a process of affix substitution with one of *MEN-* allomorphs (Verhaar, 2010; Sneddon et al., 2010; Ramlan, 2009). Verbs with *MEN-* can be extended to circumfixes *MEN-/-kan* or *MEN-/-i* to realise causative (e.g., *panas* ‘hot’ - *memanaskan* and *memanasi* ‘to make something hot’) or beneficiary semantics (e.g., *ajar* ‘to teach’ - *mengajarkan* and *mengajari* ‘to teach to someone’) Kroeger (2007); Sutanto (2002). The structures with *MEN-/-kan* and *MEN-/-i* requires a goal, a patient, a beneficiary, a theme, a location, or an instrument as an argument (Arka et al., 2009; Sutanto, 2002; Toma-

sowa, 2007). Furthermore, *-i* expresses iterative (e.g., *lempar* 'to throw' - *melempari* 'to throw repeatedly'), applicative (e.g., *kirim* 'to send' - *mengirimi* 'to send to someone'), or intensifier semantics (e.g., *pukul* 'to hit' - *memukuli* 'to hit over and over again') (Arka et al., 2009; Tomasowa, 2007). However, derived nouns with *PEN-* do not carry the *-i* or *-kan* suffixes, even though semantically they may correspond to verbs with these suffixes. For example, *pemanas*, 'heater' is paradigmatically related to *memanaskan* 'to heat', and not related to the verb *memanas* 'to become hot'.

Base Word	Base Translation	Base Word Class	Verb Word	Verb Translation	Noun Word	Noun Translation	Semantic Role
palsu	fake	adj	memalsukan	to falsify	pemalsu	counterfeiter	agent
panas	hot	adj	memanaskan	to heat	pemanas	heater	instrument
sakit	sick	adj			penyakit	disease	causer
pancing	fishing rod	n	memancing	to fish	pemancing	fisherman	agent
uap	steam	n	menguapi	to steam	penguap	steamer	instrument
pantau	to observe	v	memantau	to observe	pemantau	observer	agent
baca	to read	v	membaca	to read	pembaca	reader	instrument

Table 2.3: Examples of corresponding *MEN-* and *PEN-*

2.3 Nominalisation with *PE-*

PE- is described by Sneddon et al. (2010) and Ramlan (2009) as invariant of *PEN-*. Table 2.4 lists some examples of *PE-* attaching to nouns, verbs, or adjectives to express agents, instruments or patients (Sneddon et al., 2010; Ramlan, 2009). As shown in the table, *PE-* does not follow *PEN-*.'s nasalisation rules. As I mentioned in Section 2.2, *PEN-* becomes PEN_{pen}- when it attaches to the stem initialised by /j/, as in *penjajah* 'colonizer'. However, Indonesian has *pejalan* 'pedestrian' and *pejuang* 'fighter', instead of **penjalan* and **penjuang* (see Table 2.4). This is the essential difference between *PEN-* and *PE-*; that *PE-* is not following the nasalisation rules of *PEN-* (Sneddon et al., 2010; Ramlan, 2009; Putrayasa, 2008).

Furthermore, *PE-* attaches to verbs with the prefix *BER-* and *di-* by a process of affix substitution as shown in Table 2.5 (Sneddon et al., 2010; Verhaar, 2010; Putrayasa, 2008). However, Ramlan (2009) mentioned that only several verbs with *BER-* correlate to *PE-*. *BER-*, which has BER_{be}- and BER_{bel}- as infrequent allomorphs, primarily creates verbs expressing reciprocity, reflexivity, or stativity (Kridalaksana, 2007; Ramlan, 2009; Putrayasa, 2008; Chaer, 2008; Sneddon et al., 2010). In addition, Tjia (2015) noted that *BER-* is a middle prefix ex-

Base Word	Base Translation	Noun Word	Noun Translation	Base Word Class	Semantic Role
sakit	sick	pesakit	sick person	adj	patient
tualang	adventure	petualang	adventurer	adj	agent
jalan	road	pejalan	pedestrian	n	agent
kasih	love	pekasih	love poisson	n	instrument
sapa	greeting	pesapa	addressee	n	patient
tanda	command	petanda	signified	n	patient
juang	to fight	pejuang	fighter	v	agent
lari	to run	pelari	runner	v	agent

Table 2.4: Examples of *PE-* attaching to different base word class to create a different semantic role

pressing an intransitive verb, especially for emotion and position (e.g., *berlari* '(in the process of) running' or *bersakit* '(in the process of being) sick').

Ber- can be extended with the suffixes *-kan* and *-an*. A verb with *BER-/-kan* and *BER-/-an* circumfixes express 'having X' (e.g., *dasar* 'base' - *berdasarkan* 'on the basis of') or reciprocative (e.g., *gandeng* 'to hold hand' - *bergandengan* 'to hold hands with each other'), respectively (Sneddon et al., 2010). *Di-* is a prefix used to create passive construction and can be extended to the suffix *-kan* and *-i*. It has also been a common knowledge that *MEN-* and *di-* are highly correlated due to their respective function as active and passive verbal prefixes (e.g., *mengirim* 'to send', *dikirim* 'to be sent', *memanaskan* 'to make something hot' - *dipanaskan* 'to be made hot' and *melempari* 'to throw repeatedly' - *dilempari* 'to be thrown by something repeatedly' (Sneddon et al., 2010; Ramlan, 2009; Kridalaksana, 2007; Putrayasa, 2008; Dardjowidjojo, 1983; Chaer, 2008; Benjamin, 2009; Ermanto, 2016; Subroto, 2012; Sugerman, 2016). Although the corresponding *BER-* and *di-* have *-i*, *-an*, or *-kan* suffix extension, derived nouns with *PE-* are paradigmatically related to verbs that do not carry the *-i* or *-kan* suffixes. For example, *petaruh* 'bidder', is related to the verb *bertaruh* 'to bid' and not to **bertaruhkan* or **bertaruhan*.

2.4 Overlapping *PEN-* and *PE-*

In some cases, *PEN-* and *PE-* can occur in similar phonological condition. Moreover, both of them could also attach to the same base words. The question then arises on how to differ-

Base Word	Base Translation	Base Word Class	Verb Word	Verb Translation	Noun Word	Noun Translation	Semantic Role
sakit	sick	adj			pesakit	sick person	patient
tinggi	high	adj			petinggi	high officials	agent
tualang	adventure	adj	bertualang	to have an adventure	petualang	adventurer	agent
jalan	road	n	berjalan	to walk	pejalan	pedestrian	agent
kasih	love	n			pekasih	love poisson	instrument
kebun	garden	n	berkebun	to do gardening	pekebun	gardener	agent
kerja	work	n	bekerja	to work	pekerja	worker	instrument
sapa	greeting	n	disapa		pesapa	addressee	patient
tanda	command	n	bertanda	to have sign	petanda	signified	patient
juang	to fight	v	berjuang	to fight	pejuang	fighter	agent
lari	to run	v	berlari	to run	pelari	runner	agent

Table 2.5: Examples of corresponding *BER-* or *di-* and *PE-*

entiate *PEN-* and *PE-* when they occur in identical phonological environment. Chaer (2008) and Ramlan (2009) explained two analogical processes on how to differentiate *PEN-* and *PE-* formations. The first is that when these prefixes attach to the same base word, *PEN-* and *PE-* will form an agent-patient relationship as in *penyuruh* 'commander' - *pesuruh* 'who is commanded'. This analogical process then creates some others agent-patient paradigm between *PEN-* and *PE-* (e.g., *penatar* 'speaker in a seminar' - *petatar* 'participant in a seminar', *penyuluh* 'person who gives information' - *pesuluh* 'person who is given information', *pengubah* 'changer' - *peubah* 'which is changed'). Secondly, due to the existence of *petinju* 'boxer', words for certain sports tend to use forms with *PE-*, such as *pegolf* 'golfer', *petembak* 'shooter (athlete)' and *petenis* 'tennis player'. This theory provides a reasonable explanation as to why both *PEN-* and *PE-* attach to the same stem (e.g., *tinju* 'to punch' - *petinju* 'boxer' - *peninju* 'someone who punches', *tembak* 'to shoot' - *petembak* 'shooter (athlete)' - *penembak* 'someone who shoots', *selam* 'to dive' - *peselam* 'diver (athlete)' - *penyelam* 'someone who dives', *terjun* 'to skydive' - *peterjun* 'skydiver (athlete)' - *penerjun* 'someone who sky dives' and *dayung* 'to paddle' - *pedayung* 'paddler (athlete)' - *pendayung* 'someone who paddles') that *PE-* is semantically more specific to the athlete of the sport.

Sneddon et al. (2010) and Benjamin (2009) added that in cases where *PEN-* and *PE-* occur with the same base, thus they express similar meanings (e.g., from *sulap* 'magic' - *pesulap* and *penyulap*, both mean 'magician'). There are also cases in which *PE-* and *PEN-* emerge within the same stem and reflect different semantics. A form with *PEN-* expresses agent, causer, or instrument whereas a form with *PE-* expresses patient or agent (e.g., *siar* 'to announce/to sail' - *penyiar* 'radio announcer' - *pesiar* 'a cruise ship' and *tanda* 'sign' - *penanda*

'a sign' - *petanda* 'a hint', *ajar* 'to teach' - *pengajar* 'teacher' - *pelajar* 'student', *tempur* 'to combat' - *penempur* 'armament' - *petempur* 'combatant').

Sawardi (2015) endorsed the analogical process between the agentive *PEN-* and the patient *PE-* and further concluded that this phenomenon is a measurement of the transitiveness of a verb. Sawardi (2015) also stated that *PEN-* can be an indicator of ergativity in Indonesian. He claimed that if an intransitive verb can be nominalised using *PE-*, then the subject argument needed in the syntactical structure will be an agent (e.g., *berenang* 'to swim' - *perenang* 'swimmer'). His main point was that all *PE-*, regardless of whether it corresponds to *BER-*, is considered *PEN-* because it is derived from an intransitive verb. Thus, unlike other theories which state that *pekerja* 'worker', *pelari* 'runner', *perenang* 'swimmer', *pelayar* 'sailor' are *PE-*, in Sawardi (2015), these words are *PEN-*. He only categorised *PE-* as those whose semantic role is that of patient (e.g., *petatar* 'participant in a seminar', *pesuluh* 'person who is given information'). This claim, however, was applied only to a small amount of data. Besides, *PEN-*'s function as an instrument was not discussed.

2.5 Nominalisation with *PER-*

PER- is a nominalising prefix forming agents or patients. Compared to *PEN-* and *PE-*, which are productive in creating nouns, *PER-* is a non-productive nominalising prefix (Dardjowidjojo, 1983; Ramlan, 2009). There are only a few examples of nouns with this prefix (e.g., *tapa* 'to live as an ascetic' - *pertapa* 'hermit', *segi* 'angle' - *persegi* 'square', *antara* 'between' - *perantara* 'mediator', *tanda* 'sign' - *pertanda* 'a sign', *lambang* 'symbol' - *perlambang* 'symbol').

There are two views as to whether *PE-* and *PER-* are different. The first perceives *PER-* as invariant from *PE-* which means they need to be treated as two different prefixes (Benjamin, 2009; Sneddon et al., 2010; Ramlan, 2009). The basic premise that makes *PER-* different from *PE-* is that *PER-* is unproductive, somewhat archaic, and limited to only a few words. The second view treats *PER-* as a form similar to *PE-* (Putrayasa, 2008; Subroto, 2012; Chaer, 2008; Ermanto, 2016). Putrayasa (2008) argued that the /r/ deletion in *PER-* to become

PE- is a diachronic process. Subroto (2012) and Ermanto (2016) also stated that both *PE-* and *PER-* are derived from the verbal prefix *BER-* (e.g., *bertapa* 'to do ascetic' - *pertapa* 'hermit' and *berdagang* 'to trade' - *pedagang* 'trader').

PER- can also function as a causative prefix (e.g., *besar* 'big' - *perbesar* 'to make bigger' and *istri* 'wife' - *peristri* 'to make her a wife') (Ramlan, 2009; Rajeg, 2013). I will not discuss the causative *PER-* due to its function as a verbal prefix, although Benjamin (2009) stated that the agent and causative *PER-* might have a historical correlation as in *pejalan* 'pedestrian' which was derived originally from causative *perjalan* and "seems to imply the replication of whatever it is that the agent *PE-* is doing or has in mind - which is an appropriate way to derive a 'causative' morphology".

Chaer (2008) elaborated further on *PER-* allomorphs as follows:

1. *-r* is omitted before *-r* or if the first syllable contains *-er-*

   PER- + *ringan* 'light' → *peringan* 'to make something lighter'

   PER- + *rendah* 'low' → *perendah* 'to make something lower'

   PER- + *runcing* 'sharp' → *peruncing* 'sharpener'

   PER- + *ternak* 'to farm' → *peternak* 'farmer'

   PER- + *kerja* 'to work' → *pekerja* 'worker'

2. *-r* becomes *-l* only with the stem *ajar* 'to study'

   PER- + *ajar* 'to study' → *pelajar* 'student'

3. *-r* appears elsewhere

   PER- + *kaya* 'rich' → *perkaya* 'to become richer'

   PER- + *kecil* 'small' → *perkecil* 'to make something smaller'

PER- + *lambat* ‘slow’ → *perlambat* ‘to make something slower’

PER- + *cepat* ‘fast’ → *percepat* ‘to make something faster’

However, Chaer (2008)’s formulation for the phonological condition for *PER-* is somewhat confusing. In his *PER-* examples above, he included the instrument *PEN-* as in *peruncing* ‘sharpener’, agent *PE-* as in *pekerja* ‘worker’, and causative *PER-* as in *perkaya* ‘to become richer’.

2.6 Discussion

There are three possible classifications of the nominalising prefix in Indonesian using *PEN-*, *PE-* and *PER-*. The first classification states that Indonesian nouns could be formed by attaching *PEN-*, *PE-* and *PER-* prefixes (Sneddon et al., 2010; Sugerman, 2016; Ramlan, 2009). The second classifies that nouns are only derived from the prefix *PE-* and *PER-*, whereas *PEN-* is the variant of *PE-* (Dardjowidjojo, 1983; Kridalaksana, 2007). If it is true that *PE-* and *PEN-* are the same prefixes, in my opinion, it needs to be reconsidered because they are not in a complementary distribution. Indonesian has a condition, in which *PE-* and *PEN-* attach to the same base words (e.g., *ubah* ‘to change’ - *pengubah* ‘changer’ - *peubah* ‘which is changed’ and *tinju* ‘to punch’ - *petinju* ‘boxer’ - *peninju* ‘someone who punches’) Chaer (2008); Ramlan (2009). The final classification was given by Putrayasa (2008); Chaer (2008); Subroto (2012); Alwi et al. (2003), and Ermanto (2016) to treat *PER-* as a similar form of *PE-* due to their shared characteristics. Researchers therefore believed that *PE-* is the modern version of *PER-* as both are related to *BER-* (e.g., *pertapa* ‘hermit’ - *bertapa* ‘to do ascetic’ vs. *petani* ‘rice farmer’ - *bertani* ‘to farm’). However, if it is the case that *PER-* and *PE-* are the same prefix from a diachronic perspective, I should be able to find two forms showing a transformation, such as *pertapa* to *petapa*, meaning ‘hermit’, and both forms would be acceptable. In fact, forms such as *petani* ‘rice farmer’ or *petinju* ‘boxer’ do not show any transformation at all; there are no **pertani* or **pertinju*. Considering these different arguments among linguists, I argue that there

is still no clear consensus as to what constitutes the major nominalising categories in the Indonesian language.

Dardjowidjojo (1983) mentioned that a new formation through the process of analogy, as proposed by Chaer (2008), makes *PEN-* the most productive prefix. Given that *PEN-* is claimed to be the most productive nominalising prefix and *PER-* as the unproductive one, a question arises regarding the general use of the term productivity, which has not yet been well defined. Indeed, studies on the productivity of word formation have provided solutions to questions related to morphology in both written and spoken language, context-governed spoken language, and everyday conversations (Baayen, 1993; Baayen and Lieber, 1991; Baayen and Renouf, 1996; Baayen and Neijt, 1997; Plag, 1999). In the cases of *PEN-*, *PE-* and *PER-* prefixes, it is not clear which definition of productivity is being used by Dardjowidjojo (1983). Furthermore, Kridalaksana (2007) and Ramlan (2009) claimed that a formation can be more productive than others. However, they do not state whether the productivity parameter is based on the frequency of usage, new formation, or even its regularity (e.g., their process of analogy) in the nominalisation.

In addition to *PEN-* allomorphs' phonological condition, I notice that the theories do not describe the phonological condition because it is the first letter of the stem typography and has nothing to do with either place or manner of articulation. Overall, it can be concluded that:

1. *PENpeng-* occurs when it is combined with stem initialized by vowels, velar-stop (e.g., /g/, /k/), velar fricative (e.g., /h/) and uvular fricative (e.g., /χ/) consonants

2. *PENpem-* occurs when it is combined with stem initialized by bilabial stop (e.g., /b/, /p/) and voiceless labio–dental (e.g., /f/) consonants

3. *PENpen-* occurs when it is combined with stem initialized by alveolar stop (e.g., /d/, /t/) and alveolar fricative (e.g., /tʒ/, /dʒ/, /ʃ/, /ʒ/) consonants

4. *PENpeny-* occurs when it is combined with stem initialized by alveolar fricative (e.g., /s/) consonant

5. *PENpe-* occurs when it is combined with stem initialized by nasal (e.g., /m/, /n/, /ɲ/, //), glide (e.g., /w/, /j/) and liquid (e.g., /r/, /l/) consonants

6. *PENpenge-* occurs whenever *PEN-* attaches to single syllable stem

A problem arises when distinguishing between *PE-* and PEN_{pe}- as the allomorph of *PEN-* because both formations can appear in the same phonological condition. For example, there may be confusion around whether the word *pelatih* 'trainer' is *PEN-* or *PE-* as Indonesian has *melatih* 'to train' and *berlatih* 'to practice'. In this case, native Indonesian can use their intuition and tell that *pelatih* is *PEN-* word as it correlates to the verb *melatih* 'to train' and not *berlatih* 'to practice'. This issue regarding the overlapping phonological condition between *PEN-* and *PE-* has not been addressed until now.

Base Word	Base Translation	Base Word Class	Noun Word	Noun Translation	PE-	PEN-	Allomorph PEN-	Semantic Role	Verb Word	Verb Translation
lari	to run	v	pelari	runner	TRUE	FALSE		agent	berlari	to run
musik	music	n	pemusik	musician	TRUE	FALSE		agent	bermusik	to play music
runding	discussion	n	perunding	who are in discussion	TRUE	FALSE		agent	berunding	to have a discussion
wisata	to travel	v	pewisata	traveller	TRUE	FALSE		agent	berwisata	to travel
lukis	to paint	v	pelukis	painter	FALSE	TRUE	pe	agent	melukis	to paint
minta	to ask for	v	peminta	demander	FALSE	TRUE	pe	agent	meminta	to ask for
rintis	pioneer	n	perintis	pioneer	FALSE	TRUE	pe	agent	merintis	to pioneer
wawancara	interview	n	pewawancara	interviewer	FALSE	TRUE	pe	agent	mewawancara	to interview

Table 2.6: Examples of *PEN-* and *PE-* occuring in the same phonological condition

2.7 Future research

Conducting a corpus-based study on these prefixes is undoubtedly feasible. There is a large Indonesian corpus that forms part of the Leipzig Corpora Collection at `https://www.r-project.org/conferences.html` which comprises a variety of written registers (the web, newspapers, Wikipedia) dating from the years 2008 - 2012 (Goldhahn et al., 2012). With a total dataset of 36.608.669 word tokens from the corpus, productivity can be measured. Moreover, it may be possible to support qualitative theories using this quantitative data. From this corpus, we could run MorphInd, the Indonesian morphological parser Larasati et al. (2011), to compile all the possible *PEN-*, but not *PE-* and *PER-*. Table 2.7 shows that MorphInd identifies some words correctly, such as *perintis* 'pioneer', *pelukis* 'painter', *pewawancara* 'interviewer', and *peminta* 'demander' contain *PEN-* prefix. However, the parser is not able to identify *PE-* in *petapa*

'hermit', *pekerja* 'worker' and *pejalan* 'pedestrian'. MorphInd also misidentified *pelari* 'runner' and *pemusik* 'musician'. Thus, MorphInd lacks precision in identifying *PE-* and *PER-*. Hence, the output of the parser still needs to be manually checked and corrected. The lack of precision in identifying *PE-* and *PER-* generally occurs in some tools on stemming Indonesian (Suhartono et al., 2014; Asian et al., 2005; Adriani et al., 2007; Oktarino et al., 2016; Setiawan et al., 2016). However, work conducted by Pisceldo et al. (2008) distinguished between *PEN-* and *PER-* (*PE-* is included in *PER-*). All data preprocessing and analyses could be run in R (R Team, 2017, 2015), an open-source programming language for statistical computation.

Base Word	Base Translation	Noun Word	Noun Translation	Pe-	PeN-	Parser
rintis	pioneer	perintis	pioneer		TRUE	peN+rintis<v>_NSD
lukis	paint	pelukis	painter		TRUE	peN+lukis<v>_NSD
wawancara	interview	pewawancara	interviewer		TRUE	peN+wawancara<n>_NSD
minta	to ask for	peminta	demander		TRUE	peN+minta<v>_NSD
tapa	to do ascetic	pertapa	hermit			pertapaX−
minta	to ask for	peminta	demander		TRUE	peN+minta<v>_NSD
kerja	work	pekerja	worker	TRUE		pekerja<n>_NSD
jalan	road	pejalan	pedestrian	TRUE		pejalan<n>_NSD
lari	running	pelari	runner	TRUE		peN+lari<n>_NSD
musik	music	pemusik	musician	TRUE		peN+musik<n>_NSD

Table 2.7: Examples of the output of the MorphInd parser

Given that there is an issue in PEN_{pe}- and *PE-*, an experimental linguistics would be a useful way to address this issue, which are not yet resolved by theories. For example, studies conducted by Tomaschek et al. (2013, 2014) found that word frequency has a significant effect on vowel length, vowel quality, and vowel articulation in speech production. Specifically, they found that the higher the word frequency, the more the speaker will have language experience. This increases the proficiency of the speakers, enabling them to anticipate the tongue movement for high-frequency words. They also found differences in vowel realisations in high and low-frequency German words using articulography. For example, the higher the word frequency, the longer the articulation of long vowels and the shorter the articulation of short vowels. Regarding innovative application in experimental linguistics, it would be enlightening to see how *PEN-*, *PE-* and *PER-*, which are claimed to differ in productivity, are articulated differently by native Indonesians. In the experiment, word frequency from the corpus as well as base frequency and verbs with *MEN-* or *BER-* could be taken into account.

The affix substitution between nouns and verbs derivation can also be investigated using Blevins (2016) word in paradigm structure, in which ”the organisation of morphological system presupposes that words are construed as parts of patterns”. In Indonesian, it is generally known that *PEN-* and *PE-* have a paradigmatic relation with *MEN-* and *BER-* verbal prefixes, respectively. If it is indeed the case they are correlated, this offers a new approach to exploring the allomorphy given that both *PEN-* and *MEN-* have 6 allomorphs (e.g., PEN_{pen}-, PEN_{pem}-, PEN_{peng}-, PEN_{peny}-, PEN_{penge}-, PEN_{pe}- and MEN_{men}-, MEN_{mem}-, MEN_{meng}-, MEN_{meny}-, MEN_{menge}-, MEN_{me}-). This paradigm of *MEN-* and *PEN-* is regularly displayed in Indonesian. Such a paradigmatic relation is supported by Benjamin (2009) and Tjia (2015) who stated that *MEN-* is a very agentive and actor-oriented verbal prefix. However, they did not discuss in detail how *MEN-* and *PEN-* are paradigmatically correlated. They assumed that, because of the high agentivity of prefix *MEN-*, verbs with *MEN-* create subject nominalisation by substituting the prefix *PEN-*. This finding might be expanded to a hypothesis of the paradigmatic relation between *MEN-* and *PEN-*. The prediction is that if they are under the same paradigm, allomorphs in *PEN-* will mirror allomorphs of *MEN-*. From this, a new hypothesis can be tested; whether the productivity of the verbal prefix with *MEN-* is reflected through *PEN-* and, if so, is this also the case with *PE-* and *BER-*?

2.8 Conclusion

Theories about *PEN-*, *PE-*, and *PER-* provide many qualitative descriptions as to their form and meaning without any consensus on the classification of these prefixes. Among the theories reviewed, there were four classifications of the nominalising prefix in Indonesian: (1) *PE-* and *PER-*, (2) *PEN-* and *PE-*, (3) *PEN-* and *PER-*, and (4) *PEN-*, *PE-* and *PER-*. Furthermore, an issue arises when one of the *PEN-* allomorphs, PEN_{pe}-, cannot be distinguished from *PE-* due to their similar appearance in the phonological environment. Some researchers have discussed that when *PEN-* and *PE-* are in a contest, there are two ways to determine them. The first way is to ascertain which verbal prefix they correspond to; *PEN-* is with *MEN-* and *PE-* is with *BER-*. The second method is to check the availability of the analogical process underlying the

agent-patient semantic role between *PEN-* and *PE-*, or the athlete semantic specialisation which exists only in *PE-*.

Although there have been many qualitative descriptions and theories regarding these prefixes, some questions remain. Despite the debate on the morphological status of these nominalising prefixes, the measurement of productivity among these three prefixes is unclear. Therefore, further research on quantitative and experimental linguistics will provide new perspectives on Indonesian morphology. Corpus-based analyses as well as word frequency effect in sound production might be two possible forms of research that can be conducted in this respect. Furthermore, the new concept of word-in-paradigm can be used to analyse the verb-noun corresponding prefixes of *PE-* and *MEN-*, as well as *PE-* and *BER-*. Often these forms are used to help establish a lack of appropriate theories or to reveal that current theories are inadequate for explaining emerging research problems.

Acknowledgement

This study was funded by Indonesia Endowment Fund for Education (*Lembaga Pengelola Dana Pendidikan*) (No. PRJ-1610/LPDP/2015).

Bibliography

Adriani, M., Nazief, B., Asian, J., and Tahaghoghi, S. (2007). Stemming Indonesian: A confix–stripping approach. *ACM Transactions on Asian Language Information Processing*, 6(4):Article 13.

Alwi, H., Dardjowidjojo, S., Lapoliwa, H., and A.M., M. (2003). *Tata Bahasa Baku Bahasa Indonesia*. Balai Pustaka, Jakarta, 3rd edition.

Arka, I. W., Dalrymple, M., Mistica, M., and Mofu, S. (2009). A linguistic and computational morphosyntactic analysis for the applicative -i in Indonesian. In Butt, M. and King, T. H., editors, *International Lexical Functional Grammar Conference (LFG)*, pages 85–105. CSLI Publications.

Asian, J., Williams, H. E., and Tahaghoghi, S. M. M. (2005). Stemming Indonesian. In Estivill-Castro, editor, *The 28th Australasian Computer Science Conference (ACSC 2005)*, volume 38. Australian Computer Society, Inc.

Baayen, R. (1993). On Frequency, Transparency, and Productivity. In Booij, G. and van Marle, J., editors, *Yearbook of morphology*, pages 181–208. Kluwer.

Baayen, R. and Lieber, R. (1991). Productivity and English derivation: A corpus-based study. *Linguistics*, 29:801–844.

Baayen, R. and Renouf, A. (1996). Chronicling the times: Productive lexical innovations in an English newspaper. *Language*, 72:69–96.

Baayen, R. H. and Neijt, A. (1997). Productivity in context: A case study of a Dutch suffix. *Linguistics*, 35(3):565–587.

Benjamin, G. (2009). Affixes, Austronesian and iconicity in Malay. *Bijdragen tot de Taal-, Land- en Volkenkunde*, 165(2–3):291–323.

Blevins, J. P. (2016). *Word and paradigm morphology*. Oxford University Press, Oxford.

Chaer, A. (2008). *Morfologi Bahasa Indonesia (Pendekatan Proses)*. PT Rineka Cipta, Jakarta.

Dardjowidjojo, S. (1983). *Some Aspects of Indonesian Linguistics*. Djambatan, Jakarta.

Ermanto (2016). *Morfologi Afiksasi Bahasa Indonesia Masa Kini: Tinjauan dari Morfologi Derivasi dan Infleksi*. Kencana, Jakarta.

Goldhahn, D., Eckart, T., and Quasthoff, U. (2012). Building large monolingual dictionaries at the Leipzig Corpora Collection: From 100 to 200 languages. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation*, pages 1799–1802.

Kridalaksana, H. (2007). *Kelas Kata dalam Bahasa Indonesia*. Gramedia Pustaka Utama, Jakarta, second edition.

Kroeger, P. R. (2007). Morphosyntactic vs. morphosemantic functions of Indonesian –kan. In Zaenen, A., Simpson, J., King, T. H., Jane, G., Maling, J., and Manning, C., editors, *Architectures, Rules, and Preferences: Variations on Themes of Joan Bresnan*, number 184 in CSLI Lecture Notes, pages 229–251. CSLI Publications, Stanford, California.

Larasati, S., Kuboň, V., and Zeman, D. (2011). Indonesian morphology tool MorphInd: Towards an Indonesian corpus. In C., M. and M., P., editors, *Systems and Frameworks for Computational Morphology*, volume 100, pages 119–129. Springer.

Oktarino, A. B., Winahyu, D. T., Halim, A., and Suhartono, D. (2016). Generating affixed words from a root word and getting lemma from affixed word in Bahasa: Indonesian language. *International Journal of Knowledge Engineering*, 2(3):132 – 136.

Pisceldo, F., Mahendra, R., Manurung, R., and Arka, I. W. (2008). A two-level morphological analyser for the Indonesian language. In *In Proceedings of the 2008 Australasian Language Technology Association Workshop ALTA 2008*, pages 142–150.

Plag, I. (1999). *Morphological Productivity: Structural Constraints in English Derivation*. Mouton de Gruyter, Berlin.

Putrayasa, I. B. (2008). *Kajian Morfologi: Bentuk Derivasional dan Infleksional*. PT Refika Aditama, Bandung.

R Team, D. C. (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

R Team, S. (2015). *RStudio: Integrated Development for R. RStudio*. RStudio, Inc., Boston, MA.

Rajeg, G. P. W. (2013). Metonymy in Indonesian prefixal word formation. *Lingual: Journal of Language and Culture*, 1(2):64–81.

Ramlan, M. (2009). *Morfologi: Suatu Tinjauan Deskriptif*. CV Karyono, Yogyakarta.

Sawardi, F. (2015). Perilaku keterpilahan (split-S) Bahasa Indonesia. *Nuansa Indonesia*, XVII(1):36 – 44.

Setiawan, R., Kurniawan, A., Budiharto, W., Kartowisastro, I. H., and Prabowo, H. (2016). Flexible affix classification for stemming Indonesian language. *2016 13th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology ECTI-CON*, pages 1 – 6.

Sneddon, J. N., Adelaar, A., Djenar, D. N., and Ewing, M. C. (2010). *Indonesian: A Comprehensive Grammar*. Routledge, New York, second edition.

Subroto, E. (2012). *Pemerian Morfologi Bahasa Indonesia: Berdasarkan Perspektif Derivasi dan Infleksi Proses Afiksasi*. Yuma Pressino, Surakarta.

Sugerman (2016). *Morfologi Bahasa Indonesia: Kajian ke Arah Linguistik Deskriptif*. Penerbit Ombak, Yogyakarta.

Suhartono, D., Christiandy, D., and Rolando, R. (2014). Lemmatization technique in Bahasa: Indonesian language. *Journal of Software*, 9.

Sutanto, I. (2002). Verba berkata dasar sama dengan gabungan afiks men-i atau men-kan. *Makara, Sosial-Humaniora*, 6(2):82–87.

Tjia, J. (2015). Grammatical relations and grammatical categories in Malay: The Indonesian prefix men- revisited. *Wacana*, 16(1):105 – 132.

Tomaschek, F., Tucker, B. V., Wieling, M., and Baayen, R. H. (2014). Vowel articulation affected by word frequency. In *10th International Seminar on Speech Production*, pages 425–428.

Tomaschek, F., Wieling, M., Arnold, D., and Baayen, R. H. (2013). Word frequency, vowel length and vowel quality in speech production: An ema study of the importance of experience. In *INTERSPEECH*, pages 1302–1306.

Tomasowa, F. H. (2007). The reflective experiential aspect of meaning of the affix -i in Indonesian. *Linguistik Indonesia*, 25(2):83–96.

Verhaar, J. W. M. (2010). *Asas-Asas Linguistik Umum*. Gadjah Mada University Press, Yogyakarta.

Chapter 3

A study in productivity and allomorphy

This chapter has been published as Denistia, K. and Baayen, H. (2019). The Indonesian prefixes PE- and PEN-: A study in productivity and allomorphy. *Morphology*, 29(3):385-407.

Abstract

This study examines two nominalizing prefixes in Indonesian: *PE-* and *PEN-*, which derive nouns from verbs with a range of meanings similar to that found in *-er* suffix in English. The prefix *PE-* is form-invariant, whereas *PEN-* has several nasal allomorphs. Given their similarity in form and function, the question arises of whether *PE-* and *PEN-* are allomorphs. We conducted a corpus–based analysis of their productivity, using the written Indonesian corpus in the Leipzig Corpora Collection. In this corpus, *PEN-* is apparently more productive than *PE-*. Interestingly, the frequency of words with *PEN-* correlates significantly with the productivity of the corresponding base verbs. In addition, *PEN-* is more integrated into the verbal system; verbs that have *PEN-* are part of larger verb families. *PEN-* attaches almost exclusively to verbs and creates nouns denoting agents and instruments. By contrast, *PE-* creates nouns denoting agents and patients and attaches not only to verbs but also to nouns and adjectives. For derived words with *PE-*, there is no significant correlation between the frequency of the nominalization and the frequency of its base. *PE-* also does not participate in the linearity of the productivity

of the allomorphs of base and derived words that characterizes *PEN-*. Words with *PE-* are also more often input to further reduplication and inflectional variants than is the case for *PEN-*. This corpus–based research thus illustrates that affixes can have different qualitative and quantitative properties, although at first blush they look like allomorphs. Our analyses justify their treatment in the Indonesian literature as separate prefixes.

Keywords: Indonesian, productivity, productivity paradox, allomorphy, paradigmatics

3.1 Introduction

The question addressed in this study is whether two phonologically very similar prefixes of Indonesian are allomorphs or rather independent prefixes. According to the classical definition of allomorphy, variants of a morpheme which have the same underlying form, which share the same meaning, and are in complementary distribution, are classified as allomorphs (Bloomfield, 1933; Alber, 2011). When two different affixes express roughly the same semantics, they are referred to not as allomorphs but as rival affixes (Aronoff and Anshen, 2017). Conversely, when the same form signifies completely different semantic functions, as in the case for English *-s* (third person singular vs. plural inflection vs. third person genitive, ... Plag et al., 2017), we have affix homonymy. Less clear-cut are cases where formatives are obviously similar in form as well as in meaning, without the form similarity being phonologically conditioned. For instance, Peters (2004) argued that English *-er* and *-eer* are allomorphs where the choice of *-eer* is semantically conditioned on the referent being from the semantic field of war. Baayen et al. (2013) discussed the Russian prefixes *pere-* and *pre-*, which are etymologically related but express subtly different semantics.

Endresen (2014) provides detailed discussion of the limitations of the classical definition of allomorphy. She points out that there are counterexamples where other parameters should be taken into account, such as subtle differences in meaning as exhibited by the Russian affix pairs *s-* vs. *so-* 'together', *o-* vs. *ob-* 'around', *pere-* vs. *pre-* 'across', *vz-* vs. *voz-* 'up', and *vy-* vs. *iz-* 'out of'. The two Indonesian prefixes that are the subject of this study

likewise raise the question of whether these prefixes are allomorphs, given their phonological similarity, or separate prefixes. As pointed out by Denistia (2018), Indonesian linguists mainly have described the two morphs as independent prefixes (Ramlan, 2009; Sneddon et al., 2010), but there are also studies that take them to be allomorphs (Dardjowidjojo, 1983; Kridalaksana, 2008). Since these two prefixes are similar in form, but not phonologically conditioned, and since they are similar, but not identical in meaning, the classical criteria for allomorphy are only approximately satisfied. Thus, the present study is a corpus based investigation into what Endresen (2014) refers to as non-standard allomorphy. Specifically, we examine in detail the differences in the semantics of *PE-* and *PEN-*, the differences in their productivity, and the differences in the extent to which derived words with *PE-* and *PEN-* are input to further inflection. In our analyses, the paradigmatic relations between base words and derived words are especially informative.[2]

In what follows, we first introduce some basic aspects of Indonesian verb morphology and deverbal nominalization. In the next section, we introduce the databases that inform our analyses. We then present our analyses and conclude with a discussion of the results obtained.

3.2 Indonesian verb morphology and deverbal nominalization

The morphology of Indonesian is characterized by productive processes of affix substitution. In this study, we are interested in two prefixes that create nouns from verbs through affix substitution, and which express a range of semantic functions (e.g. agent, instrument, patient, Sneddon et al. (2010, pp.30-33)). One prefix, henceforth *PEN-*, forms nouns from verbs with the prefix *MEN-* (e.g. *penari* 'dancer' – *menari* 'to dance'). In what follows, for notational clarity, we write prefixes in upper case and their allomorphs as subscripts. *PEN-* and *MEN-* have six allomorphs: PEN_{peng}-, PEN_{pen}-, PEN_{pem}-, PEN_{pe}-, PEN_{peny}-, PEN_{penge}-, and MEN_{meng}-, MEN_{men}-, MEN_{mem}-, MEN_{me}-, MEN_{meny}- and MEN_{menge}-. Sukarno (2017); Ramlan (2009) and

[2]For discussion of paradigmatic relations in derivational morphology, see (Marle, 1986; Stekauer, 2014).

Sugerman (2016) summarized the phonological conditioning of these allomorphs as follows:

- *PEN_{peng}-*, *MEN_{meng}-* occurs with base words beginning with a vowel or a velar obstruent /g/, /k/, /h/, or /kh/,
- *PEN_{pen}-*, *MEN_{men}-* occurs with base words beginning with a alveolar or palatal obstruent /d/, /t/, /c/, /j/, /sy/, or /z/,
- *PEN_{pem}-*, *MEN_{mem}-* occurs with base words beginning with a labial consonant /b/, /p/, or /f/,
- *PEN_{pe}-*, *MEN_{me}-* occurs with base words beginning with a nasal, a semivowel, or a liquid /m/, /n/, /ng/, /ny/, /w/, /j/, /r/, or /l/,
- *PEN_{peny}-*, *MEN_{meny}-* occurs with base words beginning with /s/, and
- *PEN_{penge}-*, *MEN_{menge}-* occurs with monosyllabic base words.

The nasal allomorphy of Indonesian *MEN-* and *PEN-* is an example of classical phonologically conditioned allomorphy.

A second prefix, henceforth *PE-*, forms nouns from verbs with the prefix *BER-*, again through affix substitution (e.g. *petani* 'farmer' – *bertani* 'to farm'), see Ramlan (2009); Ermanto (2016); Sneddon et al. (2010); Putrayasa (2008); Dardjowidjojo (1983); Benjamin (2009). *BER-* has *BER_{be}-* and *BER_{bel}-* as infrequent allomorphs. *BER-* primarily creates verbs expressing reciprocity, reflexivity, or stativity (see Kridalaksana (2007); Ramlan (2009); Putrayasa (2008); Chaer (2008); Sneddon et al. (2010) for other meanings). *BER_{be}-* occurs with stems beginning with /r/ or with stems the first syllable of which ends with /r/, as in *risiko* 'risk', *berisiko* 'to run the risk' and *kerja* 'work', *bekerja* 'to work'. *BER_{bel}-* only occurs with the base word *ajar* 'to teach', *belajar* 'to study' (Sugerman, 2016). If *PE-* is regarded as an allomorph of *PEN-*, its conditioning is not phonological, as for the allomorphs of *PEN-*, but morphologi-

cal: *PEN-* is paradigmatically related to verbs with *MEN-* and *PE-* is paradigmatically related to verbs with *BER-*.[3]

The base words for the verbs and their nominalizations can be verbs, nouns, and adjectives. There is no consistent difference in lexical meaning between simple base verbs and derived verbs (e.g. *buru* 'to hunt' – *berburu* 'to hunt'), although the derived forms may show different syntactic and aspectual behaviour (e.g. *buru* 'to hunt' – *memburu* 'to hunt continuously') (Nuriah, 2004). The simple verb is typically used in imperatives.

Verbs with *MEN-* can be extended with the suffixes *-i* and *-kan*. *MEN-* typically renders a verb explicitly transitive. The suffixes *-i* and *-kan* add a further argument, either a beneficiary or a causer, while often at the same time expressing intensification or iteration (Arka et al., 2009; Sutanto, 2002; Tomasowa, 2007; Kroeger, 2007; Sneddon et al., 2010).

1. transitives and ditransitives

    a) *tulis* 'to write', *menulis* 'to write something', *menulisi* 'to write something on something'

    b) *tulis* 'to write', *menulis* 'to write something', *menuliskan* 'to write something on behalf of someone'

2. causatives

    a) *panas* 'hot', *memanas* 'to become hot', *memanasi* 'to heat up something'

    b) *panas* 'hot', *memanas* 'to become hot', *memanaskan* 'to apply heat to something'

[3]We use the term paradigmatically related to denote systematic relationships between elements in absentia. Although in derivation, paradigmatic relations are less tightly knit compared to typical inflectional paradigms such as Latin or Estonian (Dressler, 1989), derivation also can show paradigmatic organisation (Stekauer, 2014). For the importance of paradigmatic organisation for linking elements in compounds, as well as for stress assigment in compounds, see Krott et al. (2009) and Plag (2006) respectively.

3. transitives and beneficiaries

    a) *ajar* ‘to teach’, *mengajar* ‘to teach something’, *mengajari* ‘to teach someone something’

    b) *ajar* ‘to teach’, *mengajar* ‘to teach something’, *mengajarkan* ‘to teach something to someone’

    c) *kirim* ‘to send’, *mengirim* ‘to send something’, *mengirimi* ‘to send something to someone’

4. iteration and intensification

    a) *lempar* ‘to throw’, *melempar* ‘to throw something’, *melempari* ‘to throw something repeatedly at something’

    b) *pukul* ‘to hit’, *memukul* ‘to hit something’, *memukuli* ‘to hit something hard over and over again’

Verbs with *BER-* do not combine with the *-i* suffix, but are found with *-kan* or *-an* to express possession (5, 6) and reciprocity (7, 8):

5. *dasar* ‘base’, *berdasarkan* ‘be grounded in’

6. *alamat* ‘address’, *beralamatkan* ‘to have an address’

7. *gandeng* ‘to hold hands’, *bergandengan* ‘to hold hands with each other’

8. *cium* ‘to kiss’, *berciuman* ‘to kiss each other’

Derived nouns with *PEN-* do not carry the *-i* or *-kan* suffixes, even though they may correspond to verbs with these suffixes. For instance, *penerbang*, ‘pilot’, is paradigmatically related to *menerbangkan* ‘to fly an aircraft’ rather than to the verb *menerbangi*, ‘to fly in something’, with the suffix *-i* marking location. Importantly, the verb *menerbang* does not exist but only the verbs *terbang*, ‘fly’, *menerbangkan* and *menerbangi*.

Occasionally, one finds both *PEN-* and *PE-*. There are 5 cases in which the form with *PE-* semantically refers to a profession and the form with *PEN-* does not, as listed in (9). There are also some cases in which the form with *PEN-* expresses agent, causer, or instrument and the form with *PE-* expresses patient or agent. In this case, 7 instances are attested in our database, as listed in (10).

9. *PEN-* and *PE-* formations that both express agents

    a) *tembak* 'to shoot', *penembak* 'someone who shoots' and *petembak* 'shooter' (athlete)

    b) *tinju* 'to punch', *peninju* 'someone who punches' and *petinju* 'boxer' (athlete)

    c) *terjun* 'to sky dive', *peterjun* 'sky diver' (athlete) and *penerjun* 'someone who sky dives'

    d) *selam* 'to dive', *penyelam* 'someone who dives' and *peselam* 'diver' (athlete)

    e) *dayung* 'to paddle', *pendayung* 'someone who paddles' and *pedayung* 'paddler' (athlete)

10. *PEN-* and *PE-* formations expressing different semantic roles

    a) *ajar* 'to teach', *pengajar* 'teacher' (agent) and *pelajar* 'student' (patient)

    b) *kasih* 'to love', *pengasih* 'lover' (agent) and *pekasih* 'love poison' (instrument)

    c) *sakit* 'to be sick', *penyakit* 'disease' (causer) and *pesakit* 'a person with disease' (patient)

    d) *sapa* 'to greet', *penyapa* 'a person who greets' (agent) and *pesapa* 'a person who is greeted' (patient)

    e) *siar* 'to announce/to sail', *penyiar* 'radio announcer' (agent) and *pesiar* 'a cruise ship' (instrument)

f) *tanda* ‘sign’, *penanda* ‘a sign’ (agent) and *petanda* ‘a hint’ (patient)

g) *tempur* ‘to combat’, *penempur* ‘armament’ (agent) and *petempur* ‘combatant’ (instrument)

We compiled a database containing 3090 words with *PE-* and *PEN-*. Since *PEN-* and *PE-* share the form *pe-*, the question arises of how to assign occurences of the form *pe-* to either *PEN-* or *PE-*. In 235 out of 240 potentially ambiguous forms, inspection of the paradigmatic relation with the corresponding base verb, either a verb with *MEN-* or a verb with *BER-*, the noun can be unambiguously assigned to be *PEN-* or *PE-*. Five words remain ambiguous: *pewushu*, ‘wushu athlete’, *perindang*, ‘provider of shadow’, *pemagang*, ‘probabitioner’, *pemuda*, ‘young male’, and *pemudik*, ‘homecomer’. The semantics of *pewushu* clarify that it belongs to *PE-*, the prefix that is used to denote professional athletes. The remaining 4 words are truly ambiguous, but are most likely, given their semantics, belong to the class of *PEN-* formation. For instance, *perindang* realises a causative reading, which, as we shall see below, is predominantly expressed by *MEN-*.

The goal of this paper is to clarify the morphological status of *PE-* and *PEN-*, allomorphs or separate prefixes, through a quantitative survey of their productivity, their paradigmatic relations with their base verbs, and the extent to which these derived nouns are input for further inflection. Indonesian inflection comprises several bound morphs: *-ku*, *-mu*, and *-nya* for first, second, and third person singular possessives or objects, *ku-* and *kau-* for first and second person subjects (Sneddon et al., 2010). In the Indonesian literature, these bound morphs are referred to as clitics, as they are phonologically reduced forms of free pronouns (Kridalaksana, 2008). There are also two suffixes that attach to verbs or nouns to express emphasis (*-lah* and *-pun*) or questioning (*-kah*). In what follows, we will refer to these morphs as inflectional, as they do not give rise to new onomasiological units but rather modify existing words much in the same way as adverbs modify verbs in English. Indonesian also has reduplication, which is used to express the plural for nouns and realizes various semantics function on verbs and adjectives, including intensification and iteration (Sugerman, 2016; Chaer, 2008; Rafferty, 2002; Dalrymple and Mofu, 2012). Following Booij (1996), we distinguish between

inherent and contextual inflection. Agreement marking on verbs (e.g, *ku-* and *kau-*) exemplifies contextual inflection, which is syntactically governed. Inherent inflection is more similar to word formation and hence in some languages can feed derivation and compounding. For instance, in Dutch, plural nouns can appear as left constituents in compounds (Schreuder et al., 1998). Reduplication in Indonesian is inherent inflection: it is not governed by syntactic context (*marah-marah* 'very angry', *anak-anak* 'children', *berhenti-berhenti* 'to stop repeatedly'), and can feed further inflection, as in *memukul-mukuli*, 'to hit intensively over and over again', which has as parse [[[meN + [*pukul*]$_{\mathrm{N}}$]$_{\mathrm{V}}$ + *pukul*]$_{\mathrm{V}}$ + i]$_{\mathrm{V}}$. We shall see below that derived words with *PE-* are more often input to these inflectional processes than derived words with *PEN-*. We will argue that the joint quantitative evidence justifies to analyse *PE-* and *PEN-* as two distinct prefixes rather than as allomorphs. In the next section, we introduce the databases that we derived from a 36 million token corpus of written Indonesian (Goldhahn et al., 2012).

3.3 Materials

We created a database from the Indonesian corpus that is part of the Leipzig Corpora Collection at `http://corpora2.informatik.uni-leipzig.de/download.html`, accessed in April 2016. This corpus comprises a variety of written registers (the web, newspapers, Wikipedia) dating from the years 2008 - 2012 (Goldhahn et al., 2012). There are 112.025 different word types in this corpus, that occur in 2.759.800 sentences, to a total of 36.608.669 word tokens.

The words in the corpus were morphologically analyzed using the MorphInd parser, which has an overall accuracy of 84.6% (Larasati et al., 2011) and it was run in single word mode, i.e, compounds were not parsed. Prior to running the parser, the 200 words with *PE-* or *PEN-* that contained a typo were corrected manually. The MorphInd parser's results for *PE-* and *PEN-* were checked and corrected manually against the online version of *Kamus Besar Bahasa Indonesia*(hereafter called the dictionary), a comprehensive dictionary of Indonesian (`http://kbbi.kemdikbud.go.id`; accessed on June 2016), to verify the morphological status and semantics of the *PE-* and *PEN-* words. We made use of the fourth edition, published in

2012, which has more than 90,000 lemmas (Alwi, 2012). The language it records is formal; it omits words that are considered slang or foreign. Where the dictionary and the MorphInd are in conflict, we followed the dictionary. Where the dictionary does not provide information on the word category of the base, we followed the MorphInd parser. The precision of the parser for these words was 0.98 and its recall was 0.82, using the dictionary as the gold standard complemented with manual verification for out-of-vocabulary words.

Word	Translation	Allomorph	Base	Parser
pemerintah	government	pem	perintah	peN+perintah<n>_NSD
pemain	player	pe	main	peN+main<v>_NSD
petugas	officer		tugas	petugas<n>_NSD
pekebun	gardener		kebun	pekebun<x>_X–
pengusut	investigator	peng	usut	peN+kusut<a>_NSD
pengelas	welder	penge	las	peN+kelas<n>_NSD

Table 3.1: Examples of the output of the MorphInd parser

Sample output of the parser is shown in Table 3.1: a morphological segmentation is provided where available, as well as a word category label. Table 3.1 shows that MorphInd identifies *pemerintah* and *pemain* correctly. However, it is not able to identify *PE-* in *petugas* and *pekebun*. In some cases, the base identified by the parser is incorrect. For instance, *pengusut* is formed from *usut* (to investigate) [PEN_{peng}- + *usut*], but MorphInd identifies its base as *kusut* (tangled) [PEN_{peng}- + *kusut*]. MorphInd also is not always able to accurately identify single syllable base words. In the above examples, this is illustrated by *pengelas* (welder) which derives from *las* (weld), [PEN_{penge}- + *las*], and not *kelas* (classroom), [PEN_{peng}- + *kelas*]. Therefore, the output of the parser was manually checked and corrected when necessary.

We processed the data using the R (version 3.3.2) programming language (R Team, 2017) in R Studio (R Team, 2015). The databases and the R scripts used to construct these databases are available online at `http://bit.ly/PePeNProductivity`. In what follows, we first present the database with Indonesian verbs, and then proceed to the database with derived nouns with *PE-* and *PEN-*.

3.3.1 The database of Indonesian verbs

Indonesian has deverbal morphology for active, passive, causative, and transitive semantics among others, see Table 3.2 for examples. From the corpus, we retrieved all verbs recognized by the MorphInd parser and brought these together in a database. The total number of types in the database is 26996. Table 3.3 illustrates that for each verb, we provide information on the derived word's frequency in the corpus, the parse provided by MorphInd, the base word, the word category of the base, and the affix or affixes in the verb. When particles (e.g. *-lah*, *-kah*, *-pun*) or affixes (e.g. *ku-*, *-ku*, *kau-*, *-mu*, *-nya*) are found attached to a verb (Sneddon et al., 2010; Sugerman, 2016), this form is listed with its own entry. [4]

The database comprises 2489 simple verbs and 24507 affixed verbs (3665 verbs with suffixes, 11562 verbs with prefixes, and 9280 verbs with both prefix and suffix). We observed 27 verb constructions of which 13 are reported in the literature (Hidajat, 2014; Fortin, 2006; Sneddon et al., 2010; Benjamin, 2009; Arka et al., 2009; Sudaryanto, 1993; Kridalaksana, 2007). In our corpus, there are 2 attested verb constructions (e.g. *terke-/-an* and *terper-/-an*) that are not productive (1 token and 8 tokens respectively). Table 3.2 lists the 25 productive constructions.

As our specific interest is in nouns with *PE-* and *PEN-*, we extracted from this database all the verbs that correspond to these nominalizations and that carry the prefix *BER-* or *MEN-*. To this new database, henceforth the `MeBer Database`, we added information on the frequency of the base words of these complex verbs, whether the verbal prefix is *MEN-* or *BER-* and also the allomorph of *MEN-*. Whereas all nominalizations with *PEN-* have a corresponding verb with *MEN-*, there is one simple verb, *sohor* 'to be famous', that has a corresponding nominalization with *PE-*, *pesohor* 'a famous person', without having a corresponding verb with *BER-*. This verb-noun pair is not in the `MeBer Database`, but in a separate database (`SimpleWords`) which also specifies the frequency of the base verb and the frequency of the derived noun (see Sneddon et al. (2010) for discussion of such exceptional pairs).

[4] The column `MorphologicalVariation` specifies the related particles or affixes. English translations in the tables of this paper are provided for convenience but are not part of the databases.

Base Word	Base Translation	Derived Verb	Translation	Affix Semantics
dasar	base	berdasarkan	to have the base	having X
temu	to meet	bertemu	to meet each other	reciprocative
lanjut	to continue	berkelanjutan	to countinuously continue	continous - intransitive
anggap	to assume	beranggapan	to have assumption	having X
alat	tools	berperalatan	having many kinds of tools	having many kinds of X
bantu	to help	membantu	to help	active-transitive
timbang	to weight	mempertimbangkan	to consider	active-transitive-causative
ajar	to teach	mengajarkan	to teach someone	active-transitive-beneficiary
ingat	to remember	memperingati	to commemorate	active-transitive-causative
mudah	easy	mempermudah	to make something easier	active-transitive-causative
henti	stop	memberhentikan	to make someone stop	active-transitive-causative
pukul	to hit	memukuli	to hit repeteadly	active-transitive-iterative
temu	to find	ditemukan	to be found	passive
bawa	to carry	dibawa	to be carried	passive
setuju	to agree	disetujui	to be agreed	passive
luas	wide	diperluas	to be made wider	passive-causative
juang	to fight	diperjuangkan	to be fought for	passive
senjata	weapon	dipersenjatai	to be equipped with weapon	having many kinds of X
tidur	to sleep	tertidur	to sleep unintentionally	accidental
pisah	to be separated	terpisahkan	can be separated	intransitive
tanding	to beat	tertandingi	can be defeated	intransitive
tawa	laughter	tertawaan	to laugh intensively at something	intransitive
lihat	to look	kelihatan	to make something visible	adversative
kata	to say	katakan	to make someone say something	causative
temu	to meet	temui	to make someone meet someone else	causative

Table 3.2: Examples of simple and complex verbs in Indonesian, and affix combinations in complex verb as attested in the corpus

Word	Translation	Frequency	MorphInd	Base Word	Base Word Class	Morphological Variation	Affix
menjadi	become	154758	meN+jadi<a>_VSA	jadi	a		meN-
bersama	be together	32206	ber+sama<a>_VSA	sama	a		ber-
terkait	be connected to	25561	ter+kait<v>_VSP	kait	v		ter-
ujarnya	his/her saying	22957	ujar<v>_VSA+dia<p>_PS3	ujar	v	3rdSingPronoun	simple
turun	go down	18970	turun<v>_VSA	turun	v		simple
diduga	be suspected	11240	di+duga<v>_VSP	duga	v		di-

Table 3.3: Examples of entries in the verb database

All the data in `MeBer Database` were compiled computationally from the output of the MorphInd and subsequently checked manually using the dictionary. In total, there are 8484 words with the *MEN-* prefix and 3582 words with the *BER-* prefix. These counts include forms with the suffixes *-i, -kan* or *-an*. To this database, we added some words such as *beserta* 'to be together with', *belajar* 'to study', *beternak* 'to farm', *bekerja* 'to work', and *beterbangan* 'to fly randomly' and their inflectional variants, forms which MorphInd did not recognize but that we happened to identify in the course of this study. The MorphInd parser also does not recognize verbs with the allomorph *menge-*. For the 18 nominalizations with PEN_{penge}-, we manually searched for the occurrences of the corresponding verbs and added these together

with their frequency counts to the MeBer database. Finally, a total of 297 verbs with *MEN-* and 14 verbs with *BER-* were not recognized by the parser, and were corrected manually on the basis of the dictionary.[5]

Word	Translation	Frequency	MeN	MeN Allomorph	Base Word	Base Frequency	Base Word Class	MorphInd	Morphological Variation
mengatakan	to say	119115	TRUE	meng	kata	151552	n	meN+kata<n>+kan_VSA	
melakukan	to do	76116	TRUE	me	laku	1689	adj	meN+laku<a>+kan_VSA	
memiliki	to have	62317	TRUE	me	milik	12427	v	meN+milik<v>+i_VSA	
membuat	to make	46242	TRUE	mem	buat	8431	v	meN+buat<v>_VSA	
memberikan	to give	43803	TRUE	mem	beri	1295	v	meN+beri<v>+kan_VSA	
berada	to be	36567	FALSE		ada	196988	adj	ber+ada<a>_VSA	
bersama	to be together	32206	FALSE		sama	49719	adj	ber+sama<a>_VSA	
berdasarkan	to be the basis	19248	FALSE		dasar	13180	n	ber+dasar<n>+kan_VSA	
berbeda	to be different	17895	FALSE		beda	1861	adj	ber+beda<a>_VSA	
berlangsung	to be on going	17558	FALSE		langsung	37225	adj	ber+langsung<a>_VSA	
melakukannya	to do it	2800	TRUE	me	laku	1689	adj	meN+laku<a>+kan_VSA+dia<p>_PS3	3rdSingPronoun
membuatnya	to make it	2775	TRUE	mem	buat	8431	v	meN+buat<v>_VSA+dia<p>_PS3	3rdSingPronoun
meningkatnya	the increase	2311	TRUE	men	tingkat	22098	n	meN+tingkat<n>_VSA+dia<p>_PS3	3rdSingPronoun
berkurangnya	the decrease	638	FALSE		kurang	18761	adj	ber+kurang<a>_VSA+dia<p>_PS3	3rdSingPronoun
bertambahnya	the addition	593	FALSE		tambah	5457	v	ber+tambah<v>_VSA+dia<p>_PS3	3rdSingPronoun

Table 3.4: Examples of entries in the MeBer database

3.3.2 The PePeN Database

We brought together the *PE-* and *PEN-* words in a lexical database, henceforth the PePeN database. This database also includes the noun with *PE-* that have a simple verb as the base. In this way, we obtained a total of 3090 words, 267 with *PE-*, 2818 with *PEN-*, and 4 words with the unproductive variant *PER-* (Benjamin, 2009).[6] There are 34 words that the MorphInd parser did not analyze.

All derived words were annotated manually for semantic role (agent, instrument, causer, patient, and location), and checked (for at least one token) against both the dictionary and usage in the corpus. As in English, where *-er* nominalizations may express multiple semantic roles (Booij, 2010; Booij and Lieber, 2004) (e.g. *printer*, which has both an agent and instrument reading), Indonesian *PE-* and *PEN-* formations can have multiple interpretations (see Table 3.5). In this study, we did not distinguish between impersonal agent[7] and instrument. Although it is well known that *PEN-* create agents, patients, and instruments (Sneddon et al., 2010), we observed a few cases of causer (e.g. *penyakit* 'disease') and location (e.g.

[5]We suspect that the base words of the *MEN-* and *BER-* verbs were not in MorphInd's dictionary.

[6]There are four such forms in our database, *pertapa* and *petapa*, and their reduplicated variants *petapa-petapa*, *pertapa-pertapa*.

[7]Booij (1986) uses the term impersonal agent for the meaning 'radio station' of the Dutch word *zender* which also has an agentive reading, 'one who sends', and an instrumental reading, 'transmitter'.

penghujung 'the end') in our database. It is possible, even likely, that semantic roles are in use in the corpus without being registered in the database, as manual verification of all 579564 tokens with *PE-* or *PEN-* in the corpus was infeasible. In the database, words with more than one semantic role have multiple entries in the database, with one row for each role (cf. Table 3.5). The frequencies listed in rows of Table 3.5 are those of the overall frequency of the word and are not broken down by semantics.

Word	Translation	Semantic Role
pembanding	who compares	agent
pembanding	tool to compare	instrument
pembanding	who is compared	patient
penggerak	tool to move	instrument
penggerak	who moves	agent
pemicu	a trigger	agent
pemicu	who triggers	instrument
pewaris	heir	agent
pewaris	who gives inheritance	patient

Table 3.5: Examples of semantic role

The PePeN Database thus provides the following information:

1. Word frequency: the token frequency of the derived word in the corpus,
2. Allomorph: the form of the *PEN-* prefix; where the allomorph does not follow the rules as given in Chaer (2008); Sneddon et al. (2010), e.g. *penglihat* 'seer' is expected to be *pelihat*, this is marked in the 'notes' column of the database as `AllomorphDeviation`,
3. Base word,
4. Word category of the base word,
5. Base word frequency: the token frequency of the base word in the corpus,
6. MorphInd output as illustrated in Table 3.1,
7. Semantic role of the derived noun with respect to the base word (agent, instrument, patient, ...),

8. Morphological variation: reduplications, particles (e.g.*-lah, -pun, per-*) or affixes (e.g.*-ku, -mu, -nya*), if present,

9. Typo: whether the form in the corpus had a spelling error (corrected in the database, frequency counts include the frequency of the corrected typos); when several spelling alternants are in use, this is indicated in the `FreeVariance` column of the database as illustrated in Table 3.7

Entries of this database are listed in Table 3.6.

Word	Translation	Frequency	Allomorph	Base Word	Base Word Class	Base Frequency	Semantic Role	Morphological Variation
pemerintah	government	78047	pem	perintah	n	4315	agent	
pelaku	doer	17776	pe	laku	n	1689	agent	
penyakit	disease	12042	peny	sakit	adj	20454	causer	
pengusaha	enterpreneur	8053	peng	usaha	n	18041	agent	
pendukung	supporter	5960	pen	dukung	v	710	agent	
pelajar	student	3421		ajar	n	729	patient	
penasihat	advisor	2050	pe	nasihat	n	386	agent	
penyebabnya	his/her causer	1106	peny	sebab	n	18271	agent	Poss3rdSing

Table 3.6: Example entries in the PePeN database

Word	Translation	Frequency	TypoRevision	FreqOfTypo	FreeVariance
pebowling	bowling player	23	peboling	23	TRUE
pengonsumsi	consumer	12	pengkonsumsi	12	TRUE
pemain	player	34704	pemian,pemaen, pemasin,pemein, pemaik,pemailn, pemainn,pemiain, pemjain	20,4, 2,2, 1,1, 1,1, 1	FALSE

Table 3.7: Example entries in the PePeN database illustrating spelling variants and typos (*pemain* is the second most frequent *PEN-* nominalizations in the database)

3.4 Analysis

3.4.1 Productivity of *PE-* and *PEN-* derived nouns

Prefix	Tokens	Types	Hapaxes
PEN	498484	2221	588
PE	81083	184	45

Prefix	Tokens	Types	Hapaxes
penge-	535	18	6
peny-	38533	244	75
peng-	83515	628	173
pen-	91985	546	142
pe-	138165	417	103
pem-	145696	364	89

Table 3.8: Counts of tokens, types, and hapaxes for *PE-* and *PEN-* (upper table) for the six allomorphs of *PEN-* (lower table)

The *PE-* and *PEN-* prefixes differ in their productivity. As shown in the upper panel of Table 3.8, *PEN-* occurs with more tokens, more types, and more hapax legomena compared to *PE-*. Further detail is provided by the lower panel of Table 3.8, which shows the numbers of tokens, types, and hapaxes for *PEN-* allomorphs and *PE-*.

Figure 3.1 presents rank-frequency plots for *PE-* and *PEN-* (left panel), and for *PE-* and all allomorphs of *PEN-* (right panel), using logarithmic scales (Zipf, 1935, 1949). The left panel clarifies that the highest ranked words with *PEN-* also exceed in frequency the highest ranked words with *PE-*. Nevertheless, the productivity index V1/N (Baayen, 2009) remains greater for *PEN-* (0.00118) than for *PE-* (0.00055). The second panel of Figure 3.1 shows that four of the six allomorphs of *PEN-* have rank-frequency curves that lie above the rank-frequency curve of *PE-*. The curve for PEN_{peny}-, crosses the curve for *PE-* around rank 50, but still shows many more low-frequency formations. The only allomorph that is less productive than *PE-* is PEN_{penge}-, an allomorph that attaches to monosyllabic words and which appears in the corpus with only 18 types.

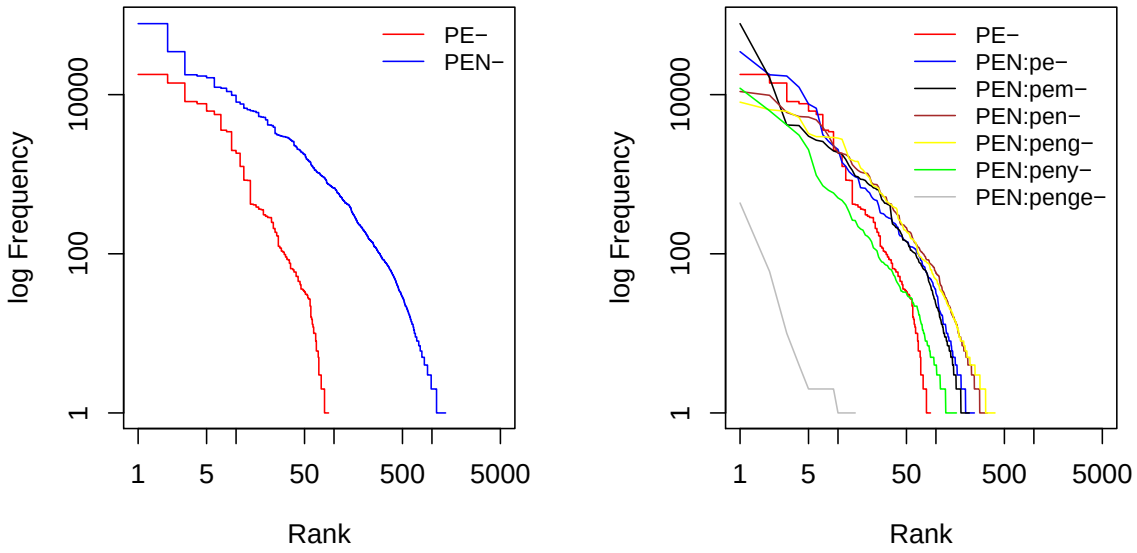


Figure 3.1: Rank-frequency curves for *PE-* and *PEN-* (left panel), and for *PE-* and sum of the allomorphs of *PEN-*'s frequency (right panel). *PE-* is less productive than *PEN-*, and it is also less productive than the allomorphs of *PEN-*, with the exception of PEN_{penge}-, which is attested with only 18 types

Given the similarity of *PE-* and *PEN-* form, the question arises of whether it makes sense to consider *PE-* as a low productivity allomorph of *PEN-*. To address this question, we examined the counts of types and hapax legomena for *PE-* and the allomorphs of *PEN-* as a function of the number of base verbs with *BER-* and base verbs with allomorphs of *MEN-*. The panel of Figure 3.2 shows that the rate at which base verbs give rise to derived nouns is the same (according to a regression model) for all allomorphs of *MEN-* and that *PE-* patterns as an outlier, both with respect to type counts and with respect to hapax legomena. It is remarkable that the rate at which hapaxes and types appear is so constant across the allomorphs of *PEN-* and *MEN-*. From this, we draw the conclusion that the outlier *PE-* is best understood as a formative in its own right. We note here that Indonesian *PEN-* and *MEN-* offer a remarkable window on the relation between base productivity and derived productivity.

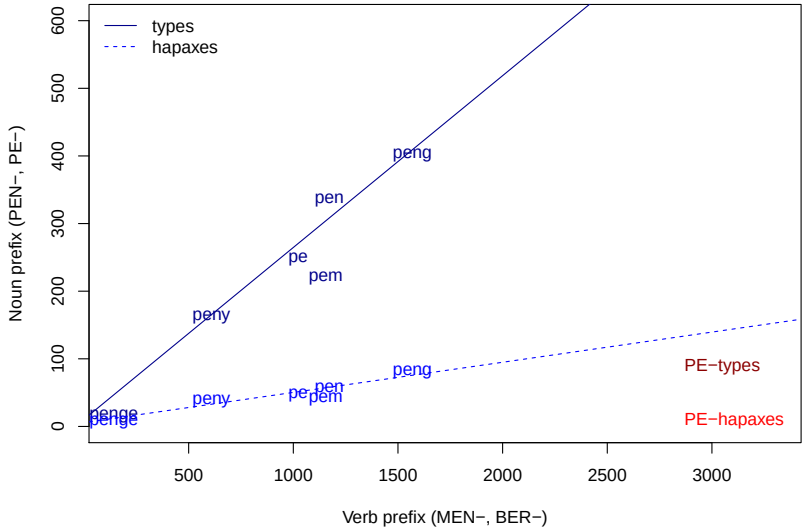


Figure 3.2: Counts of types for base verbs (horizontal axis) and counts of types and hapaxes for *PE-* and *PEN-* (vertical axis); solid and dashed lines represent regression lines to the *PEN-* allomorphs for counts of types and counts of hapax legomena respectively

	agent	causer	instrument	location	patient
PE-	170	0	3	0	15
pe-	316	0	94	1	6
pem-	271	0	91	0	2
pen-	412	0	130	3	1
peng-	474	0	149	4	1
penge-	13	0	5	0	0
peny-	185	6	53	0	0

	agent	causer	instrument	location	patient
PE-	39	0	0	0	6
pe-	75	0	25	1	2
pem-	62	0	27	0	0
pen-	108	0	33	0	1
peng-	136	0	37	0	0
penge-	3	0	3	0	0
peny-	59	1	15	0	0

Table 3.9: Cross-tabulation of *PE-* and the allomorphs of *PEN-* by semantic role. Upper table: counts of types; lower table: counts of hapax legomena

Further evidence that *PE-* is not an allomorph of *PEN-* emerges when we take the semantic roles of the derived noun into account. Table 3.9 cross-tabulates *PE-* and the allomorphs of *PEN-* by the semantic roles of these nouns in our database; Figure 3.3 provides the corresponding visualisation for the three roles that are most frequent: agent, patient, and instrument. Both *PE-* and *PEN-* create agent nouns. *PE-* shows some productivity for patient nouns, of which there are proportionally very few among the nouns with *PEN-*. (The numbers are small, but this asymmetry is significant according to a chi-squared test, $\chi^2_{(1)} = 81.32, p < 0.0001$; interestingly, the few patient nouns with *PEN-* are realised with the allomorph *pe-*, however, the proportion of patient hapaxes is much lower (0.02 for *PEN-* and 0.13 for *PE-*, $p < 0.015$, proportion test). Conversely, *PEN-* is productive for instruments, which are virtually absent for *PE-*. This may be one of the reasons that *PEN-* is more productive than *PE-*. For *PEN-*, a chi-squared test indicates that the ratios of agents to instruments are proportional across all allomorphs ($\chi^2_{(5)} = 1.01, p > 0.1$ and $\chi^2_{(5)} = 5.48, p > 0.1$ for both types and hapax legomena). The uniformity of semantic functions accross the allomorphs of *PEN-* is perfectly in line with the fact that these allomorphs are phonologically conditioned. Conversely, the lack of productivity for instruments that characterizes *PE-*, and its (limited) productivity for patient nouns that is strongly attenuated for *PEN-* is a further indication that *PE-* is unlikely to be an allomorph of *PEN-*. Thus, Indonesian *PEN-* and *PE-* show the kind of semantic specialisation that led Baayen et al. (2013) to conclude that Russian *pere-* and *pre-* are not allomorphs but independent prefixes.

The counts underlying Table 3.9 and Figure 3.3 are based on a type definition that distinguishes between forms of the noun with different possessive suffixes or suffixes expressing emphasis, as well as noun plurals. When such variants are collapsed into a single type, the pattern of results on the ratios of agents to instruments across all allomorphs remains similar ($\chi^2_{(5)} = 0.75, p > 0.1$ and $\chi^2_{(5)} = 5.11, p > 0.1$ for both types and hapax legomena). However, the number of distinct types for patient nouns with *PE-* reduces to 5, each of which occurs more than once. Thus, *PE-* appears to be well-entrenched for a handful of patient nouns, but does not show real productivity here.

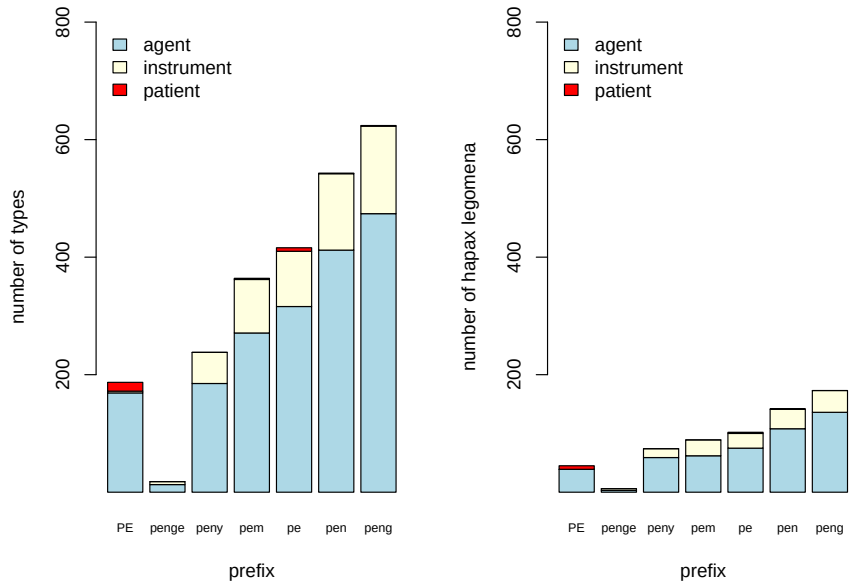


Figure 3.3: Counts of types (left panel) and hapax legomena (right panel) broken down by semantic role, for *PE-* and the allomorphs of *PEN-*. Both prefixes support agents, but *PE-* shows limited productivity for patient nouns, whereas *PEN-* shows additional productivity for instruments

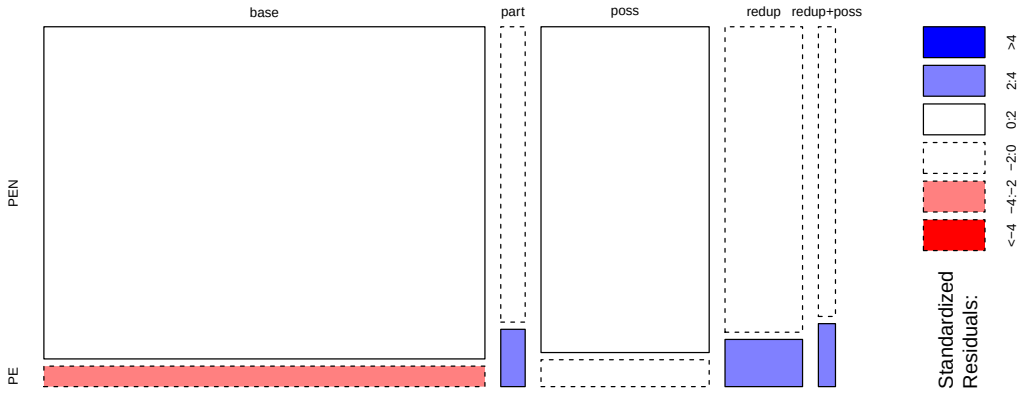


Figure 3.4: Mosaic plot for the cross-classification of *PE-* and *PEN-* by type of inflection. The colour coding represents the Pearson residuals, which clarify where the observed counts are greater (purple) or smaller (pink) than the expected values. A chi-squared test confirms that *PE-* and *PEN-* distribute differently over inflectional types ($\chi^2_{(4)} = 36.59, p < 0.0001$)

	PE	pe	pem	pen	peng	penge	peny
base	84	247	218	325	396	15	160
BoundMorpheme	3	0	0	1	1	0	0
Particle	13	20	14	16	15	0	2
Possession	42	93	86	133	140	2	55
Possession+Particle	1	1	1	1	7	0	1
Reduplication	34	43	36	61	56	1	23
Reduplication+Particle	0	0	0	0	1	0	0
Reduplication+Possession	10	13	9	9	12	0	3

Table 3.10: Counts of variants types for *PE-* and allomorphs of *PEN-*. The base represents the non-variant forms. Particles, possessive suffixes, and plural reduplications dominate the counts

Krott et al. (1999) reported the paradoxical finding that words with less productive affixes tend to be used more as base words for further word formation. A similar observation holds for *PE-* and *PEN-*, but now for inflection rather than word formation. Inflectional variation is well illustrated by the noun *pengikut* 'follower', which is attested in the corpus with 9 variants: *pengikutku* 'my follower', *pengikutmu* 'your follower', *pengikutnya* 'his/her follower'; reduplication as in *pengikut-pengikut* 'followers'; reduplication and affixes as in *pengikut-pengikutmu* 'your followers', *pengikut-pengikutnya* 'his/her followers'; affixes and particles as in *pengikutmupun* 'your follower' (contrastive your, i.e., your, not somebody else's follower), *pengikutnyapun* 'his/her follower' (contrastive), *pengikutnyalah* 'his/her follower' (contrastive in imperative mood). Table 3.10 shows the counts of the different kinds of inflections types for *PE-* and *PEN-*. In our corpus, particles (e.g. *-lah, -pun*), possessive suffixes (e.g. *-ku, -mu, -nya*), and plural reduplications are used most often. Figure 3.4 presents a mosaic plot for the cross-classification of *pe* and *PEN-* by type of inflection. The mosaic plot shows that inflected forms of *PE-* are overrepresented for particles, plurals, and combinations of plurals and possessives. In other words, the less productive prefix, *PE-*, is used more intensively as input for further inflection than is the case for *PEN-*. This is likely to be due to the greater entrenchment of words with *PE-* in the mental lexicon, which makes them more readily available for more further affixation. Thus, the same principles that Krott et al. (1999) reported for derivation in Germanic languages generalize to inflection in Indonesian.

3.4.2 The base verbs of *PEN-* and *PE-*: *MEN-* and *BER-*

Several studies call attention to the tight relation between *PE-* and *PEN-* and their verbal base words (Putrayasa, 2008; Chaer, 2008; Ramlan, 2009; Kridalaksana, 2007; Dardjowidjojo, 1983). We therefore inspected the productivity of verb formation, focusing on monomorphemic words as potential base words. In our database, a total of 5581 such monomorphemic words is attested, with 3617 simple nouns, 943 simple adjectives, and 1021 simple verbs. As shown in Table 3.2, a large number of affixes is available for creating verbs from nouns, adjectives, and verbs. For this study, the number of different complex verb forms will be referred to as a monomorphemic word's verb family size. The verb family size measure includes inflectional variants of the verbs in its counts. Plots of this verb family size against base frequency show that, as expected, a higher base frequency predicts a greater verb family size. Interestingly, the functional form of this relation is different for base words that give rise to nouns with *PEN-*, and those that do not. This is illustrated in Figure 3.5 (see also Table 3.11), which present the results of a GAM (Generalized Additive Model, `MGCV` package version 1.8-17, Wood (2006, 2011)) with a poisson link fitted to the verb family size with centered log base frequency as the predictor. The increase of verb family size with base frequency is greater when *PEN-* is present, as can be seen by comparing the right panel with the left. In the right panel, we see a linear increase, whereas in the left panel, there is no increase at all for the lowest frequency base words. For the larger part of the range of the base word frequencies, the verb family size is larger if the verb family has a noun with *PEN-*. We also considered the base words with *PE-* in the verb family, but as the resulting curve was not significantly different from that of base words with verb families that did not have either nominalization, the two sets were merged into one defined by the absence of *PEN-* in the verb family. Apparently, base productivity and derived productivity are interacting for *PEN-*, but independent for *PE-*.

Figure 3.6 [8] presents mosaic plots for the cross-classification of word category and the presence of *PE-* or *PEN-* in a monomorphemic base word's verb family. The mosaic plot in the left panel concerns base words that have at least one formation in their verb family

[8] This plot is created using `VCD` package version 1.4.4 (Zeileis et al., 2007)

A. parametric coefficients	Estimate	Std. Error	t-value	p-value
intercept	0.7790	0.0106	73.4636	< 0.0001
type = *PEN-*	0.9446	0.0158	59.6832	< 0.0001
B. smooth terms	edf	Ref.df	F-value	p-value
log base freq, type != *PEN-*	3.7165	3.9521	982.0015	< 0.0001
log base freq, type = *PEN-*	1.0003	1.0005	512.5895	< 0.0001

Table 3.11: GAM summary for partial effects for verb family size regressed on centered log base frequency, for morphological families including derived nouns with *PEN-* and without *PEN-* but possibly including nouns with *PE-*

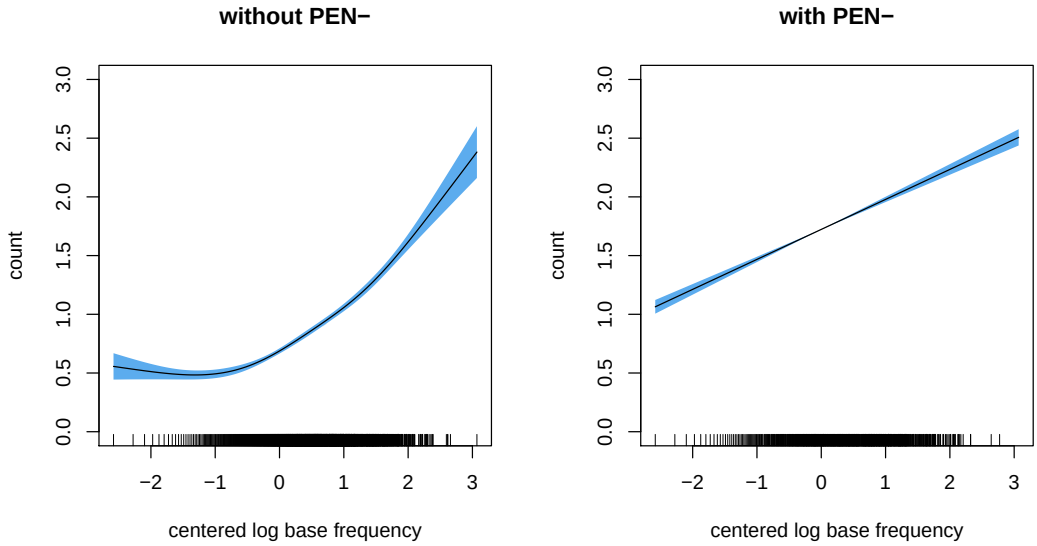


Figure 3.5: Partial effects for verb family size regressed on centered log base frequency, for morphological families without nouns with *PEN-* but possibly including nouns with *PE-* (left panel) and for morphological families including derived nouns with *PEN-* (right panel)

(i.e. with neither *PE-* and *PEN-*, with *PE-*, or with *PEN-*). The plot shows that simple words that give rise to affixed verbs but not to any formations with *PE-* or *PEN-* are overrepresented for nouns, and that base words that have *PEN-* in their verb family are overrepresented for verbs, unsurprisingly ($\chi^2_{(4)} = 839.97, p < 0.0001$). These overrepresentations are indicated by the residuals (Zeileis et al., 2007). The right panel concerns monomorphemic base words for which the verb family size is zero. Again, we see that base words that have *PEN-* in their verb family are overrepresented for verbs ($\chi^2_{(4)} = 288.58, p < 0.0001$). No such overrepresentation is visible for *PE-*. Whereas the literature on *PE-* and *PEN-* holds that *PEN-* is derived from verbs with *MEN-*, our corpus data indicate that *PEN-* actually can attach to simple words that

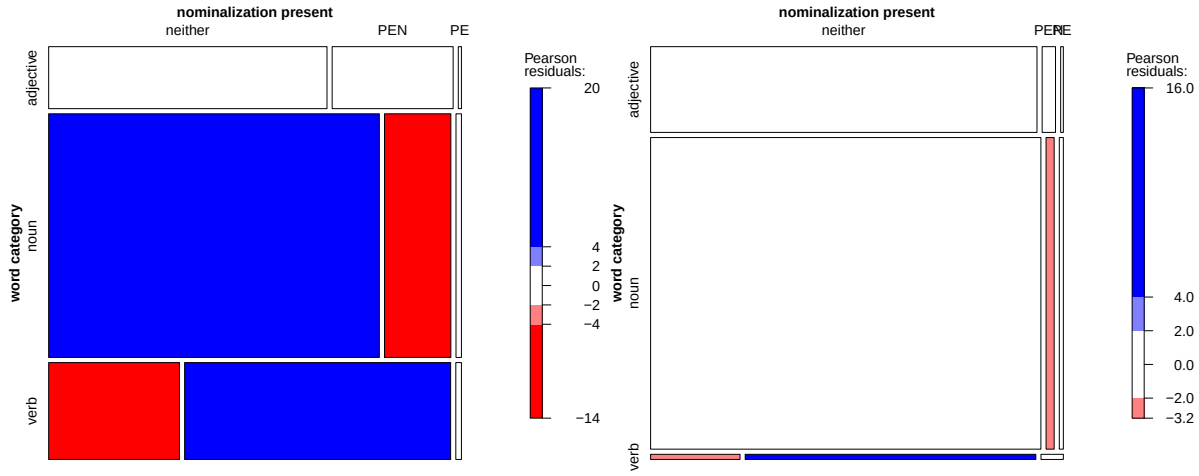


Figure 3.6: Left panel: mosaic plot for the type counts of verbs derived from monomorphemic words cross-classified by the word category of the monomorphemic word and the presence of *PE-* or *PEN-* in its verb family. Right panel: corresponding mosaic plot for the type counts of monomorphemic words that do not have any derived verbs attested in the corpus. The colour coding represents the Pearson residuals, which clarify where the observed counts are greater (blue) or smaller (red) than the expected values

do not have a corresponding verb with *MEN-*, even though the total number of instances is small (45). It is possible that the relevant *MEN-* verbs are in use in the language, but not attested in our corpus. Alternatively, it is conceivable that these *MEN-* verbs only have a virtual existence as possible words.

Prefix	Tokens	Types	Hapaxes
MEN-: menge-	1704	26	4
MEN-: meny-	187756	519	91
MEN-: me-	538078	1074	173
MEN-: mem-	558348	977	246
MEN-: men-	706181	1476	190
MEN-: meng-	722637	1102	292
BER-	801052	2869	760

Table 3.12: Counts of tokens, types, and hapaxes for six *MEN-* allomorphs (e.g. *menge- meny-, me-, mem, -men, meng-*) and *BER-*

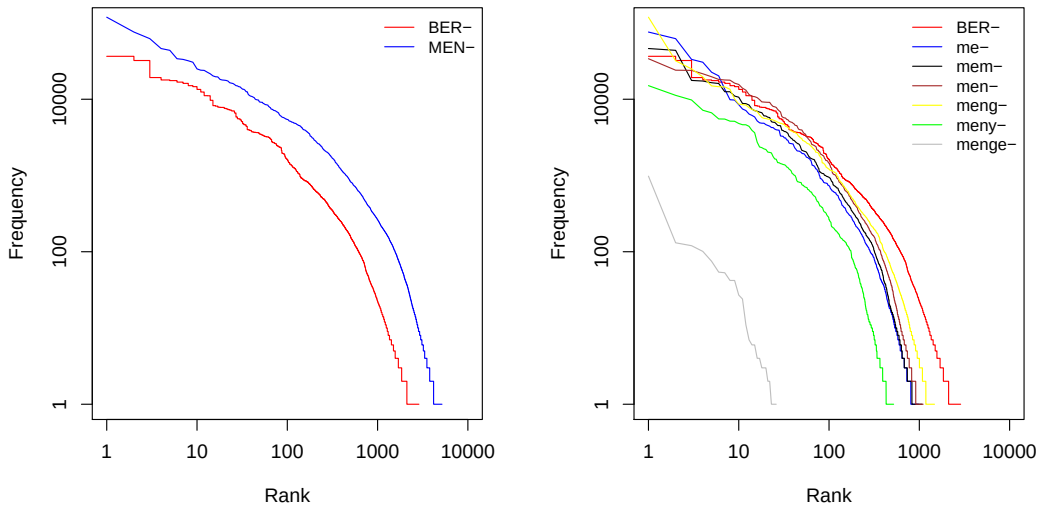


Figure 3.7: Rank-frequency plots for *MEN-* and *BER-* distributions. The x-axis represents rank and y-axis represents frequency of occurrence in the corpus. The lines in the left panel illustrate that *MEN-* is more productive than *BER-*. However, *BER-* becomes the most productive prefix when it is compared to the individual allomorphs of *MEN-* (right panel)

We have seen that *PEN-* is more productive than *PE-* and more tightly integrated into the verbal system. This raises the question of whether the reduced productivity of *PE-* might be due to reduced productivity of the verbal prefix *BER-*. Indeed, verbs with *MEN-* are more productive overall than verbs with *BER-* (2714704 tokens with *MEN-* vs. 801052 tokens with *BER-*, 5174 types with *MEN-* vs. 2869 types with *BER-*, and 996 hapax legomena with *MEN-* vs. 760 hapax legomena with *BER-*); see also Table 3.12 and the rank-frequency plot for *BER-* and *MEN-* in the left panel of Figure 3.7. However, when considering the allomorphs of *MEN-* separately, it turns out that *BER-* is more productive than any of these allomorphs, as shown in the right panel of Figure 3.7. Although *BER-* is more productive than any of the allomorphs of *MEN-*, it is not the case that *PE-* is proportionally more productive than any of the allomorphs of *PEN-*. It follows that the modest productivity of *PE-* is not a straightforward consequence of the lack of productivity of *BER-*. This conclusion receives further support from the presence of a significant correlation between the frequency of the *MEN-* base and the *PEN-* nominalization ($r_s = 0.4397, p < 0.0001$) and the absence of such a correlation for *BER-* and *PE-* ($r_s = 0.1908, p = 0.1711$).

3.5 General discussion

We have presented a quantitative investigation of the use of two nominalizing prefixes of Indonesian: *PE-* and *PEN-*. Although quite similar in form, nouns with *PE-* are described by literature as derived from verbs with the prefix *BER-*. Conversely, nouns with *PEN-* typically originate from verbs with the prefix *MEN-*, and show the same allomorphy in the same conditioning contexts as these prefixed verbs. In this paper, we addressed three questions. First, do *PE-* and *PEN-* differ with respect to their degree of productivity? Second, how does their productivity relate paradigmatically to the productivity of their base words? Third, given the similarity in form of *PE-* and *PEN-*, should they be taken to be allomorphs? To answer these questions, we examined the use of these nominalizations and their base words in a corpus of written Indonesian.

With regards to their productivity, *PEN-* is clearly more productive than *PE-* by any measure of productivity. In fact, *PE-* is less productive than any of the allomorphs of *PEN-*, with as only exception the allomorph PEN_{penge}-, for which only 18 words are attested. *PEN-* is productive for agents and instruments, whereas *PE-* is productive for agent nouns and to some small extent for patient nouns. Nouns with *PE-* and *PEN-* reveal the same productivity paradox that was reported by Krott et al. (1999) for derivation and compounding. Krott et al. observed that less productive morphological categories are used more intensively as input for further word formation. In our data, we likewise find that the less productive prefix, *PE-*, appears with more variants compared to *PEN-*.

Whereas words with *PE-* are more readily accessible for further inflection compared to *PEN-* (see Figure 3.4), words with *PEN-* emerge as paradigmatically more entrenched. Verbs to which *PEN-* attaches tend to allow for more verbal affixation than is the case for verbs to which *PE-* attaches (see Figure 3.5). Furthermore, the productivity of the allomorphs of *PEN-* mirrors the productivity of the allomorphs of their base words with *MEN-* (see Figure 3.2). The proportionalities that govern the types and hapaxes of the allomorphs of *MEN-* and *PEN-* does not extend to *BER-* and *PE-*. In fact, *PE-* is surprisingly uncommon with base verbs with *BER-*, which is not what standard descriptions in the literature — *PEN-* is derived from *MEN-*,

PE- is derived from *BER-* (Chaer, 2008; Ramlan, 2009; Ermanto, 2016; Sneddon et al., 2010; Putrayasa, 2008; Dardjowidjojo, 1983; Benjamin, 2009) — would lead one to expect.

It is well known that the productivity of an affix can vary depending on the structure of its base words (Aronoff, 1976; Baayen and Renouf, 1996). Nevertheless, it is surprising to see an almost perfect linear relation between the productivity of the allomorphs of *MEN-* and the productivity of the allomorphs of *PEN-*, both with respect to types and with respect to hapax legomena. This linear relationship strongly supports analyses according to which the variant forms of *PEN-* and *MEN-* are allomorphs. Our examination of the use of *PE-* and *PEN-* in written Indonesian revealed some novel uses that have not been noted in the preceding literature on allomorphy.

This raises the question of whether *PE-* should be considered to be yet another allomorph of *PEN-*. Several observations argue against this possibility. First, *PE-* does not participate in the linear dependence that characterizes the productivity of the allomorphs of *MEN-* and *PEN-*. Second, our data indicate that *PEN-* has a strong preference for verbs as base words, but *PE-* does not show such a preference. Third, a monomorphemic base word's verb family tends to be larger when this verb family gives rise to a nominalization with *PEN-*, but no such tendency is present for *PE-*. Fourth, the frequencies of words with *PEN-* enter into a significant correlation with the frequency of the base words, but no such correlation is present for *PE-*: the formations with *PE-* have become independent of their base words. Finally, *PE-* is proportionally overrepresented for patient nouns, whereas *PEN-* creates primarily instruments in addition to agents.

That allomorphy is to some extent a matter of degree is well known (Baayen et al., 2013; Endresen, 2014). Obviously, *PE-* is highly similar in form to *PEN-*, in fact, it is identical to one of its allomorphs (although it is possible that phonetically the two are different, see Plag et al. (2017) for durational differences between the realisations of English *-s* depending on the semantics functions expressed). Yet, even though *PE-* and *PEN-* are largely in complementary distribution, they differ substantially in their productivity, both quantitatively and qualitatively, as well as in their entrenchment in the verbal system of Indonesian.

Acknowledgement

This study was funded by Indonesia Endowment Fund for Education (*Lembaga Pengelola Dana Pendidikan*) (No. PRJ-1610/LPDP/2015).

Bibliography

Alber, B. (2011). Past participles in Mòcheno: Allomorphy, alignment and the distribution of obstruents. In Putnam, M. T., editor, *Studies on German-Language Islands*, 123, pages 33–64. John Benjamins Publishing Company, Amsterdam.

Alwi, H. (2012). *Kamus Besar Bahasa Indonesia*. Gramedia Pustaka Utama, Jakarta, fourth edition.

Arka, I. W., Dalrymple, M., Mistica, M., and Mofu, S. (2009). A linguistic and computational morphosyntactic analysis for the applicative -i in Indonesian. In Butt, M. and King, T. H., editors, *International Lexical Functional Grammar Conference (LFG)*, pages 85–105. CSLI Publications.

Aronoff, M. (1976). *Word Formation in Generative Grammar*. MIT Press, Cambridge, Mass.

Aronoff, M. and Anshen, F. (2017). Morphology and the lexicon: lexicalization and productivity. In Spencer, A. and Zwicky, A. M., editors, *The Handbook of Morphology*, pages 237–247. John Wiley & Sons, Inc.

Baayen, R. (2009). Corpus linguistics in morphology: Morphological productivity. In Lüdeling, A. and Kyto, M., editors, *Corpus Linguistics. An International Handbook*, pages 900–919. Mouton De Gruyter, Berlin.

Baayen, R., Janda, L. A., Nesset, T., Dickey, S., Endresen, A., and Makarova, A. (2013). Making choices in Russian: Pros and cons of statistical methods for rival forms. *Russian Linguistics*, 37:253–291.

Baayen, R. H. and Renouf, A. (1996). Chronicling the times: Productive lexical innovations in an English newspaper. *Language*, 72:69–96.

Benjamin, G. (2009). Affixes, Austronesian and iconicity in Malay. *Bijdragen tot de Taal-, Land- en Volkenkunde*, 165(2–3):291–323.

Bloomfield, L. (1933). *Language*. George Allen & Unwin Ltd, London.

Booij, G. (2010). *Construction Morphology*. Oxford. OUP.

Booij, G. and Lieber, R. (2004). On the paradigmatic nature of affixal semantics in English and Dutch. *Linguistics*, 42:327–357.

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

Booij, G. E. (1996). Inherent versus contextual inflection and the split morphology hypothesis. In Booij, G. E. and Marle, J. v., editors, *Yearbook of Morphology 1995*, pages 1–16. Kluwer Academic Publishers, Dordrecht.

Chaer, A. (2008). *Morfologi Bahasa Indonesia (Pendekatan Proses)*. PT Rineka Cipta, Jakarta.

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

Dardjowidjojo, S. (1983). *Some Aspects of Indonesian Linguistics*. Djambatan, Jakarta.

Denistia, K. (2018). Revisiting the Indonesian prefixes peN-, pe2-, and per-. *Linguistik Indonesia*, 36(2):145–159.

Dressler, W. (1989). Prototipical differences between inflection and derivation. *Zeitschrift für sprachwissemnschaft und kommunikationsforschung*, 42:3–10.

Endresen, A. (2014). *Non-standard allomorphy in Russian prefixes: Corpus, experimental, and statistical exploration*. PhD thesis, Faculty of Humanities, Social Sciences and Education,

The Artic University of Norway.

Ermanto (2016). *Morfologi Afiksasi Bahasa Indonesia Masa Kini: Tinjauan dari Morfologi Derivasi dan Infleksi*. Kencana, Jakarta.

Fortin, C. R. (2006). Reconciling meng- and NP movement in Indonesian. *Berkeley Linguistics Society and the Linguistic Society of America*, (2):47–58.

Goldhahn, D., Eckart, T., and Quasthoff, U. (2012). Building large monolingual dictionaries at the Leipzig Corpora Collection: From 100 to 200 languages. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation*, pages 1799–1802.

Hidajat, L. (2014). A distributed morphology analysis of Indonesian ke-/-an verbs. *Linguistik Indonesia*, 32(1):11–31.

Kridalaksana, H. (2007). *Kelas Kata dalam Bahasa Indonesia*. Gramedia Pustaka Utama, Jakarta, second edition.

Kridalaksana, H. (2008). *Kamus Linguistik*. PT Gramedia Pustaka Utama, Jakarta, 4th edition.

Kroeger, P. R. (2007). Morphosyntactic vs. morphosemantic functions of Indonesian –kan. In Zaenen, A., Simpson, J., King, T. H., Jane, G., Maling, J., and Manning, C., editors, *Architectures, Rules, and Preferences: Variations on Themes of Joan Bresnan*, number 184 in CSLI Lecture Notes, pages 229–251. CSLI Publications, Stanford, California.

Krott, A., Robert, S., and Baayen, R. (1999). Complex words in complex words. *Linguistics*, 37:905–926.

Krott, A., Robert, S., and Baayen, R. (2009). Semantic influence on linkers in Dutch noun-noun compounds. *Folia Linguistica*, 36:7–22.

Larasati, S., Kuboň, V., and Zeman, D. (2011). Indonesian morphology tool MorphInd: Towards an Indonesian corpus. In C., M. and M., P., editors, *Systems and Frameworks for Computational Morphology*, volume 100, pages 119–129. Springer.

Marle, J. v. (1986). The domain hypothesis: the study of rival morphological processes. *Linguistics*, 24:601–627.

Nuriah, Z. (2004). The relation of verbal Indonesian affixes men- and -kan with argument structure. Master's thesis, Utrecht University, Netherland.

Peters, P. (2004). *The Cambridge Guide to English Usage*. Cambridge University Press, Cambridge.

Plag, I. (2006). The variability of compound stress in English: structural, semantic and analogical factors. *English Language and Linguistics*, 10(1):143–172.

Plag, I., Homann, J., and Kunter, G. (2017). Homophony and morphology: The acoustics of word-final S in English. *Journal of Linguistics*, 53(1):181–216.

Putrayasa, I. B. (2008). *Kajian Morfologi: Bentuk Derivasional dan Infleksional*. PT Refika Aditama, Bandung.

R Team, D. C. (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

R Team, S. (2015). *RStudio: Integrated Development for R. RStudio*. RStudio, Inc., Boston, MA.

Rafferty, E. (2002). Reduplication of nouns and adjectives in Indonesian. *Papers from the Tenth Annual Meeting of the Southeast Asian Linguistics Society*, pages 317–332.

Ramlan, M. (2009). *Morfologi: Suatu Tinjauan Deskriptif*. CV Karyono, Yogyakarta.

Schreuder, R., Neijt, A., Van der Weide, F., and Baayen, R. H. (1998). Regular plurals in Dutch compounds: Linking graphemes or morphemes? *Language and cognitive processes*, 13:551–573.

Sneddon, J. N., Adelaar, A., Djenar, D. N., and Ewing, M. C. (2010). *Indonesian: A Comprehensive Grammar*. Routledge, New York, second edition.

Stekauer, P. (2014). Derivational paradigms. In Lieber, R. and Štekauer, P., editors, *The Oxford Handbook of Derivational Morphology*, pages 354–369. Oxford University Press.

Sudaryanto (1993). *Metode dan Aneka Teknik Analisis Bahasa: Pengantar Penelitian Wahana Kebudayaan secara Linguistis*. Duta Wacana University Press, Yogyakarta.

Sugerman (2016). *Morfologi Bahasa Indonesia: Kajian ke Arah Linguistik Deskriptif*. Penerbit Ombak, Yogyakarta.

Sukarno (2017). The behaviours of the general nasal /N/ in Indonesian active prefixed verbs. *International Journal of Language and Linguistics*, 4(2):48 – 52.

Sutanto, I. (2002). Verba berkata dasar sama dengan gabungan afiks men-i atau men-kan. *Makara, Sosial-Humaniora*, 6(2):82–87.

Tomasowa, F. H. (2007). The reflective experiential aspect of meaning of the affix -i in Indonesian. *Linguistik Indonesia*, 25(2):83–96.

Wood, S. (2006). *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC.

Wood, S. (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.

Zeileis, A., Meyer, D., and Hornik, K. (2007). Residual–based shadings in visualizing (conditional—) independence. *Journal of Computational and Graphical Statistics*, 16(3):507–525.

Zipf, G. (1935). *The Psycho-Biology of Language*. Houghton Mifflin, Boston.

Zipf, G. (1949). *Human Behaviour and the Principle of the Least Effort. An Introduction to Human Ecology*. Hafner, New York.

Chapter 4

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

This chapter has been accepted, pending minor revision, for publication in *Corpus Linguistics and Linguistic Theory*: Denistia, K., Shafaei-Bajestan, E., and Baayen, H. "Exploring semantic differences between the Indonesian prefixes *PE-* and *PEN-* using a vector space model".

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 se-

mantic 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: distributional semantics, cosine similarity, similarity judgments, Indonesian, morphology

4.1 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, 2008). 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 format their allomorphs as subscripts: PEN_{peng}-, PEN_{pen}-, PEN_{pem}-, PEN_{peny}-, PEN_{penge}-; and one non-nasalized allomorph PEN_{pe}-. The second prefix, *PE-*, is clearly similar in form, and has been argued to be very similar also in meaning as PEN_{pe}- (Nomoto, 2006)[9].

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., Dardjowidjojo (1983); Nomoto (2006, 2017); Putrayasa (2008); Benjamin (2009); Ramlan (2009); Sneddon et al. (2010); Ermanto (2016)), through a process of affix substitution e.g. *petani* 'farmer'–*bertani* 'to farm' and *penari* 'dancer'–*menari* 'to dance'). Similar to *PEN-*

[9]Nomoto (2006) first described *PE-* as *PER-*, however, later in Nomoto (2017), he referred to *PER-* as *PE-*.

, *MEN-* has also six phonologically-conditioned allomorphs: MEN_{meng}-, MEN_{men}-, MEN_{mem}-, MEN_{meny}-, MEN_{menge}-, and MEN_{me}-.

Verbs with *MEN-* can be extended with the suffixes *-i* and *-kan* (Sutanto, 2002; Tomasowa, 2007; Kroeger, 2007; Sneddon et al., 2010). 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') (Ramli, 2006; Arka et al., 2009). 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.

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/addressor' 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-*. 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 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 Modeling (DSM), cf. 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 Firth (1957), "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 (their co-occurrences with other words in large corpora) is brought together in high-dimensional vectors, also known as word embeddings (Turney and Pantel, 2010). Thanks to the 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 judgments (e.g., Lund and Burgess (1996); Lowe and McDonald (2000); McDonald and Brew (2004)), and studies of morphological processing (Kuperman and Baayen (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 4.2, 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 judgments. Finally, we discuss the results obtained and conclude the study in Section 4.4.

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 Team, 2017). The databases and the `R` scripts are available online at `http://bit.ly/PePeNSemVector`.

4.2.1 Indonesian lemmatized database

Using the morphological analyses provided by MorphInd, we lemmatized the LCC corpus. In a preliminary processing step preceeding lemmatization, we lower-cased all words and excluded

numbers, punctuation marks, and the 15 highest frequency stop words[10]. 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; Rafferty (2002); Chaer (2008); Dalrymple and Mofu (2012); Sugerman (2016)). Several examples illustrating the output of the lemmatization process are shown in Table 4.1.

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.

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

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.

'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.'

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

Table 4.1: Examples of the lemmatization

4.2.2 Modeling 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[11]. Cosine similarity was employed to measure the degree of semantic similarity of two lemmas. Let vectors $\overrightarrow{v}$ and $\overrightarrow{w}$ be two n dimensional vectors representing two lemmas. The cosine similarity of $\overrightarrow{v}$ and $\overrightarrow{w}$ is the cosine of the angle θ between $\overrightarrow{v}$ and $\overrightarrow{w}$, and is equal to the inner product of the vectors, after being length-normalized (see equation 4.1). Thus, similarity judgment is based on the orientation, and not the magnitude, of the vectors.

$$sim(\vec{v},\vec{w}) = cos(\theta) = \frac{\vec{v}.\vec{w}}{\|\vec{v}\|\|\vec{w}\|} = \frac{\sum_{i=1}^{n} v_i w_i}{\sqrt{\sum_{i=1}^{n} v_i^2}\sqrt{\sum_{i=1}^{n} w_i^2}}. \tag{4.1}$$

4.2.3 Datasets

Using the cosine similarity, we constructed two datasets, henceforth the CosSim database and the PePeNCos database[12]. 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

[11] We used a skipgram model, with a window size of 5, a vector size of 200, and no hierarchical softmax. Items occurring less than 5 times in the corpus were discarded beforehand.

[12] In this study, we did not distinguish between impersonal agent and instrument. Impersonal agent is the term used by Booij (1986) for the meaning 'radio station' of the Dutch word *zender* which also has an agentive reading, 'one who sends', and an instrumental reading, 'transmitter'.

Lemma1	Lemma2	Cos	PrefixL1	PrefixL2	BaseWordL1	BaseWordL2	SemRoleL1	SemRoleL2
menjadi 'to become'	dalam 'inside'	-0.07	*MEN-*		jadi	dalam		
bekerja 'to work'	abadi 'eternal'	-0.07	*BER-*		kerja	abadi		
mengatakan 'to say'	menjadi 'to become'	0.09	*MEN-*	*MEN-*	kata	jadi		
melakukan 'to do'	bekerja 'to work'	0.19	*MEN-*	*BER-*	laku	kerja		
pemerintah 'government'	dalam 'inside'	-0.08	*PEN-*		perintah	dalam	agent-instrument	
petugas 'officer'	pemerintah 'government'	0.08	*PE-*	*PEN-*	tugas	perintah	agent	agent-instrument

Table 4.2: Examples of entries in the CosSim database

base word; and the word category of the base word. Example entries of this database are listed in Table 4.2.

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.

The PePeNCos database is a subset of the CosSim database and contains 116 derived words with *PE-* and 1,584 derived words with *PEN-*. This specific subset is worth separate mention since it is later useful for our analyses of the cosine similarity between base words and derived words. The database specifies the cosine similarity of the derived word and corresponding base 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; Ramlan, 2009; Sneddon et al., 2010; Nomoto, 2017). Table 4.3 presents some examples of entries in this 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

Table 4.3: Examples of entries in the PePeNCos database

4.2.4 Semantic similarity ratings

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 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 of 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; 2 *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 val-

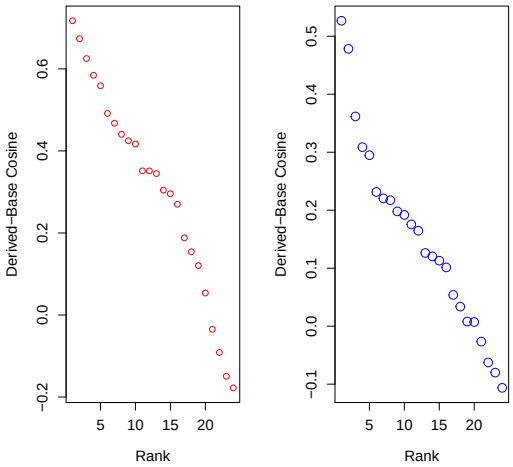


Figure 4.1: Rank distribution of cosine similarities of words with *PE-* (left panel) and words with *PEN-* (right panel) with their respective base words, as used in the semantic similarity judgment task

ues (see Figure 4.1) 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[13]. Example entries are listed in Table 4.4.

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

Table 4.4: Examples of entries of the database with human similarity ratings. Part: participant

[13]The supplementary materials are accessible at `https://osf.io/3w4hc/`

4.3 Analysis

In what follows, we first compare the semantic similarities within and between the sets of words with *PE-* and *PEN-* (Section 4.3.1). In Section 4.3.2, 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 4.3.3). Section 4.3.4 investigates semantic similarity for base words and their prefixed derivatives, and Section 4.3.5 concludes with comparing the corpus-based semantic similarities with human ratings of semantic similarity.

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

Figure 4.2, left panel, presents box plots 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 4.5. Although the distributions show considerable overlap, differences in mean cosine similarity do reach significance for the between prefix comparisons (*PE-/PEN-*) and within-prefix 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 ($p = 0.049$, see the left panel of Figure 4.2).

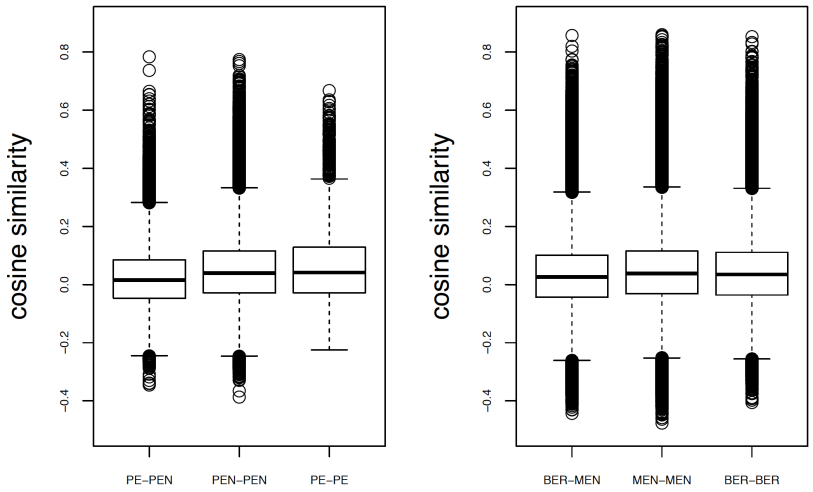


Figure 4.2: Boxplots for the distributions of cosine similarities. Left panel: cosine similarities between *PE-* and *PEN-*, within *PEN-* and within *PE-*. Within and between prefix cosine similarities, group means are significantly different only for between prefix comparisons (left panel). Right panel: cosine similarities between *MEN-* and *BER-*, within *MEN-* and within *BER-*. For these base words, all pairs of group means are significantly different

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 'mosturizer'	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

Table 4.5: 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

4.3.2 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 Dardjowidjojo (1983); Putrayasa (2008); Benjamin (2009); Ramlan (2009); Sneddon et al. (2010); Ermanto (2016); Nomoto (2017)), we investigate whether pairs of verbs 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-*. The Kruskal-Wallis rank sum test ($\chi^2_{(2)} = 34699, p < 0.0001$) and

Bonferroni-corrected pairwise tests clarified that the mean for *BER-/MEN-* pairs (0.032) was significantly smaller than those for the within-prefix pairs ($p < 0.0001$ for both comparisons); see also right panel in Figure 4.2. 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.

4.3.3 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 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 5 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 4.3, 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 ($\chi^2_{(2)} = 362.41, p < 0.0001$) 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), were 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 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-*, we expect to observe differences in mean cosine similarity between the mixed group and agents on the one hand, and the mixed group and instruments on the other hand. Specifically, we expect the mixed group to be intermediate with respect to the agents and the instruments.

A Kruskal-Wallis rank sum test ($\chi^2_{(2)} = 6895.1$, $p < 0.0001$) 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, $p < 0.0001$); see the right panel of Figure 4.3. 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 4.3 is due to the specific use of *PE-* for athletes. In order to investigate this possibility, we therefore 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 4.4, 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 ($\chi^2_{(3)} = 525.99$, $p < 0.0001$) and Bonferroni-corrected pairwise tests clarified that the mean cosine similarities of pairs within the *PE-*$_{athletes}$ set are significantly higher than those for the pairs of words in the other sets of agent nouns ($p < 0.0001$). When both *PE-*$_{athletes}$ and *PE-*$_{non\text{-}athletes}$ are merged into one set, the mean cosine similarity decreases to 0.049; see the left panel of Figure 4.3. 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 4.4, 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 $p < 0.0001$).

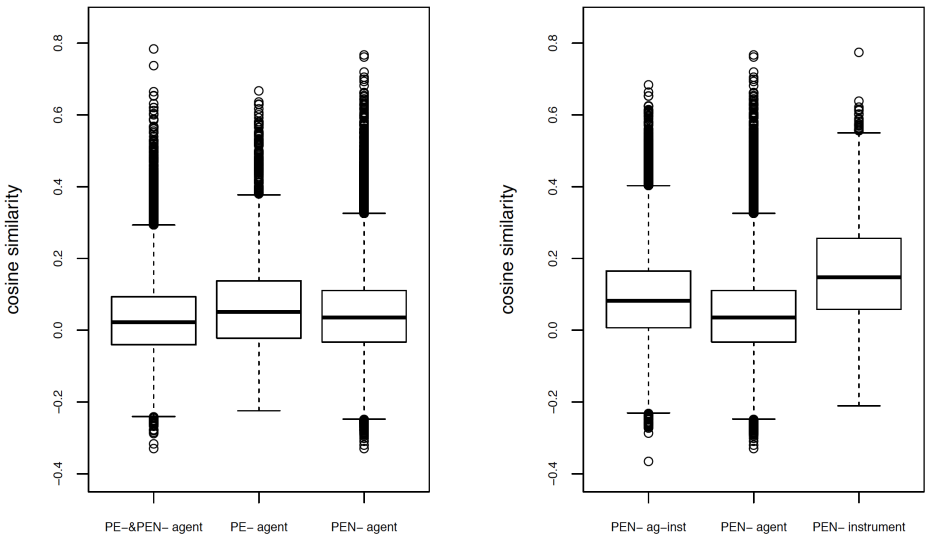


Figure 4.3: Boxplots for the distributions of cosine similarities for cross-prefix pairs of words with *PE-* and *PEN-* expressing agents, as well as for within-prefix pairs expressing agents (left panel). The right panel compares the distributions of cosine similarities for words with *PEN-*, comparing 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

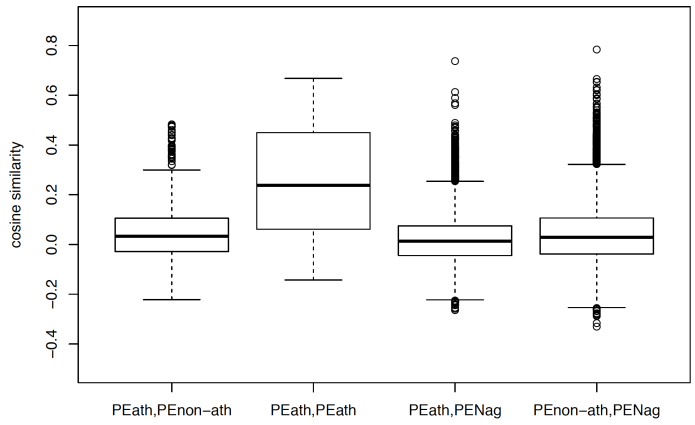


Figure 4.4: Boxplots for the cosine similarity for *PE-* partition into nouns for athletes and nouns for non-athletes, and agent nouns with *PEN-*

4.3.4 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 4.5 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 = 44626, $p < 0.0001$) 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 = 2488, $p = 0.648$ for verbs; W = 790, $p = 0.1329$ for adjectives; but W = 5932, $p < 0.0001$ for nouns). The right panel of Figure 4.5 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.

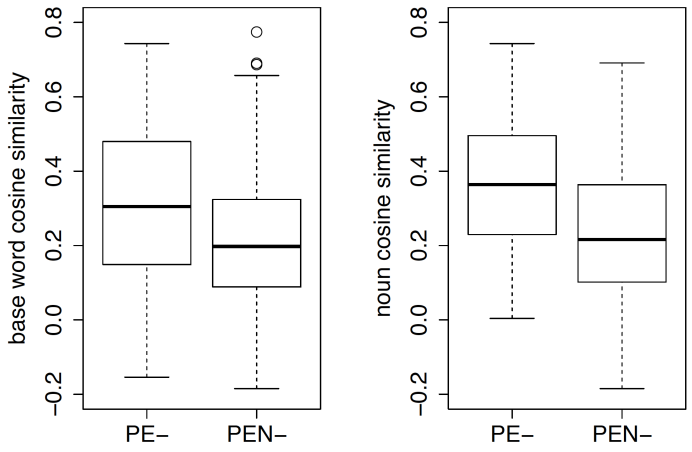


Figure 4.5: Boxplots for the distributions of cosine similarities for word pairs consisting of the base and the derived word (left panel) and the noun base and the derived word (right panel). Mean cosine similarity is higher for *PE-* compared to *PEN-* in both comparisons

4.3.5 Modelling human judgment for base-derived pairs cosine similarity

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 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 to 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, 11 never used more than 3 options of the 5 options available on the scale. 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.

Table 4.6 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 4.6. 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.

A. parametric coefficients	Estimate	Std. Error	t-value	p-value
Intercept (*PE-*)	3.5894	0.0821	43.7323	< 0.0001
B. smooth terms	edf	Ref.df	F-value	p-value
s(CosineSimilarity) [*PE-*]	2.7263	3.3935	17.8613	< 0.0001
s(CosineSimilarity) difference curve *PEN-*	8.3073	9.1748	8.3258	< 0.0001
s(FrequencyBase,FrequencyDerived)	14.4390	14.9135	28.6596	< 0.0001
Random intercepts participant	67.7460	71.0000	21.1922	< 0.0001

Table 4.6: GAMM fitted to the ratings elicited for 48 pairs of *PE-* and *PEN-* nominalizations and their base words

As can be seen by comparing the left and centre panels of Figure 4.6, 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 *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. Apparently, the cosine similarity measure is more precise for *PE-* than for *PEN-*. This is probably due to the semantic specialization of *PE-* for athletes.

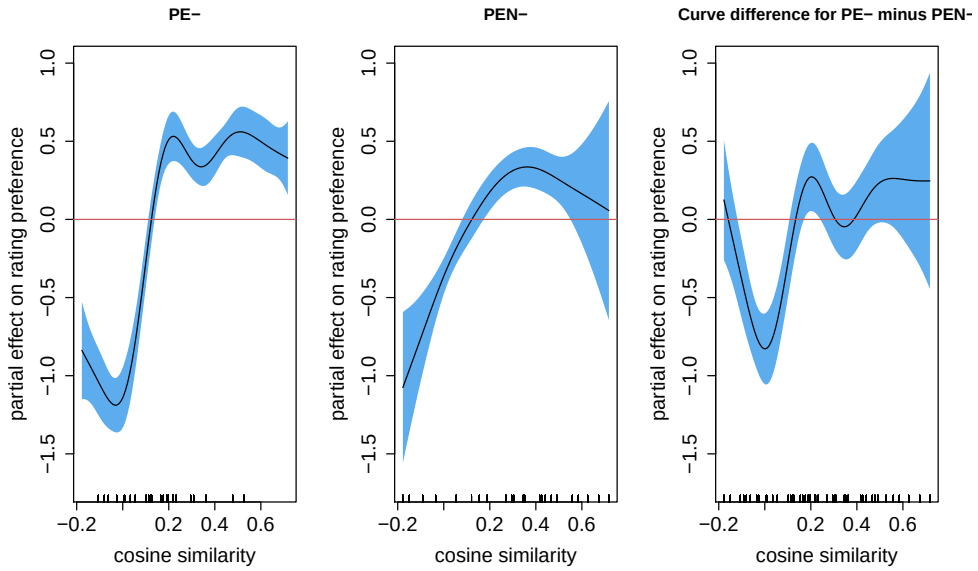


Figure 4.6: Partial effects for cosine similarity as a predictor of human ratings for *PE-* (left panel) and *PEN-* (middle panel). Right panel: the difference curve which, when added to the curve of *PEN-*, yields the curve of *PE-*

4.4 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 (Sneddon et al., 2010; Rajeg, 2013), have their own quantitative semantic profiles; if so, this would provide further support for *PE-* and *PEN-* being separate affixes rather than allomorphs (Ramlan, 2009; Sneddon et al., 2010; Nomoto, 2017; Denistia and Baayen, 2019). 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 judgments.

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-* (Dardjowidjojo, 1983; Putrayasa, 2008; Benjamin, 2009; Ramlan, 2009; Sneddon et al., 2010; Ermanto, 2016; Nomoto, 2017; Denistia and Baayen, 2019). 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 (Sutanto, 2002; Chaer, 2008; Tomasowa, 2007; Kroeger, 2007; Putrayasa, 2008; Arka et al., 2009; Sneddon et al., 2010). *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 judgments 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 4.3). Furthermore, the semantic similarities of pairs of base and derived words are greater for *PE-* than for *PEN-* (Figure 4.5). 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 (Sneddon et al., 2010; Ramlan, 2009), 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 favored, 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.

Acknowledgement

This study was funded by Indonesia Endowment Fund for Education (*Lembaga Pengelola Dana Pendidikan*) (No. PRJ-1610/LPDP/2015).

Bibliography

Alfonseca, E., Hall, K., and Hartmann, S. (2009). Large-scale computation of distributional similarities for queries. In *Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, volume Companion Volume: Short Papers, pages 29–32. Association for Computational Linguistics.

Alwi, H. (2012). *Kamus Besar Bahasa Indonesia*. Gramedia Pustaka Utama, Jakarta, fourth edition.

Arka, I. W., Dalrymple, M., Mistica, M., and Mofu, S. (2009). A linguistic and computational morphosyntactic analysis for the applicative -i in Indonesian. In Butt, M. and King, T. H., editors, *International Lexical Functional Grammar Conference (LFG)*, pages 85–105. CSLI Publications.

Baayen, R. (2009). Corpus linguistics in morphology: Morphological productivity. In Lüdeling, A. and Kyto, M., editors, *Corpus Linguistics. An International Handbook*, pages 900–919. Mouton De Gruyter, Berlin.

Baayen, R. H., Chuang, Y.-Y., Shafaei-Bajestan, E., and Blevins, J. P. (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*, pages 1–39.

Baayen, R. H. and Moscoso del Prado Martin, F. (2005). Semantic density and past-tense

formation in three Germanic languages. *Language*, 81(3):666–698.

Baayen, R. H. and Neijt, A. (1997). Productivity in context: A case study of a Dutch suffix. *Linguistics*, 35(3):565–587.

Benjamin, G. (2009). Affixes, Austronesian and iconicity in Malay. *Bijdragen tot de Taal-, Land- en Volkenkunde*, 165(2–3):291–323.

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

Chaer, A. (2008). *Morfologi Bahasa Indonesia (Pendekatan Proses)*. PT Rineka Cipta, Jakarta.

Cheung, J. C. K. and Penn, G. (2013). Probabilistic domain modelling with contextualized distributional semantic vectors. *Association for Computational Linguistics (ACL)*, pages 392–401.

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

Dardjowidjojo, S. (1983). *Some Aspects of Indonesian Linguistics*. Djambatan, Jakarta.

Denistia, K. (2018). Revisiting the Indonesian prefixes peN-, pe2-, and per-. *Linguistik Indonesia*, 36(2):145–159.

Denistia, K. and Baayen, H. (2019). The Indonesian prefixes PE- and PEN-: A study in productivity and allomorphy. *Morphology*, 29(3):385–407.

Ermanto (2016). *Morfologi Afiksasi Bahasa Indonesia Masa Kini: Tinjauan dari Morfologi Derivasi dan Infleksi*. Kencana, Jakarta.

Fam, R., Lepage, Y., Gojali, S., and Purwarianti, A. (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*, Tsukuba, Japan.

Firth, J. R. (1957). *Studies in Linguistic Analysis*, chapter A synopsis of linguistic theory 1930–1955, pages 1–32. Basil Blackwell, Oxford.

Goldhahn, D., Eckart, T., and Quasthoff, U. (2012). Building large monolingual dictionaries at the Leipzig Corpora Collection: From 100 to 200 languages. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation*, pages 1799–1802.

Jalaluddin, N. H. and Syah, A. H. (2009). Penelitian makna imbuhan pen- dalam Bahasa Melayu: Satu kajian rangka rujuk silang. *GEMA Online Journal of Language Studies*, 9(2):57–72.

Kridalaksana, H. (2007). *Kelas Kata dalam Bahasa Indonesia*. Gramedia Pustaka Utama, Jakarta, second edition.

Kridalaksana, H. (2008). *Kamus Linguistik*. PT Gramedia Pustaka Utama, Jakarta, 4th edition.

Kroeger, P. R. (2007). Morphosyntactic vs. morphosemantic functions of Indonesian –kan. In Zaenen, A., Simpson, J., King, T. H., Jane, G., Maling, J., and Manning, C., editors, *Architectures, Rules, and Preferences: Variations on Themes of Joan Bresnan*, number 184 in CSLI Lecture Notes, pages 229–251. CSLI Publications, Stanford, California.

Kuperman, V. and Baayen, R. H. (2009). Semantic transparency revisited. In *Presentation at the 6th International Morphological Processing Conference*.

Landauer, T. and Dumais, S. (1997). A solution to plato's problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*, 104(2):211–240.

Larasati, S., Kuboň, V., and Zeman, D. (2011). Indonesian morphology tool MorphInd: Towards an Indonesian corpus. In C., M. and M., P., editors, *Systems and Frameworks for Computational Morphology*, volume 100, pages 119–129. Springer.

Lazaridou, A., Marelli, M., Zamparelli, R., and Baroni, M. (2013). Compositional-ly derived representations of morphologically complex words in distributional semantics. In *Pro-*

ceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), volume 1, pages 1517–1526.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, 140:5–55.

Lowe, W. and McDonald, S. (2000). The direct route: Mediated priming in semantic space. Technical report, The University of Edinburgh.

Lund, K. and Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*, 28(2):203–208.

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

McCarthy, D., Koeling, R., Weeds, J., and Carroll, J. (2007). Unsupervised acquisition of predominant word senses. *Computational Linguistics*, 33(4):553–590.

McDonald, S. and Brew, C. (2004). A distributional model of semantic context effects in lexical processing. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, page 17. Association for Computational Linguistics.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

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

Nomoto, H. (2006). A study on complex existential sentences in Malay. Master's thesis, Universiti Bahasa Asing Tokyo, Tokyo.

Nomoto, H. (2017). The syntax of Malay nominalization. In Razak, R. A. and Yusoff, R., editors, *Aspek Teori Sintaksis Bahasa Melayu*, pages 71–117. Dewan Bahasa dan Pustaka, Kuala Lumpur.

Pantel, P. (2005). Inducing ontological co-occurrence vectors. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 125–132. Association for Computational Linguistics.

Pastika, I. W. (2012). Klitik -nya dalam Bahasa Indonesia. *Adabiyat*, 11(1):122–142.

Putrayasa, I. B. (2008). *Kajian Morfologi: Bentuk Derivasional dan Infleksional*. PT Refika Aditama, Bandung.

R Team, D. C. (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Rafferty, E. (2002). Reduplication of nouns and adjectives in Indonesian. *Papers from the Tenth Annual Meeting of the Southeast Asian Linguistics Society*, pages 317–332.

Rajeg, G. P. W. (2013). Metonymy in Indonesian prefixal word formation. *Lingual: Journal of Language and Culture*, 1(2):64–81.

Ramlan, M. (2009). *Morfologi: Suatu Tinjauan Deskriptif*. CV Karyono, Yogyakarta.

Ramli, M. S. (2006). Imbuhan dan penandaan tematik dalam Bahasa Melayu. *Jurnal Melayu*, 2:47–54.

Rubenstein, H. and Goodenough, J. B. (1965). Contextual correlates of synonymy. *Communications of the ACM*, 8(10):627–633.

Sneddon, J. N., Adelaar, A., Djenar, D. N., and Ewing, M. C. (2010). *Indonesian: A Comprehensive Grammar*. Routledge, New York, second edition.

Sugerman (2016). *Morfologi Bahasa Indonesia: Kajian ke Arah Linguistik Deskriptif*. Penerbit Ombak, Yogyakarta.

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, D. (2016). *Pedoman Umum Ejaan Bahasa Indonesia*. Badan Pengembangan dan Pembinaan Bahasa Kementerian Pendidikan dan Kebudayaan, 4 edition.

Sutanto, I. (2002). Verba berkata dasar sama dengan gabungan afiks men-i atau men-kan. *Makara, Sosial-Humaniora*, 6(2):82–87.

Tomasowa, F. H. (2007). The reflective experiential aspect of meaning of the affix -i in Indonesian. *Linguistik Indonesia*, 25(2):83–96.

Turney, P. D. and Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, 37:141–188.

Wood, S. (2006). *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC.

Wood, S. (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.

Chapter 5

Affix substitution in Indonesian: A computational modeling approach

This chapter has been submitted to *Linguistics: An Interdisciplinary Journal of the Language Sciences*: Denistia, K. and Baayen, H. "Affix substitution in Indonesian: A computational modeling approach".

Abstract

Indonesian has two noun-forming prefixes, *PE-* and *PEN-*, that are similar in both form and meaning. These prefixes, which are probably not allomorphs (Denistia and Baayen, 2019), often stand in a paradigmatic relation to verbal base words with the prefixes *BER-* and *MEN-* (Dardjowidjojo, 1983). The central question addressed in the present study is whether the form similarities between *PEN-* (and its allomorphs) and *MEN-* (and its allomorphs) make this prefix easier to learn compared to *PE-*. To address this question, we made use of a computational model of lexical processing in the mental lexicon, the 'discriminative lexicon' (DL) model introduced by (Baayen et al., 2019). We trained this model on 2517 word forms that were inflected or derived variants of 99 different base words. Of these 2517 word forms, 109 were nouns with *PE-* and 221 words were nouns with *PEN-*. Both the production and the compre-

hension networks of the model performed with very high accuracy for both prefixes. However, the model was able to provide more precise predictions for *PE-* as compared to *PEN-*, contrary to our initial expectations, implying that *PE-* should have a processing advantage compared to *PEN-*. There are two reasons for why *PE-* is learned more robustly than *PEN-*. First, *PE-* words tend to be longer and hence have more discriminative triphones. Second, due to cue competition with *MEN-*, the prefixal triphones of *PEN-* are less effective cues than those of *PE-*. A measure of functional load is proposed that helps clarify the relative importance of the triphones in the prefixes and those straddling the boundary between prefix and stem. Our results shed further light on the productivity paradox (Krott et al., 1999), on the role of junctural phonotactics (Seidenberg, 1987; Hay, 2003; Hay and Baayen, 2003), and on the (dis)functionality of affix substitution.

Keywords: computational modeling, paradigmatic relations, linear discriminative learning, junctural phonotactics, Indonesian morphology

5.1 Introduction

In Indonesian, there are two nominalizing prefixes: *PE-* and *PEN-*, which derive nouns with a range of similar meanings (agent, instrument, patient, location, causer), see Booij (1986) for a discussion of affixal polysemy. The prefix *PEN-* is described in the literature as having six phonologically-conditioned allomorphs which are in complementary distribution (Ramlan, 2009; Sugerman, 2016; Sukarno, 2017). The *N* in *PEN-* denotes the nasal assimilation that characterizes most of the allomorphs of this prefix: PEN_{peng}-, PEN_{pen}-, PEN_{pem}-, PEN_{peny}-, PEN_{penge}-, and one non-nasalized allomorph PEN_{pe}-, which precedes base words with initial liquids or glides. This last *PEN-* allomorph, PEN_{pe}-, is indistinguisable in form from the second prefix investigated in this study, *PE-*. Qualitative studies (Ramlan, 2009; Sneddon et al., 2010) argue that *PE-* and *PEN-* are independent prefixes. On the other hand, Dardjowidjojo (1983) and Kridalaksana (2008) take them to be allomorphs.

Many nouns with *PEN-* are derived by affix substitution from verbs with a prefix *MEN-* that is characterized by a similar set of allomorphs as *PEN-* (Dardjowidjojo, 1983; Nomoto, 2006, 2017; Putrayasa, 2008; Benjamin, 2009; Ramlan, 2009; Sneddon et al., 2010; Ermanto, 2016). For example, the word *penari* 'dancer' is derived from the verb *menari* 'to dance'. A recent corpus study (Denistia and Baayen, 2019) revealed that that the productivity of the allomorphs of *PEN-* mirrors the productivity of the allomorphs of *MEN-*. *PE-* and its base words, on the other hand, do not show such a correlation. This is one of the reasons that Denistia and Baayen (2019) conclude that *PEN-* and *PE-* are not allomorphs.

The kind of affix substitution exhibited by *MEN-* and *PEN-* is not restricted to Indonesian, but also is found in other Austronesian languages. For instance, in Tagalog, the prefix *ma-* is a question marker for agents and the prefix *pa-* is the question marker for instruments (Dempwolff, 1934). Affix pairs that differ with respect to the initial consonant (stop versus corresponding nasal) are widespread in Austronesian languages (Kager et al., 1996; Blust, 2004; Lombardi, 2001; Halle and Clements, 1983). This raises the question of whether this kind of word formation is beneficial for learning. Returning to Indonesian *PEN-* and *MEN-*, *pengajar* 'teacher' and *mengajar* 'to teach a lesson' are derived from the same base *ajar* 'lesson'. The form similarity of the two prefixes, and the fact that they show the same kind of nasal assimilation, constitutes a pocket of regularity in the morphology of Indonesian, which may facilitate learning. However, the two prefixes only differ minimally between themselves: [p] and [m] differ only in manner of articulation. This places a high discrimination load on this manner feature, which is an idiosyncratic property within this pocket of regularity. Blevins et al. (2017) argue that there is a trade-off between predictability on the one hand, and discriminability on the other hand, with regularity facilitating prediction and irregularity supporting good discrimination. Thus, the systematicity in form variation that characterizes *PEN-* and *MEN-* might facilitate learning, whereas the minimal difference between the verb and noun prefixal forms can be detrimental for discrimination.

In what follows, we address the question of how this trade-off between systematicity and discriminability works out. We do so by comparing *PE-* with *PEN-*. In contrast to *PEN-* and *MEN-*, where we have a clear pocket of regularity (see Table 5.1), *PE-* is on its own,

with no systematic paradigmatic form similarities. To carry out this comparison between *PE-* and *PEN-*, we will focus on the functional load of their triphones, i.e., phones but with their left and right immediate context. Martinet (1952) argued that the functional load of phones is specific to the phonological system of a given language. The computational quantification of functional load is usually implemented at the phone level, by comparing minimal pairs (Wedel et al., 2013; Oh et al., 2015). In the present study, however, we will operationalize functional load using the theory of the discriminative lexicon (DL Baayen et al., 2019). Within this theory of the mental lexicon, linear discriminative learning (LDL) is the computational engine for mapping forms onto meanings (comprehension) and meanings onto forms (production). LDL is a computational formalization of Word and Paradigm Morphology (Matthews, 1974, 1991; Blevins, 2003, 2006, 2016; Baayen et al., 2018; Chuang et al., 2019). For studying paradigmatic relations in the lexicon (for the more general importance of paradigmatic relations, see also van Marle, 1984; Stekauer, 2014; Hathout and Namer, 2019), it turns out to be a useful tool.

Noun	English Noun	Verb	English Verb	Noun	English Noun	Verb	English Verb
pencinta	who loves something	**men**cinta	to love	**pe**cinta	lover	**ber**cinta	to make love
peninju	who punches	**men**inju	to punch	**pe**tinju	boxer	**ber**tinju	to do boxing
pengecek	checker	**menge**cek	to check	**pe**tani	rice farmer	**ber**tani	to do rice farming
pelukis	painter	**me**lukis	to paint	**pe**lari	runner	**ber**lari	to run
pengajar	teacher	**meng**ajar	to teach	**pe**kasih	love potion	kasih	love
penyumbang	donator	**meny**umbang	to donate	**pe**suruh	who is commanded	suruh	order
pembaca	reader	**mem**baca	to read	**pe**golf	golf player	golf	golf

Table 5.1: Examples of paradigmatic parallelism for *PEN-* and *MEN-*, and for *PE-* and *BER-* and *PE-* and other base words. Nasal allomorphy is restricted to word pairs with *PEN-* and *MEN-*.

The remainder of this study is structured as followed. We first introduce LDL as our computational engine for probing the paradigmatics of *PE-* and *PEN-*. We then present the dataset that we constructed and on which we trained the model. Following this, we present our computational analyses of the learnability of *PE-* and *PEN-*. We conclude with a general discussion.

5.2 Linear discriminative learning

Linear discriminative learning provides a computational framework for setting up mappings between numeric vectors representing words' forms and numeric vectors representing words' meanings. These mappings can be conceptualized as building on two-layer networks without any hidden layers, or equivalently as using the mathematics of multivariate multiple regression. The performance of linear discriminative learning has been studied for English (Baayen et al., 2019), German (Baayen and Smolka, 2020) and Estonian (Chuang et al., 2019). It has also been successfully used to study the lexical processing of auditory nonwords (Chuang et al., 2020) and to model a double dissociation in aphasia (Heitmeier and Baayen, 2020). We will use the toy lexicon in Table 5.2 to illustrate how LDL works.

Lexeme	Word	Animacy	Concreteness	SemanticRole
ajar	ajar	inanimate	abstract	
ajar	pengajar	animate	concrete	agent
tani	petani	animate	concrete	agent

Table 5.2: An example lexicon with three word forms and their inflectional features.

When modeling comprehension, the model has to learn a mapping from words' forms to their meanings. The form representations that we use are based on triphones, which are context-sensitive phones. As the Indonesian spelling system is very transparent, we approximated triphones by letter trigrams. For example, for the word *ajar* 'lesson', we obtain the triphones `#aj, aja, jar, ar#`. Here, the # symbol denotes a word boundary. Equation (**??**) shows the form matrix $\boldsymbol{C}$,

$$\boldsymbol{C} = \begin{array}{c} \\ \textit{petani} \\ \textit{pengajar} \\ \textit{ajar} \end{array} \begin{array}{c} \begin{array}{cccccccccccccc} \text{\#pe} & \text{pet} & \text{eta} & \text{tan} & \text{ani} & \text{ni\#} & \text{pen} & \text{eng} & \text{nga} & \text{gaj} & \text{aja} & \text{jar} & \text{ar\#} & \text{\#aj} \end{array} \\ \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{bmatrix} \end{array}, \tag{5.1}$$

for the lexicon shown in Table 5.2. The i-th row of $\boldsymbol{C}$ for specifies for word i which triphones it contains. When a triphone is present, it is coded with 1, if a triphone is absent in that word, it is coded with 0. In this way, we obtain numeric vectors for words' forms.

The next step is to set up numeric vectors for these words' meanings. Numeric semantic vectors are widely used in distributional semantics, and can be derived in many ways from text corpora (see, e.g., Landauer and Dumais, 1997; Mikolov et al., 2013). In this study, following Chuang et al. (2019, 2020); Baayen et al. (2018), we make use of simulated semantic vectors. These vectors are constructed as follows. First, every elementary semantic feature in Table 5.2, henceforth referred to as lexomes, is coupled with a vector of random numbers that follow a normal distribution. For the lexomes in Table 5.2, these randomly generated vectors can look like those in matrix $\boldsymbol{A}$.

$$\boldsymbol{A} = \begin{array}{c} \\ \textit{animate} \\ \textit{inanimate} \\ \textit{concrete} \\ \textit{abstract} \\ \textit{agent} \\ \textit{tani} \\ \textit{ajar} \end{array} \begin{array}{c} \begin{array}{cccccccccccccc} \text{S1} & \text{S2} & \text{S3} & \text{S4} & \text{S5} & \text{S6} & \text{S7} & \text{S8} & \text{S9} & \text{S10} & \text{S11} & \text{S12} & \text{S13} & \text{S14} \end{array} \\ \begin{bmatrix} 2.548 & -0.417 & -0.421 & -0.719 & -2.106 & 1.993 & 0.386 & 1.101 & 1.531 & -1.125 & -0.682 & 1.388 & -1.598 & 0.203 \\ 1.132 & 1.968 & 1.425 & 1.525 & 1.308 & 1.009 & 1.696 & 2.041 & 1.475 & 1.728 & 3.828 & 1.626 & 3.515 & 1.847 \\ 0.511 & 0.297 & 0.186 & 0.308 & 0.400 & 1.302 & -0.525 & 2.306 & 2.557 & -0.569 & 0.224 & -0.999 & -1.144 & -0.479 \\ 1.628 & -0.688 & 0.006 & 0.090 & 1.529 & 1.181 & 0.360 & 0.957 & -1.240 & -1.043 & 1.117 & 2.229 & 0.624 & 1.429 \\ 2.098 & 1.124 & 1.564 & 1.173 & 1.865 & 1.508 & 0.892 & 0.248 & 1.524 & 1.655 & 1.963 & 0.672 & 2.146 & 0.931 \\ 1.514 & 2.015 & 0.311 & 1.115 & 1.304 & 0.577 & 2.242 & -0.218 & -0.022 & 1.178 & 0.557 & 2.370 & 2.764 & 0.144 \\ -0.486 & 0.123 & -2.523 & -0.876 & 0.248 & -3.041 & -2.960 & 1.025 & -0.777 & -0.389 & 0.553 & -1.853 & -1.281 & -0.557 \end{bmatrix} \end{array} \tag{5.2}$$

In order to obtain the semantic vector of a given word form, we take the pertinent row vectors from $\boldsymbol{A}$ and sum them. For instance, the semantic vector of *pengajar* 'teacher' is just the sum of $\overrightarrow{animate} + \overrightarrow{concrete} + \overrightarrow{agent} + \overrightarrow{ajar}$. Thus, the value on the first semantic dimension for *pengajar*, 4.671, is obtained by summing $2.548 + 0.511 + 2.098 - 0.486$ in the first column of matrix $\boldsymbol{A}$. This procedure is repeated for each word, and results in the semantic matrix $\boldsymbol{S}$:

$$\boldsymbol{S} = \begin{array}{c} \\ \textit{petani} \\ \textit{pengajar} \\ \textit{ajar} \end{array} \begin{array}{c} \begin{array}{cccccccccccccc} \text{S1} & \text{S2} & \text{S3} & \text{S4} & \text{S5} & \text{S6} & \text{S7} & \text{S8} & \text{S9} & \text{S10} & \text{S11} & \text{S12} & \text{S13} & \text{S14} \end{array} \\ \begin{bmatrix} 6.671 & 3.019 & 1.640 & 1.877 & 1.464 & 5.381 & 2.994 & 3.436 & 5.590 & 1.139 & 2.062 & 3.431 & 2.168 & 0.799 \\ 4.671 & 1.127 & -1.194 & -0.114 & 0.408 & 1.762 & -2.208 & 4.680 & 4.835 & -0.427 & 2.057 & -0.792 & -1.877 & 0.099 \\ 2.274 & 1.403 & -1.091 & 0.739 & 3.085 & -0.851 & -0.904 & 4.023 & -0.542 & 0.297 & 5.498 & 2.003 & 2.859 & 2.719 \end{bmatrix} \end{array} \tag{5.3}$$

Given form matrix $\boldsymbol{C}$ and semantic matrix $\boldsymbol{S}$, we can map the row vectors of $\boldsymbol{C}$ onto the row vectors of $\boldsymbol{S}$ using the transformation matrix $\boldsymbol{F}$, which can be obtained by solving

$$\boldsymbol{CF} = \boldsymbol{S}. \tag{5.4}$$

For production, we are interested in the matrix $\boldsymbol{G}$ that maps the row vectors of the semantic matrix $\boldsymbol{S}$ onto the row vectors of the form matrix $\boldsymbol{C}$:

$$\boldsymbol{SG} = \boldsymbol{C}. \tag{5.5}$$

Details on how to calculate $\boldsymbol{F}$ and $\boldsymbol{G}$ are given in Baayen et al. (2019) and Baayen et al. (2018).

The matrices $\boldsymbol{F}$ and $\boldsymbol{G}$ can be conceptualized as fully connected simple networks, without any hidden layers. The comprehension network takes form features (triphones) as input, and generates a vector of real values on the output units, thus creating a meaning in the model's semantic space. The production network takes a meaning in semantic space, and maps it to a vector that specifies, for each triphone, the amount of support this triphone receives from the word's semantics.

Just as in regression, a straight line cannot pass through all the data points, the semantic vectors that are predicted using the mapping (or network) $\boldsymbol{F}$ are approximate. Following notational conventions in statistics, we denote the predicted, and necessarily approximate, semantic vectors by $\hat{\boldsymbol{S}}$:

$$\boldsymbol{CF} = \hat{\boldsymbol{S}} \tag{5.6}$$

Likewise, the predicted form vectors are denoted as $\hat{\boldsymbol{C}}$:

$$\boldsymbol{SG} = \hat{\boldsymbol{C}} \tag{5.7}$$

The evaluation of the model's comprehension accuracy proceeds by examining how close the model's predicted semantic vectors are to the gold standard semantic vectors in $\boldsymbol{S}$. This idea is formalized by constructing the correlation matrix $\boldsymbol{R}_s$ that specifies for each row vector of the predicted semantic matrix $\hat{\boldsymbol{S}}$ how well it correlates with the semantic vectors of $\boldsymbol{S}$. The word the semantic vector $\boldsymbol{s}$ of which has the highest correlation with the predicted semantic vector $\hat{\boldsymbol{s}}$ is then chosen as the predicted meaning.

For production, the evaluation process is more complex because a predicted form vector $\hat{\boldsymbol{c}}$ specifies the amount of support for the different triphones, but this does not provide any information about the proper ordering of the triphones for the articulation of the target word. As a first step, the evaluation algorithm removes all triphones that have an a mount of semantic support less than a given threshold θ. In a second step, the algorithm constructs all possible sequences of triphones that satisfy three conditions: (1) the sequence should begin with a #-initial triphone, (2) it should end with a #-final triphone, and (3) any two consecutive triphones in the sequence should properly overlap, where proper overlap is defined as the first two phones of the second triphone being identical to the second and third phones of the first triphone. Thus, `ABC` and `BCD` properly overlap, but `ABC` and `PCD` do not. Finally, the algorithm calculates for each path the corresponding semantic vector using equation (5.6) and selects that path for articulation for which the predicted semantic vector is closest to the semantic vector targeted for production.

5.2.1 Dataset

The initial data was retrieved from Leipzig Corpora Collection available at `http://corpora2.informatik.uni-leipzig.de/download.html`, accessed on August 2016. From this corpus, which currently consists of 7,964,109 different word types and 1,206,281,985 word tokens, we first selected 99 mono-morphemic adjectives, verbs, nouns, and adverbs for which the highest counts of derived words are attested, and for which at least one derived word with *PE-* or *PEN-* is attested. Monosyllabic base words, which are usually low frequency words, were not included in our dataset as they do not have as many derivations and inflections as the selected 99 base words. As a consequence, the allomorphs of PEN_{penge}- and MEN_{menge}- were not present in our dataset. We then added these derived words to our dataset, and also included inflected forms (e.g., *-ku*, *-mu*, and *-nya* for first, second, and third person singular possessives or objects, *ku-* and *kau-* for first and second person subjects, as well as the marker of emphasis *-lah* and the question marker *-kah* (Kridalaksana, 2008; Sneddon et al., 2010). This procedure resulted in a dataset with 3010 words comprising 183 adjectives, 38 adverbs, 1396 nouns, and

1393 verbs. Among the verbs, 521 words with *MEN-* were attested in our dataset. For most of these verbs, the corresponding word with *PEN-* is included in our database. Derived words beginning with *PEN-* that do not have a corresponding verb with *MEN-* were not included. All words were checked against the *Kamus Besar Bahasa Indonesia*, a comprehensive dictionary of Indonesian (Alwi et al., 2003), available at `https://kbbi.kemdikbud.go.id` and consulted on February 20, 2020. Words that are not attested in the dictionary, but that appear in the corpus and that have a clear interpretation given their context, were also included. In the present study, we focus on the 2517 word forms that do not involve some form of reduplication. This set of words comprises 109 words with *PE-* and 221 words with *PEN-*. The dataset is available online at `https://bit.ly/PePeNwithLDL`.

5.2.2 Modeling

We made use of the implementation of LDL in the `WpmWithLdl version 1.3.21` (Baayen et al., 2018, 2019) for R, version 3.6.2, run under (R Team, 2015). Scripts documenting the modeling steps are available online at `https://bit.ly/PePeNwithLDL`.

The form matrix $\boldsymbol{C}$ that we constructed specified, for each of the 2517 words, which of 852 letter trigrams are present in that word. As the orthography of Indonesian is transparent, the letter trigrams usually provide a good approximation of phone triplets.

For the semantic matrix $\boldsymbol{S}$, we simulated numeric vectors of length 852. These vectors were constructed by adding the vectors of a word's content lexome and its inflectional and derivational lexomes. In what follows, we provide further detail on how we set up our coding of inflectional and derivational features.

Indonesian has a rich morphology. For example, from the noun *ajar* 'lesson' a total of 57 derivational and inflectional formations can be created (see Table 5.4 for example formations). For derivation, Indonesian uses both prefixation (e.g., *ter-, ber-, meN-, di-, PE-, PEN-*), suffixation (e.g., *-an, -i, -kan*), and circumfixation (e.g., *ter-/-kan, meN-/-kan, meN-/-i,*

Semantic feature	Values
Animacy	animate; animate,inanimate; inanimate
Concreteness	abstract; concrete
Voice	active; passive
Transitivity	intransitive; transitive
ObjectSemanticRole	goal; patient object; place; recipient; recipient,place; theme; theme,beneficiary; tool
Volition	abilitative; unintentional
Manner	action; applicative; causative; distributive manner; intensity; iterative; locative; random action; reciprocal; reflective; repetitive
Aspect	condition; imperfective; perfective; process; result
SubjectSemanticRole	agent; agent-instrument; causer; instrument; location; patient; professional
State	possession; regularity; shared possession; stative
Degree	comparative; intensive degree; superlative
Gradation	gradual; non gradual
ChangeOfObject	change of form; change of instrument used; change of location; change of state
BaseRelationship	to give X; to have character trait X; to produce X; to use X
PronounPerson	first; second; third
PronounFunction	object; possessive; subject
NyaFunction	NyaDefiniteDeterminer; NyaObject; NyaPossessive; NyaSubject
Mood	emphasize; imperative; polite imperative; question

Table 5.3: Inflectional and derivational features and their corresponding values. For each value (a functional lexome), a separate numeric semantic vector was generated, following a normal distribution with mean 0 and standard deviation 1.

ber-/-an). Inflection for person (*-ku, -mu, -nya*) and mood (*-lah, -kah*) makes use of suffixation only.

Table 5.3 lists the semantic features and their lexomic values that we distinguished for our dataset. We generated a separate numeric vector for each of these values. For a given word form, only a subset of the features is relevant. For instance, the prefix *MEN-* creates active-transitive verbs. Thus, the verb *mengajar* 'to teach a lesson' is specified for the content lexome *ajar* and for the function lexomes `active`, `transitive`, and `theme`. The prefix *di-* indicates the passive. So, the word *diajar* 'to be taught' is specified as having the lexomes `passive`, `transitive`, and `theme`. Further examples are given in Table 5.4.

Derived words can be ambiguous. For instance, *berpukulan* can have either a possesive reading, [ber + [[*pukul*]$_{N}$]$_{V}$ + an]$_{N}$]$_{V}$ 'to have the ability to deliver a real punch' or a reciprocal reading [ber + [*pukul*]$_{N}$] + an]$_{V}$ 'to hit each other'. In our database, we gave *berpukulan* a reciprocal interpretation because this reading is more frequent in the corpus. To give another

Word	Animacy	Concreteness	Aspect	Manner	SemanticRole	Voice	Transitivity	ObjectSemanticRole	Volition
terajar						passive	transitive		abilitative
terajarkan				causative		passive	transitive		abilitative
berpelajaran	inanimate	abstract	result	action		active	intransitive		
mengajar						active	transitive	theme	
mengajarimu				locative		active	transitive	patient object	
diajar						passive	transitive	theme	
diajarkan						passive	transitive	theme,beneficiary	
diajarkannya						passive	transitive	theme,beneficiary	
pelajar	animate	concrete			patient				
pelajarku	animate	concrete			patient				
ajarannyalah	inanimate	abstract	result						
pelajaran	inanimate	abstract	result	action					
pembelajaranmu	inanimate	abstract	process	action					
pengajar	animate	concrete			agent				
pengajarlah	animate	concrete			agent				
ajarkan						passive	transitive	theme,beneficiary	

Manner	Aspect	State	ChangeOfObject	PronounPerson	PronounFunction	NyaFunction	Mood
causative			state				
action	result	possession					
locative				second	object		
			state				
			state	third	subject	NyaSubject	
				first	possessive		
	result			third	possessive	NyaPossessive	emphasize
action	result						
action	process	regularity		second	possessive		
							emphasize
			state				imperative

Table 5.4: Examples of Indonesian derived words for the base word *ajar*

example, the circumfix *ke-/-an* can express `result` as in *tinggi* 'high' - *ketinggian* 'height', but it can also mean 'too high'. Here, we also selected the more frequent, de-adjectival, reading, following the *Kamus Besar Bahasa Indonesia*. Further justification of this choice is provided by inflection with the possessive pronouns *-ku, -mu, -nya* that are attested in the corpus.

Sometimes, derived words with the same base can have very similar meanings, an example being the pair *pelajaran* and *ajaran*, which both mean 'lesson'. Apart from that the two words occur in different social contexts (secular versus religious education), *pelajaran* has a more active reading. We therefore coded *ajaran* as having the lexomes `ajar`, `inanimate`, `abstract`, `result`, and *pelajaran* as having the lexomes `ajar`, `inanimate`, `abstract`, `action`, `result`.

The feature `BaseRelationship` is used to discriminate between words such as *mengeras* 'to become harder' and *berkeras* 'to have a strong belief about something'. Both words share the lexomes `keras` 'hard', `active`, and `intransitive`. But *berkeras* specifies a character trait rather than a physical change of state. Other examples encoded by means of the feature

`BaseRelationship`, which occurs in 40 words with the prefix *ber-*, are listed below:

1. to `give the object` designated by the base word (*korban* 'sacrifice' - *berkorban* 'to give a sacrifice')

2. to `have a characteristic property` expressed by the base word (*waspada* 'alert' - *berwaspada* 'to be alert', *sendiri* 'alone' - *bersendiri* 'to be alone')

3. to `produce the object` denoted by the base (*suara* 'voice' - *bersuara* 'to speak up', *telur* 'egg' - *bertelur* 'to lay an egg', *usaha* 'effort' - *berusaha* 'to make an effort')

4. to `use the object` expressed by the base word (*layar* 'sail' - *berlayar* 'to sail', *dayung* 'paddle' - *berdayung* 'to use paddle')

Finally, the `ChangeOfObject` feature is needed for the suffix *-kan*. This suffix typically renders a verb explicitly transitive by adding a further argument, either a beneficiary or a causer (Arka et al., 2009; Sutanto, 2002; Tomasowa, 2007; Kroeger, 2007; Sneddon et al., 2010). When *-kan* attaches to verbs, it may provide further information about the object, either notionally or physically (Soekarno, 2010). In our dataset, changes of object with the suffix *-kan* are attested for 509 words. Here are some examples:

1. change of location

    - *dekat* 'near', *dekatkan meja itu* 'get that table closer (imperative)'

    - *datang* 'to come', *dia mendatangkan Bapak Presiden Jokowi* 'he/she makes Mr. President Jokowi come'

2. change of form

    - *musik* 'music', *puisinya dimusikkan* 'the poem is put to music'

    - *hukum* 'law', *kata-katanya dihukumkan* 'his/her words are made into law'

3. change of instrument used

    - *pukul* ‘to hit’, *memukul* ‘to hit something (by hand)’, *dia memukulkan tongkat* ‘he/she hits with a stick’

4. change of state

    - *bersih* ‘clean’, *bersihkan meja itu* ‘make that table clean (imperative)’

    - *tinggi* ‘high’, *tinggikan meja itu* ‘make that table higher (imperative)’

For all content lexomes, and for the function lexomes listed in Table 5.3, a semantic vector was generated with real-valued numbers that followed a Gaussian distribution with a standard deviation of 4, and a mean that was drawn randomly from a (0, 1)-normal distribution. The semantic vector for a given word form was obtained by summing the vector of its content lexome and the semantic vectors of all its pertinent function lexomes. Finally, we added to the vector of each word a vector of numbers drawn from a (0,1) normal distribution in order to represent the individual aspects of a word’s meaning that are not captured by the vectors of the word’s constituent lexomes.

5.2.3 Accuracy

For the 2517 different words in our dataset, comprehension accuracy, evaluated on the training data, was 93.6% (160 errors). Production accuracy was 93.8% (154 errors). Thus, overall, accuracy is high.

To see where the model encountered difficulties, we zoomed in on the set of errors made. For the set of comprehension errors, the lexeme was recognized correctly in more than 98% of the cases. Accuracies for `ChangeOfObject`, `Voice`, `PronounPerson` and `PronounFunction` were 100%, 93%, 90% and 90% respectively. Accuracy was especially low for the `Aspect` (30%), for `NyaFunction` (22%), and for `SubjectSemanticRole` (0%).

For production, the lexeme was predicted 100% correctly by the model. The same 100% accuracy also holds for `Animacy`, `Voice`, `Transitivity`, `Volition`, `Manner`, `Aspect`, `State`, `Gradation`, `ChangeOfObject`, `BaseRelationship`, `PronounPerson`, `PronounFunction`, and `Mood`. `Concreteness` accuracy was 98%, `ObjectSemanticRole` was at 92%, and `SubjectSemanticRole` was at 90%. The lowest accuracy was for `NyaFunction` (75%).

Apparently, the model was challenged most by understanding and producing words with the *-nya* suffix. Interestingly, *-nya* can realize four different lexomes, depending on which base word class it attaches to and in what context it is used. When *-nya* attaches to a noun, it expresses either definiteness (`NyaDefiniteDeterminer`) or third person singular possessive (`NyaPossessive`). In addition, *-nya* can realize third person objects (`NyaObject`) as well as third person subjects (`NyaSubject`) when it attaches to a verb. This polysemy clearly renders fragile the comprehension of words with *-nya*. Nevertheless, of the 708 words with *-nya*, a total of 651 (92%) are correctly understood, and 639 (90%) are produced correctly.

Comprehension accuracy for the *PE-* and *PEN-* words was at 98% (107 out of 109 words) and 100% (221 words) respectively. The eleven comprehension errors involving words with *PE-* or *PEN-* are listed in Table 5.5. There are seven cases where one of these prefixes is incorrectly added, there is one case where a prefix is omitted, two cases where *PE-* and *PEN-* are exchanged, and one case where the old prefix *PER-* is perceived instead of *PE-*. With one exception, the targeted word is within the top five most highly ranked candidates (see the `rank target` column in Table 5.5).

targeted form	English translation	targeted prefix	predicted_form	English translation	predicted prefix	rank target
tinggi	high	-	peninggi	sth to make sb higher	PEN-	2
besarnya	the largeness	-	pembesarnya	his/her/the magnifier	PEN-	2
petanda	sth that is marked	PE-	pertanda	sth that marks	-	2
sakitnya	his/her/the illness	-	penyakitnya	his/her/the illness	PEN-	2
pendagang	long stick to carry stuffs on shoulder	PEN-	pedagang	seller	PE-	2
penyertanya	his/her/the sth/sb that comes together	PEN-	pesertanya	his/her/the participant	PE-	2
pekasih	love potion	PE-	kekasih	one's beloved	-	3
mabuk	get drunk	-	pemabuk	sb who likes to get drunk	PE-	3
ajar	lesson	-	pengajar	teacher	PEN-	4
buatlah	make (soft imperative)	-	pembuatlah	creator (emphasize)	PEN-	6
suruh	a command	-	pesuruh	sb who is commanded	PE-	305

Table 5.5: Comprehension errors involving *PE-* and *PEN-*, including omissions and intrusions.

The error made for *pekasih*, incorrectly understood as *kekasih*, is an interesting one. It has been observed (Sneddon et al., 2010; Ramlan, 2009; Chaer, 2008; Ermanto, 2016; Subroto, 2012; Sugerman, 2016) that when *PEN-* and *PE-* are both realized for the same base word, *PEN-* expresses an agentive meaning and *PE-* expresses a patient meaning. For instance, for the base word *suruh* 'command', we have *penyuruh* 'commander' and *pesuruh*, 'the one commanded', i.e., 'maid'. The targeted word *pekasih*, 'love potion', is exceptional in that it has an instrumental reading (see also Denistia and Baayen, 2019, for a discussion of the semantic roles of *PEN-* and *PE-*). *Kekasih*, 'one's beloved', on the other hand, realizes a patient reading, a semantic role that is found for *PE-* but not for *PEN-*. In other words, *kekasih* is semantically more regular than *pekasih*, and the model clearly favors the semantically more regular form.

Another interesting comprehension error is *pertanda* instead of *petanda*. The prefix *per-* is no longer productive (Dardjowidjojo, 1983; Benjamin, 2009). However, *pertanda* expresses the more common agentive, whereas *petanda* realizes the less common patient reading. Again, we see that the model is attracted towards the form expressing the semantic role that is most common for *PE-*.

targeted form	English translation	targeted prefix	predicted_form	English translation	predicted prefix	rank target
penambak	fish farmer	PEN-	petambak	fish farmer (profession)	PE-	2
pembersihnya	his/her/the cleaner	PEN-	pembersih	cleaner	PEN-	2
penerusnya	his/her/the inheritance	PEN-	peterusnya	-	-	2
pendatanya	his/her/the data collector	PEN-	pendata	data collector	PEN-	2
pendayungnya	his/her/the person who paddles	PEN-	pendayung	sb who paddles	PEN-	2
penyakitnyalah	his/her/the illness (emphasize)	PEN-	sakitnyalah	his/her/the illness (emphasize)	-	
penyapanya	his/her/the addressor	PEN-	penyapa	addressor	PEN-	
berpembersih	having a cleaner	-	pembersih	cleaner	PEN-	
penyakitnya	his/her/the illness	PEN-	penyakit	illness	PEN-	
penyampainya	his/her/the messenger	PEN-	penyampai	messenger	PEN-	

Table 5.6: Production errors for *PE-* and *PEN-*.

Production accuracy for the *PE-* and *PEN-* words was at 100% (109 words) and 96% (212 out of 221 words) respectively. Table 5.6 lists the errors made. From ten production errors, eight cases are affix omission, and one case where *PE-* and *PEN-* are exchanged (*penambak - petambak*). Among the errors, 60% of targeted words are within the top five most highly ranked candidates. Some of the errors again occur for words in which the triphone `nya` occurs twice: *penyapanya, penyakitnya*, and *penyampainya*. One of the errors, *penyapanya*, exemplifies the cost of approximating triphones with letter trigrams. This form, which is derived from *PEN-* + *sapa* 'to greet', has as targeted trigrams `#pe, pen, eny, nya, yap, apa, pan, any,`

nya and ya#. However, the proper phonetic transcription for *penyapanya* is #pə, pəɲ, eɲa, ɲap, apa, paɲ, aɲa, ɲa#. In this transcription, there is no repeated phone sequence. In other words, the phonological form of this word is more discriminative than its orthographic form.

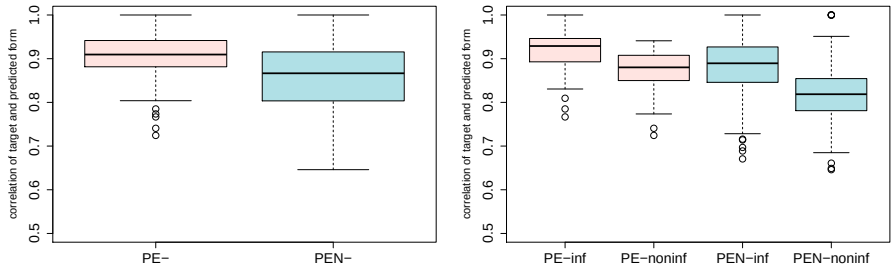
In summary, the model's accuracy for *PE-* and *PEN-* is very high. The model makes only a few errors, and in these few cases, the target words are listed among the top five candidates. Furthermore, the kind of errors that occur make sense linguistically. It is also noteworthy that the errors made are mostly existing words, and that the one case where the model produced a novel word, *peterusnya*, the word is phonotactically legal and similar to an existing word, *penerusnya*, 'the next person'. Given the good performance of the model, evaluated qualitatively in terms of whether it understands or produces the correct form, we next consider how well *PE-* and *PEN-* are learned quantitatively, and what the functional load of their triphones is.

5.3 Results

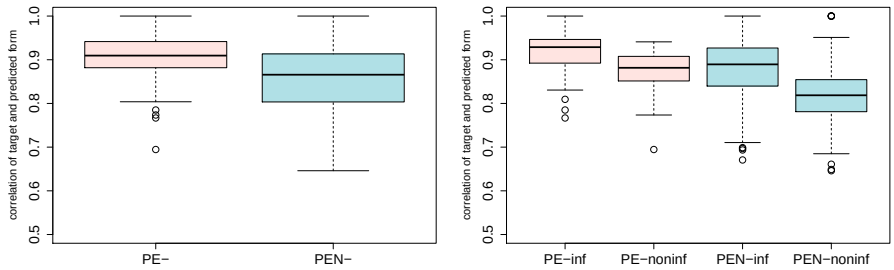
5.3.1 Quantitative differences in correlation strengths

Even though a word may be understood or produced correctly, the strength of the correlation between the predicted form vector $\hat{\boldsymbol{c}}$ and the gold standard ($\boldsymbol{c}$, production), or the strength of the correlation between the predicted semantic vector $\hat{\boldsymbol{s}}$ and the gold standard semantic vector ($\boldsymbol{s}$), can vary considerably. Figure 5.1 presents boxplots for the distribution of correlations, for comprehension (upper panels) and production (lower panels). The panels on the left side present the distributions of the correlations split by prefix. The panels on the right side split the words for a given prefix further down into uninflected and inflected forms.

For comprehension, a Wilcoxon test clarified that the mean correlation between the target and predicted form is higher for *PE-* (0.902) than for *PEN-* (0.859, $W = 17458$, $p < 0.0001$). When we subset *PE-* and *PEN-* into those words that have an inflectional exponent



(a) comprehension



(b) production

Figure 5.1: Distribution of correlations between predicted and gold standard vectors for comprehension (upper panels) and production (lower panels). For both comprehension and production, correlations are higher for *PE-* than for *PEN-*. The same pattern is visible when *PE-* and *PEN-* are subcategorized into inflected and uninflected words.

and those that do not, the same pattern is observed. For inflected *PEN-* and *PE-*, $W = 7941, p < 0.0001$, and for uninflected *PEN-* and *PE-*, $W = 1964, p < 0.0001$. For production, the pattern is virtually the same (*PE-* versus *PEN-*: $W = 17430, p < 0.0001$, inflected *PE-* and *PEN-*: $W = 1983, p < 0.0001$, uninflected *PE-* and *PEN-*: $W = 7875, p < 0.0001$).

In order to better understand why *PE-* is learned better than *PEN-*, we first removed the verbs, adverbs, and adjectives in the training data, and refitted the model. The differences shown in Figure 5.1 all disappeared, both for comprehension and for production (all $p > 0.1$). Interestingly, when only verbs with *MEN-* were removed from the training data, the mean correlation between the target and predicted forms for *PEN-* increased by 0.027 for comprehension and 0.025 for production, whereas a much reduced increase was observable for *PE-* (0.004 for comprehension and 0.002 for production). Importantly, a Wilcoxon test showed that just by removing verbs with *MEN-* from the training data, the correlations with the gold standard for *PE-* on the one hand, and those for *PEN-* on the other hand, already become very similar ($W = 13480, p = 0.0782$ for comprehension, and $W = 13721, p = 0.04$ for production). It follows that the presence of adverbs and adjectives in the training data only have a minor effect on the strength of the correlations for *PEN-* with the targeted gold standard vectors, and that the verbs with the *MEN-* are at issue.

Base word	English	Noun	Prefix	English	Verb	English	Distinct triphones	Shared triphone
ajar	lesson	pengajar	PEN-	teacher	mengajar	to teach a lesson	#pe, pen, #me, men	eng, nga, gaj, aja, jar, ar#
cinta	love	pencinta	PEN-	who keens on something	mencinta	to love	#pe, pen, #me, men	enc, nci, cin, int, nta, ta#
cinta	love	pecinta	PE-	who makes love	bercinta	to make love	#pe, pec, eci #be, ber, erc, rci	cin, int, nta, ta#
suruh	order	penyuruh	PEN-	commander	menyuruh	to give an order	#pe, pen, #me, men	eny, nyu, yur, uru, ruh, uh#
suruh	order	pesuruh	PE-	who is commanded			#pe, pes, esu	sur, uru, ruh, uh#
jalan	street	pejalan	PE-	pedestrian	berjalan	to walk	#pe, pej, eja, #be, ber, erj, rja	jal, ala, lan, an#
sakit	ill	pesakit	PE-	ill person			#pe, pes, esa, #sa	sak, aki, kit, ti#

Table 5.7: Examples of distinct and shared triphones for *PE-* and *PEN-*, and their corresponding verbal prefixes *BER-* and *MEN-*.

We can now begin to understand why *PE-* is learnt better than *PEN-*: the verbs in *MEN-* are in stronger competition with *PEN-*. This competition is illustrated in Table 5.7. When we compare nouns with *PEN-* with their paradigmatic counterparts with *MEN-*, we find

that there are two triphones that distinguish the nouns from the verbs, and that there are three triphones that the nouns and the verbs have in common. However, when we compare nouns with *PE-* with their base words (either a verb with *BER-*, or a simple nominal base), we find three or even four discriminative triphones, whereas the number of shared triphones is only two. In other words, nouns with *PE-* have more discriminative triphones compared to words with *PEN-*, whereas words with *PEN-* have more triphones that they share with their base verbs with *MEN-*.

There is one other possible reason why *PE-* is learned better than *PEN-*: words with *PE-* tend to be longer than words with *PEN-*: mean length in characters is 7.4 and 6.6 for *PE-* and *PEN-* respectively ($W = 14974, p < 0.0005$). In other words, words with *PE-* tend to have more triphones, which facilitates discrimination. Interestingly, Denistia and Baayen (2019) observed that less productive *PE-* attracts more inflectional suffixes than does more productive *PEN-*, replicating the productivity paradox observed by Krott et al. (1999). This asymmetry is also present in the current dataset, albeit as a non-significant trend. When we compare the number of words with *PE-* (109) and the number of words with *PEN-* (211) in our dataset, the probability of a word with *PE-* being inflected is 0.71, whereas for words with *PEN-*, this probability is 0.67 (however, $p = 0.529$, proportions test). Furthermore, for the 99 base words in our dataset, *PE-* attaches to fewer monomorphemic words (32) than *PEN-* (73) ($p < 0.0001$, proportions test).

5.3.2 Functional load of prefix-initial triphones

Above, we observed that the initial triphones of words with *PEN-* are crucial for distinguishing these nouns from their corresponding base verbs with *MEN-*. However, words with *PEN-* may also require discrimination from words with *PE-*, given pairs of words such as *pencinta* 'who keens on something' and *pecinta* 'who makes love'. In what follows, we explore in more detail the functional load of the triphones in the nouns with *PE-* and *PEN-*.

In order to quantify, within our discriminative approach, the functional load of a triphone, we selectively modified the model's comprehension network by setting the weights on the connections from that triphone to all outcomes to zero. In this way, we eliminate the contribution of that triphone to the predicted semantic vector $\hat{\boldsymbol{s}}$. In what follows, we refer to the semantic vector that is the result of taking out the weights for triphone τ as $\hat{\boldsymbol{s}}_\tau$. The functional load L_τ of triphone τ can now be assesed as the difference between the correlation of the original estimated vector $\hat{\boldsymbol{s}}$ with the gold standard vector $\boldsymbol{s}$ and the correlation of the gold standard vector $\boldsymbol{s}$ with the manipulated predicted vector $\hat{\boldsymbol{s}}_\tau$:

$$L_\tau = r(\boldsymbol{s},\hat{\boldsymbol{s}}) - r(\boldsymbol{s},\hat{\boldsymbol{s}}_\tau). \tag{5.8}$$

When a triphone makes an important contribution to a word's semantics, then taking it out of commission should result in a substantially reduced correlation $r(\boldsymbol{s},\hat{\boldsymbol{s}}_\tau)$, and as a consequence, its functional load L_τ will be large.

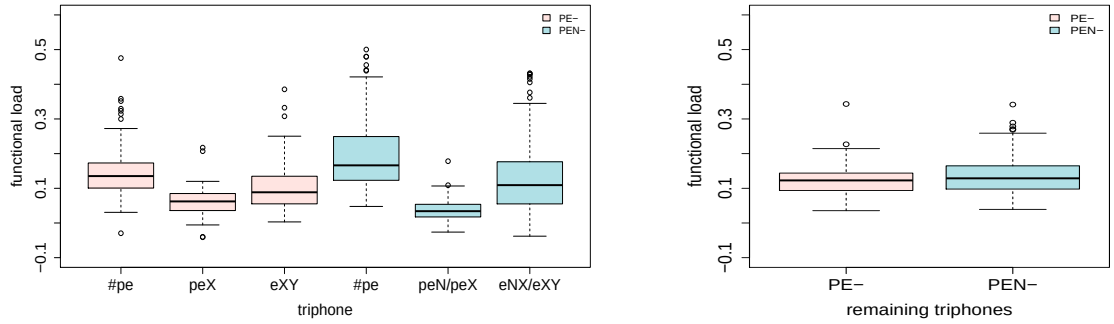


Figure 5.2: Summaries of the distribution of L_τ, using boxplots. Left panel: functional load for the first three triphones of words with *PE-* (red) and *PEN-* (blue). Right panel: average functional load of the triphones starting with the third triphone in the word up to and including the last triphone in the word.

The left panel of Figure 5.2 summarizes the distributions of the functional load of the first three triphones for *PE-* (red) and *PEN-* (blue), using boxplots. For both prefixes, the initial triphone has the largest functional load, whereas the functional load of the second triphone is the smallest. Furthermore, the differences are more pronounced for *PEN-* than for *PE-*. Wilcoxon tests clarified that the first triphone of *PE-* has a smaller functional load than

the first triphone of *PEN-* ($W = 8467, p < 0.0001$) and that the second triphone of *PE-* has a higher functional load than the second triphone of *PEN-* ($W = 17438, p < 0.0001$). There is no significant difference between the third triphones ($W = 10529, p = 0.0631$). The right panel of Figure 5.2 shows that the average functional load, calculated over the third triphone up to and including the last triphone, does not differ in the mean between *PE-* and *PEN-* ($W = 10427, p = 0.0473$). Thus, we find that the first triphone is more important for *PEN-* whereas the second triphone is more important for *PE-*. Furthermore, taking triphones in the stem out of commission affects both kinds of prefixed words equally.

What could be the reason that the first triphone has greater functional load for *PEN-* and that the second triphone has a greater load for *PE-*? To address this question, we first note that there is no significant difference for the two prefixes between the sums of the functional loads of their first and second triphones ($W = 10849, p = 0.1426$). This indicates that the two prefixes achieve a different balance of the same total functional load. An important difference between the second triphones of *PEN-* and *PE-* is that the second triphone for *PEN-*, `peN` (where `N` denotes the nasal of the pertinent allomorph) has three prefix-specific phones whereas that of *PE-*, `peX`, incorporates as its third element the first segment of the base word (in this notation, `X` denotes the first phone of the base word). As a consequence, the second triphone of *PE-* is more discriminative than that of *PEN-* (the exception being the PE_{pe}- allomorph of *PEN-*). The `peN` triphone helps reduce the set of competitors to the (still large) set of words beginning with *PEN-*, whereas `peX` reduces the set of competitors to the much smaller subset of words beginning with *PE-* and sharing the initial base word segment `X`.

Figure 5.3 presents the average functional load of the first five triphones for words with *PE-, PEN-*, and also *MEN-*. The left panel of this figure clarifies that the third triphone of words with *PEN-*, `eNX` or `eXY`, has a higher functional load compared to the second triphone: it helps reduce the set of competitors to those sharing the initial segment of the base word. At subsequent triphones further into the word, the average functional load remains fairly constant for all three prefixes.

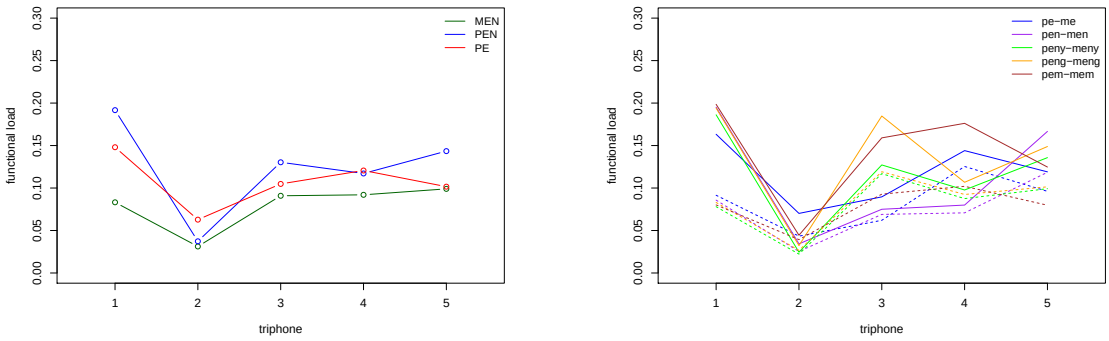


Figure 5.3: Left panel: Mean functional load of the triphones at positions 1–5 for *MEN-, PEN-*, and *PE-*. Right panel: Mean functional load of the triphones at positions 1–5 for the allomorphs of *PEN-* (solid lines) and *MEN-* (dashed lines). The low functional load for the second position, which comprises all the triphones of the prefix itself, is noteworthy.

Importantly, the frequency of the triphones is not the crucial factor determining functional load. Triphone frequencies are highest for the initial triphone `#pe`, and steadily decrease as one moves further into the word. For instance, the frequency of the triphone that fully spans one allomorph of PE_{pen}-, `pen`, 339, is higher than the mean frequency of the triphones `enX` that incorporate the first phone of the base word (`eNX`; 82.4). We return to this observation in the general discussion when we compare our discriminative approach with approaches that assume words are segmented at low-frequency boundary diphones.

The importance of specifically the initial triphone `#PE` for *PEN-* may arise because the model has to differentiate the nouns with *PEN-* not only from those with *PE-*, but also from the corresponding verbs with *MEN-*. Note that for *MEN-*, the functional load of the initial triphone is substantially smaller than that of *PEN-* ($W = 10355, p < 0.0001$). Verbs with *MEN-* occur with a wider range of inflectional and derivational affixes than is the case for *PEN-*, and hence their functional load can be spread out over more triphones. This allows the model to shift functional load forward to the initial triphone for *PEN-*.

The right panel of Figure 5.3 clarifies that the triphones that are shared by *PEN-* and *MEN-* (found at positions 3–5) show similar ups and downs in their functional load. This is probably due to the lexomes that are shared by the base verbs and the corresponding derived

nouns. A given shared triphone will support the shared semantics in a similar way for both the verb and the noun. We also note that the curve for *MEN-* is invariably located lower in the graph than the corresponding curve for *PEN-*. The reason for this is that, as mentioned above, *MEN-* occurs with a wider range of inflectional and derivational suffixes, which take their own share of the total functional load.

We should note, however, that the pattern in the left panel of Figure 5.3 presents an average for many different words, and that there can be considerable variation between words. For instance, we have not yet considered in detail the allomorphy of *PEN-*. As shown in the right panel of Figure 5.3, the different allomorphs show the same general pattern, but also exhibit considerable variation. The pattern for the PE_{pe}- allomorph is similar to that of *PE-* shown in the left panel, with a relatively high functional load for the second triphone. Furthermore, as illustrated in Figure 5.4, across different stems, functional load can vary substantially across triphone positions even when controlling for the identity of the stem. Whereas the first six panels show a pattern similar to the aggregate pattern, the lower two panels present divergent patterns.

5.4 General discussion

In this study, we addressed the question whether the prefix *PEN-* is easier to learn than its rival prefix *PE-*, thanks to *PEN-* showing a systematic relation with base words with the prefix *MEN-*. Computational modeling with linear discriminative learning revealed very high and similar accuracy for both nominal prefixes, with perhaps a small advantage in production for *PE-*. Importantly, the predicted form and meaning vectors showed stronger correlations with the targeted gold standard vectors for *PE-* as compared to *PEN-*. The presence of a difference as such dovetails well with several studies reporting qualitative and quantitative differences between these prefixes (Ramlan, 2009; Sneddon et al., 2010; Denistia and Baayen, 2019; Denistia et al., 2020). However, the present finding suggests, surprisingly, that the paradigmatic relation between *PEN-* and *MEN-* may come with a small learning disadvantage, instead of a learning

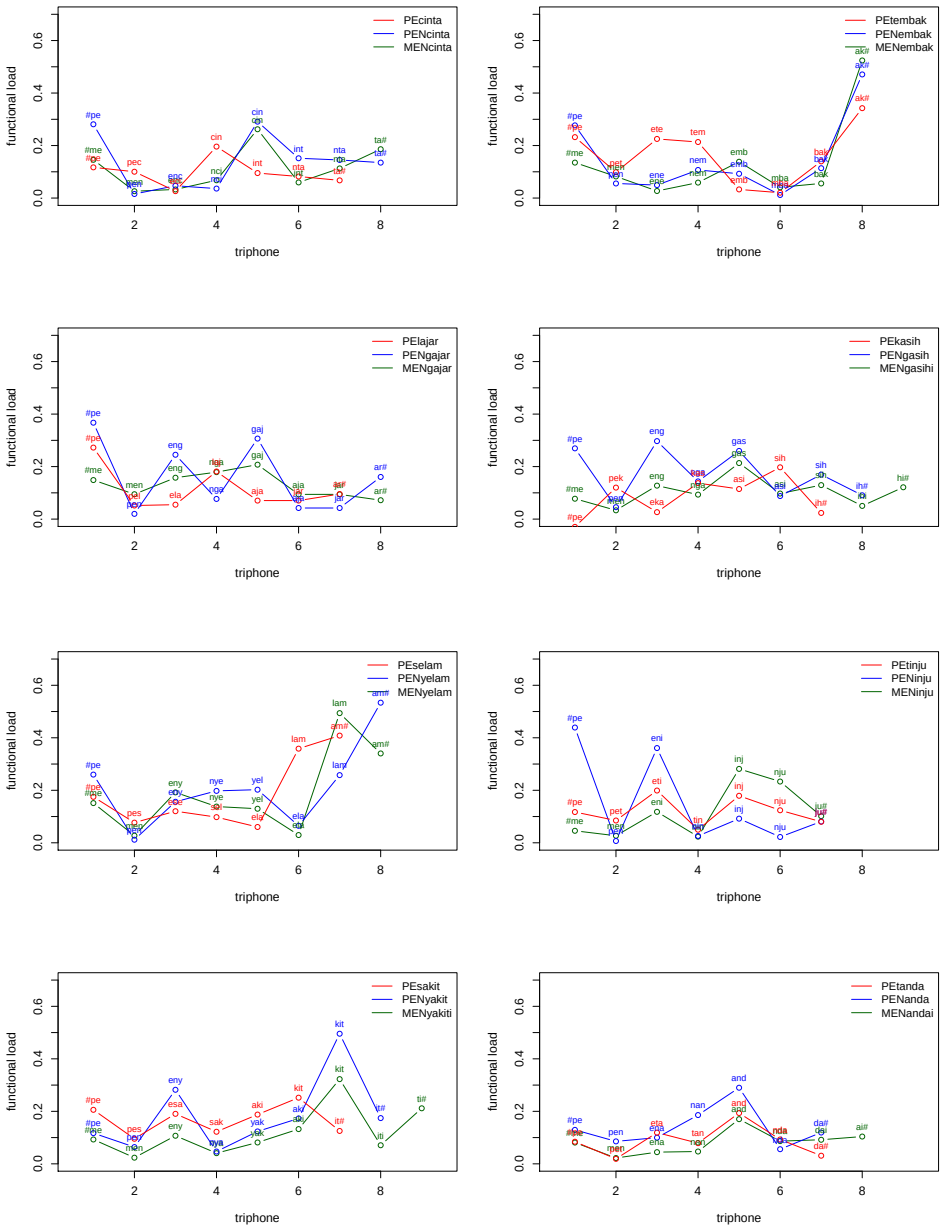


Figure 5.4: Functional load of triphones (ordered by position in the word) for word triplets with *PEN-*, *MEN-*, and *PE-* that share the same base word.

advantage. This in turn predicts that *PE-* should have a processing advantage in tasks such as word naming, visual lexical decision, and auditory lexical decision (for discrimination learning and lexical processing, see, e.g., Baayen et al., 2019; Chuang et al., 2020).

One reason that *PE-* is learned more robustly is that *PE-* has more inflected variants, which help make words with this prefix more discriminable. Denistia and Baayen (2019) observed that although *PE-* is less productive than *PEN-*, it is more often input for further word formation. This pattern exemplifies the productivity paradox reported by Krott et al. (1999): since words with less productive *PE-* are more entrenched in the lexicon, they are more readily available for further inflection. The present findings add to this understanding of the productivity paradox that the additional inflectional exponents typically found more frequently for words beginning with *PE-* makes these forms more discriminable, thereby compensating for the negative processing consequences of its lower degree of productivity.

A second reason for the more robust learning of words with *PE-* is that the triphones shared by *PEN-* and *MEN-* are in competition. For instance, the `enX` triphone cue (with `X` representing the first phone of the base word) has to compromise between the verbal and nominal meanings associated with *PEN-* and *MEN-*. Furthermore, due to the formal similarity of *PEN-* and *MEN-*, words with *PEN-* have fewer distinctive cues compared to words with *PE-*. In line with the observation of Blevins et al. (2017) that there is a trade-off between predictability and regularity, such that regularity results in better prediction while irregularity facilitates better discrimination, our study indicates that the similarity of the nominal and verbal prefixes *PEN-* and *MEN-*, which at higher levels of cognitive processing may offer an advantage for the learning, comes with a disadvantage at the lower level of implicit error-driven learning, resulting in mappings between form and meaning that are less precises for *PEN-* as compared to *PE-*.

In order to more precisely understand the mappings between meaning and form for *PEN-* and *PE-*, we developed a new measure gauging functional load: L_τ. This measure gauges to what extent the similarity between the predicted semantic vector and the targeted semantic vector decreases when a triphone τ is withheld from the model input. We observed that the functional load of the second triphone was lower than that of the first and third triphones. Fur-

thermore, the functional load for the initial triphone was slightly greater for *PEN-*, whereas that of the second triphone was slightly greater for *PE-*. Apparently, under the pressure to discriminate between both words with *PEN-* and *MEN-*, and words with *PEN-* and *PE-*, the initial triphone is used more to discriminate *PEN-* from the other prefixes, whereas the second triphone is used more to discriminate between words with *PE-* and words with the other prefixes.

In the present framework, the role of triphones at the boundary between the prefix and the stem is very different from the role boundary n-phones (typically, diphones) play in theories that assume words are segmented into prefix and stem (Seidenberg, 1987; Hay, 2003; Hay and Baayen, 2003). In these theories, it is assumed that a low-frequency diphone straddling the boundary between prefix and stem facilitates segmentation. However, the reliability of diphones as a boundary cues is questionable (Baayen et al., 2016). Importantly, from a discriminative perspective, n-phones at the juncture of prefix and stem are precisely those cues that potentially have a high functional load, the reason being that they do not occur in many other words and hence can contribute more substantially to discriminating the target word from its competitors. It is worth noting that the functional load of triphones is not proportional to their frequency. In our data, for instance, the initial triphone `#pe` is both frequent and has a high functional load, whereas the second triphone of *PE-*, `peX`, has a much lower frequency and a lower functional load, whereas the subsequent lower-frequency triphone `eXY` has a higher functional load again. In other words, triphone frequency is too crude a measure to capture the details of functional load.

The formalization of functional load proposed in the present study offers a novel way of addressing questions that traditionally are addressed by means of minimal pairs. Wedel et al. (2013), for instance, argues that functional load is a major factor in determining whether two phonemes merged or not. Their study showed that the greater the number of minimal pairs is that is associated with a phoneme, the lower the probability will be that this phoneme will merge with another phoneme. In the same vein, we expect that triphones with a higher functional load will be less likely to merge. At the same time, our operationalization of functional load makes it possible to take more subtle paradigmatic pressures into account, as illustrated for the first and

second triphones of *PEN-* and *PE-*. Due to paradigmatic pressure from *MEN-*, the functional load of the #pe triphone is higher for *PEN-* and lower for *PE-*, whereas the functional load of the second triphone is higher for *PE-* and lower for *PEN-*.

In the literature, studies on the nasal/plosive alternation in Austronesian languages have focused on the initial segment (see, e.g., Ramlan (2009); Sugerman (2016); Sukarno (2017); Kager et al. (1996); Halle and Clements (1983); Blust (2004)), and proposed a rule of nasal substitution for the nominalization. Alternatively, the *MEN-*/*PEN-* alternation can be understood as involving a rule of affix substitution (see for an extended discussion of affix substitution Marle, 1985, 1986). In the present study, which is grounded in Word and Paradigm morphology (Blevins, 2016), phonological and morphological substitution rules are not part of the theoretical toolkit, as the word is taken to be the fundamental smallest unit of analysis. Even though we did not inform our computational model about exponents and stems, the model nevertheless learned a substantial part of Indonesian morphology with a high accuracy (around 93–94%). Model accuracy for *PEN-* and *PE-* was near ceiling (around 96–100%). What our approach offers the analyst over and above what phonological or morphological substitution rules can reveal is further insight into the learnability of the prefixes and the distribution of phones' functional load in the prefix and at the prefix-stem boundary. The finding that *PEN-* is learned less robustly than *PE-*, due to more extensive cue-competition when substitution pairs are phonologically similar, suggests a possible reason for why affix substitution is relatively rare both within languages and across Austronesian languages (Dempwolff, 1934; Blust, 2004).

What sets the present approach apart from computational modeling with Analogical Modeling of Language (AML Skousen, 1989) and from nearest-neighbor approaches such as implemented in the Tilburg Memory-Based Learner (TiMBL Daelemans et al., 2007) is, first, that AML and TiMBL consider similarity at the level of form, abstracting away from semantic similarities, and second, that AML and TiMBL are classifiers. Thus, while AML or TiMBL could be used to predict which allomorph of *PEN-* is appropriate given a set of features describing the phonology of the base word, these models do not straightforwardly predict words' forms themselves. Nevertheless, both AML and TiMBL have proved valuable insight into a

range of phenomena (see, e.g., Krott, 2001; Daelemans and Van den Bosch, 2005; Eddington, 2002; Arndt-Lappe, 2011), and one feature of these models that has proved especially useful is the possibility to inspect the sets of closest neighbors that drive analogical prediction. Within the present discriminative framework, it is also possible to inspect which words are the closest neighbors, both in semantic space (comprehension) and in form space (production). Furthermore, quantitative measures can also be derived from the properties of the production and comprehension networks to predict aspects of lexical processing (see, e.g., Milin et al., 2017; Chuang and Baayen, 2020). In fact, the measure of functional load proposed in the present study may turn out to be predictive for the acoustic duration of phones in spoken Indonesian (cf. Baayen et al., 2019; Tomaschek et al., 2019). We leave exploring this possibility to future research. What we hope to have demonstrated with the present computational modeling study is that discrimination learning provides a useful new quantitative tool for understanding the interaction between form and meaning in morphology.

Acknowledgement

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

Bibliography

Alwi, H., Dardjowidjojo, S., Lapoliwa, H., and A.M., M. (2003). *Tata Bahasa Baku Bahasa Indonesia*. Balai Pustaka, Jakarta, 3rd edition.

Arka, I. W., Dalrymple, M., Mistica, M., and Mofu, S. (2009). A linguistic and computational morphosyntactic analysis for the applicative -i in Indonesian. In Butt, M. and King, T. H., editors, *International Lexical Functional Grammar Conference (LFG)*, pages 85–105. CSLI Publications.

Arndt-Lappe, S. (2011). Towards an exemplar-based model of stress in English noun-noun compounds. *Journal of Linguistics*, pages 549–585.

Baayen, R. H., Chuang, Y.-Y., and Blevins, J. P. (2018). Inflectional morphology with linear mappings. *The Mental Lexicon*, 13(2):232–270.

Baayen, R. H., Chuang, Y.-Y., Shafaei-Bajestan, E., and Blevins, J. P. (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*, pages 1–39.

Baayen, R. H., Shaoul, C., Willits, J., and Ramscar, M. (2016). Comprehension without segmentation: A proof of concept with naive discriminative learning. *Language, Cognition, and Neuroscience*, 31(1):106–128.

Baayen, R. H. and Smolka, E. (2020). Modelling morphological priming in German with naive discriminative learning. *Frontiers in Communication*, Language Sciences:1–40.

Benjamin, G. (2009). Affixes, Austronesian and iconicity in Malay. *Bijdragen tot de Taal-, Land- en Volkenkunde*, 165(2–3):291–323.

Blevins, J. P. (2003). Stems and paradigms. *Language*, 79:737–767.

Blevins, J. P. (2006). Word-based morphology. *Journal of Linguistics*, 42(03):531–573.

Blevins, J. P. (2016). *Word and paradigm morphology*. Oxford University Press, Oxford.

Blevins, J. P., Milin, P., and Ramscar, M. (2017). The zipfian paradigm cell filling problem. *Perspectives on Morphological Organization: Data and Analyses*, 10:141.

Blust, R. (2004). Austronesian nasal substitution: A survey. *Oceanic Linguist*, 43(1):73–148.

Booij, G. E. (1986). Form and meaning in morphology: the case of Dutch àgent nouns. *Linguistics*, 24:503–517.

Chaer, A. (2008). *Morfologi Bahasa Indonesia (Pendekatan Proses)*. PT Rineka Cipta, Jakarta.

Chuang, Y., Vollmer, M.-L., Shafaei-Bajestan, E., Gahl, S., Hendrix, P., and Baayen, R. H. (2020). The processing of nonword form and meaning in production and comprehension: A computational modeling approach using linear discriminative learning. *Behavior Research Methods*.

Chuang, Y. Y. and Baayen, R. H. (2020). Discriminative learning and the lexicon: Ndl and ldl. In *Oxford Research Encyclopedia of Linguistics*. Oxford University Press. (under review).

Chuang, Y.-Y., Loo, K., Blevins, J. P., and Baayen, R. H. (2019). Estonian case inflection made simple. A case study in Word and Paradigm morphology with Linear Discriminative Learning. *PsyArXiv*, pages 1–19.

Daelemans, W. and Van den Bosch, A. (2005). *Memory-based language processing*. Cambridge University Press, Cambridge.

Daelemans, W., Zavrel, J., Van der Sloot, K., and Van den Bosch, A. (2007). TiMBL: Tilburg

Memory Based Learner Reference Guide. Version 6.1. Technical Report ILK 07-07, Computational Linguistics Tilburg University.

Dardjowidjojo, S. (1983). *Some Aspects of Indonesian Linguistics*. Djambatan, Jakarta.

Dempwolff, O. (1934). *Vergleichende Lautlehre des austronesischen Wortschatzes*. Number 19 in Vergleichende lautlehre des austronesischen wortschatzes. D. Reimer.

Denistia, K. and Baayen, H. (2019). The Indonesian prefixes PE- and PEN-: A study in productivity and allomorphy. *Morphology*, 29(3):385–407.

Denistia, K., Shafaei-Bajestan, E., and Baayen, H. (2020). Exploring semantic differences between the Indonesian prefixes PE- and PEN- using a vector space model. *Corpus Linguistics and Linguistic Theory*. Accepted pending minor revision.

Eddington, D. (2002). Spanish diminutive formation without rules or constraints. *Linguistics*, 40:395–419.

Ermanto (2016). *Morfologi Afiksasi Bahasa Indonesia Masa Kini: Tinjauan dari Morfologi Derivasi dan Infleksi*. Kencana, Jakarta.

Halle, M. and Clements, G. N. (1983). *Problem Book in Phonology*. MIT Press, Cambridge.

Hathout, N. and Namer, F. (2019). Paradigms in word formation: what are we up to? *Morphology*, 29:153–165.

Hay, J. B. (2003). *Causes and Consequences of Word Structure*. Routledge, New York and London.

Hay, J. B. and Baayen, R. H. (2003). Phonotactics, parsing and productivity. *Italian Journal of Linguistics*, 1:99–130.

Heitmeier, M. and Baayen, R. H. (2020). Simulating phonological and semantic impairment of English tense inflection with Linear Discriminative Learning. *PsyArXiv*, January 7:1–29.

Kager, R., Hulst, H. v. d., and Zonneveld, W., editors (1996). *Prosody Morphology Interface*, chapter Austronesian Nasal Substitution and other NC effects. Cambridge University Press, Cambridge.

Kridalaksana, H. (2008). *Kamus Linguistik*. PT Gramedia Pustaka Utama, Jakarta, 4th edition.

Kroeger, P. R. (2007). Morphosyntactic vs. morphosemantic functions of Indonesian –kan. In Zaenen, A., Simpson, J., King, T. H., Jane, G., Maling, J., and Manning, C., editors, *Architectures, Rules, and Preferences: Variations on Themes of Joan Bresnan*, number 184 in CSLI Lecture Notes, pages 229–251. CSLI Publications, Stanford, California.

Krott, A. (2001). *Analogy in Morphology: The selection of linking elements in Dutch compounds*. University of Nijmegen, Nijmegen.

Krott, A., Schreuder, R., and Baayen, R. H. (1999). Complex words in complex words. *Linguistics*, 37(5):905–926.

Landauer, T. and Dumais, S. (1997). A solution to plato's problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*, 104(2):211–240.

Lombardi, L., editor (2001). *Segmental phonology in Optimality Theory: Constraints and Representations*, chapter Austronesian Nasal Substitution Revisited, pages 159–182. Cambridge University Press, Cambridge.

Marle, J. v. (1985). *On the Paradigmatic Dimensions of Morphological Creativity*. Foris, Dordrecht.

Marle, J. v. (1986). The domain hypothesis: The study of rival morphological processes. *Linguistics*, 24:601–627.

Martinet, A. (1952). Function, structure, and sound change. *Word*, 8:1–32.

Matthews, P. H. (1974). *Morphology. An Introduction to the Theory of Word Structure*. Cam-

bridge University Press, Cambridge.

Matthews, P. H. (1991). *Morphology*. Cambridge University Press, Cambridge.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Milin, P., Feldman, L. B., Ramscar, M., Hendrix, P., and Baayen, R. H. (2017). Discrimination in lexical decision. *PLOS-one*, 12(2):e0171935.

Nomoto, H. (2006). A study on complex existential sentences in Malay. Master's thesis, Universiti Bahasa Asing Tokyo, Tokyo.

Nomoto, H. (2017). The syntax of Malay nominalization. In Razak, R. A. and Yusoff, R., editors, *Aspek Teori Sintaksis Bahasa Melayu*, pages 71–117. Dewan Bahasa dan Pustaka, Kuala Lumpur.

Oh, Y. M., Coupé, C., Marsico, E., and Pellegrino, F. (2015). Bridging phonological system and lexicon: Insights from a corpus study of functional load. *Journal of Phonetics*, 53:153–176.

Putrayasa, I. B. (2008). *Kajian Morfologi: Bentuk Derivasional dan Infleksional*. PT Refika Aditama, Bandung.

R Team, S. (2015). *RStudio: Integrated Development for R. RStudio*. RStudio, Inc., Boston, MA.

Ramlan, M. (2009). *Morfologi: Suatu Tinjauan Deskriptif*. CV Karyono, Yogyakarta.

Seidenberg, M. (1987). Sublexical structures in visual word recognition: Access units or orthographic redundancy. In Coltheart, M., editor, *Attention and Performance XII*, pages 245–264. Lawrence Erlbaum Associates, Hove.

Skousen, R. (1989). *Analogical Modeling of Language*. Kluwer, Dordrecht.

Sneddon, J. N., Adelaar, A., Djenar, D. N., and Ewing, M. C. (2010). *Indonesian: A Comprehensive Grammar*. Routledge, New York, second edition.

Soekarno, Y. (2010). *Derivational syntax: A minimalist approach to affixation in Bahasa Indonesia predicates*. LAP LAMBERT Academic Publishing, Germany.

Stekauer, P. (2014). Derivational paradigms. In Lieber, R. and Štekauer, P., editors, *The Oxford Handbook of Derivational Morphology*, pages 354–369. Oxford University Press.

Subroto, E. (2012). *Pemerian Morfologi Bahasa Indonesia: Berdasarkan Perspektif Derivasi dan Infleksi Proses Afiksasi*. Yuma Pressino, Surakarta.

Sugerman (2016). *Morfologi Bahasa Indonesia: Kajian ke Arah Linguistik Deskriptif*. Penerbit Ombak, Yogyakarta.

Sukarno (2017). The behaviours of the general nasal /N/ in Indonesian active prefixed verbs. *International Journal of Language and Linguistics*, 4(2):48 – 52.

Sutanto, I. (2002). Verba berkata dasar sama dengan gabungan afiks men-i atau men-kan. *Makara, Sosial-Humaniora*, 6(2):82–87.

Tomaschek, F., Plag, I., Ernestus, M., and Baayen, R. H. (2019). Modeling the duration of word-final s in English with naive discriminative learning. *Journal of Linguistics*. https://psyarxiv.com/4bmwg, doi = 10.31234/osf.io/4bmwg.

Tomasowa, F. H. (2007). The reflective experiential aspect of meaning of the affix -i in Indonesian. *Linguistik Indonesia*, 25(2):83–96.

van Marle, J. (1984). *On the paradigmatic dimension of morphological creativity*. De Gruyter, Berlin, Boston.

Wedel, A., Kaplan, A., and Jackson, S. (2013). High functional load inhibits phonological contrast loss: A corpus study. *Cognition*, 128:179–186.

Chapter 6

Summary and conclusions

Research on Indonesian morphology has thus far been qualitative in nature, with researchers' knowledge of the language informing grammatical description and analysis. An important contribution of my research to Indonesian linguistics is that I have been able to show the advantages of complementing qualitative research with corpus-based approach. Corpus linguistics has a long history in English linguistics (O'Keeffe and McCarthy, 2010; Biber and Randi, 2015), but to my knowledge, Chapters 3 and 4 are among the very first corpus-based studies of aspects of the Indonesian lexicon that are based on a large corpus of Indonesian[14]. For my research, I have been able to profit from the Leipzig Corpora Collection (`http://corpora2.informatik.uni-leipzig.de/download.html`), which includes a large Indonesian corpus (Goldhahn et al., 2012). Thanks to the availability of the `MorphInd` parser (Larasati et al., 2011), all words in the corpus could be parsed. This made it possible for constructing a database of all words in the corpus with the prefixes that I have investigated, *PE-* and *PEN-*. This database is available at `http://bit.ly/PePeNProductivity`,`http://bit.ly/PePeNSemVector`,`https://bit.ly/PePeNwithLDL`.

By systematically investigating how *PE-* and *PEN-* are used in written Indonesian, some new usages of these prefixes, that thus far had escaped notice in the qualitative literature, emerged. For instance, *PE-* does not only express the semantic roles of agent and patient, but

[14]Rajeg (2019) is the other study in theoretical linguistics that is corpus-based.

it also can form nouns denoting instruments (e.g., *kasih* 'love'-*pekasih* 'love potion'). Furthermore, *PEN-* does not only realize agents and instruments, but also creates nouns specifying the causer (e.g., *sakit* 'ill'-*penyakit* 'disease') and location (e.g. *hujung* 'the end'-*penghujung* 'the very end'). I have subsequently also used the corpus to study some negation formations in Indonesian (Rajeg et al., 2018), and was able to show that *tidak* 'not' is commonly used for verbal negation, whereas *tak* 'not' occurs with verbs with the circumfix *ter-/-kan* (e.g., *tak terhindarkan* 'unavoidable'). As this study is not centered on *PE-* and *PEN-*, it is not included in this doctoral dissertation.

Thanks to the resources developed for Indonesian by computational linguists, it has been possible to study the productivity of *PE-* and *PEN-* using the quantitative measures developed by Baayen (1993). One of the most remarkable findings is that the productivity of the allomorphs of PEN- is strongly correlated with the productivity of the allomorphs of *MEN-*. I was able to show that *PE-* and its base verbs with *BER-* are outliers that do not partake in the paradigmatic proportionality that characterizes *PEN-* and *MEN-*. Furthermore, the frequencies with which *BER-* verbs are used do not correlate with the frequencies with which the corresponding *PE-* nominalizations are used. This is a surprising result given that the literature describes *PE-* as standing in the same kind of paradigmatic relation to *BER-* that is present for *PEN-* and *MEN-* (Chaer, 2008; Ramlan, 2009; Ermanto, 2016; Sneddon et al., 2010; Putrayasa, 2008; Dardjowidjojo, 1983; Benjamin, 2009).

Another methodology that I have found useful for understanding how *PE-* and *PEN-* are used is the toolkit of distributional semantics. With the `Word2Vec` software (Mikolov et al., 2013), I was able to derive so-called word embeddings from the Leipzig corpus, i.e., vectors of real numbers representing words' meanings. Within the set of words with *PEN-*, the words expressing instruments were semantically cohesive, whereas within the set of words with *PE-*, those words denoting athletes or professional players showed strong semantic similarity, in line with Chaer (2008)'s observation that *PE-* is specialized for this semantic domain. Distributional semantics has also proved useful in one other study on Indonesian morphology (Rajeg et al., 2019) that, because it does not concern *PE-* and *PEN-*, is not part of this dissertation.

The corpus-based surveys and the quantitative analyses of productivity and semantic similarity that I carried out for *PE-* and *PEN-* clarified that *PE-* is not simply an allomorph of *PEN-*, but rather an independent prefix with its own specific semantic and quantitative properties. A final question that I addressed using computational modeling is whether the similarity of words with *PEN-* to their base verbs with *MEN-*, which share the same nasal allomorphy, makes *PEN-* easier to learn than *PE-*. Nasal substitution ([m] in *MEN-* being substituted by a [p] in *PEN-*) is found also in other Austronesian languages, see, e.g., (Dempwolff, 1934), for [m] / [p] substitution. It is conceivable that these paradigmatic substitution patterns provide systematicity in the grammar that facilitates learning.

To clarify whether indeed *PEN-* is easier to learn than *PE-*, I made use of linear discriminative learning (Baayen et al., 2019). A model was created for a set of 2500 words, constructed by taking 99 simple words and adding all inflected and derived forms of these words that occur in the Leipzig corpus. The model learned to understand and produce these words with high accuracy (94%). I then subjected the words with *PEN-* and *PE-* to further analyses, and observed that in comprehension, the model was able to predict the semantic vectors of words with *PE-* more precisely than for *PEN-*. This turned out to be a direct consequence of cue competition: phones (or more precisely, triphones) that are shared by *PEN-* and *MEN-* are in competition for nominal versus verbal semantics. Although the form similarity of *PEN-* and *MEN-* may be useful for learning at higher levels of cognitive processing, it turns out to lead to somewhat decreased precision in subliminal implicit learning. A new measure for the functional load of (tri)phones made it possible to inspect in more detail how the discriminative burden is spread out over the (tri)phones of *PEN-*, *MEN-*, and *PE-*. In future work, it will be interesting to replace the simulated semantic vectors simulated by the model with corpus-derived semantic vectors.

I hope to have shown that corpus linguistics, distributional semantics, and computational modeling offer exciting new possibilities for understanding the morphology of Indonesian.

Zusammenfassung und fazit

Die indonesische Morphologie ist bisher qualüitativ erforscht worden, wobei das Wissen der Forscher selbst über die Sprache für die grammatische Beschreibung und Analyse bestimmte. Ein wichtiger Beitrag meiner Forschung zur indonesischen Linguistik ist, dass ich zeigen konnte, welche Vorteile es bringt, qualitative Forschung durch korpusbasierte Untersuchungen zu ergänzen. Die Korpuslinguistik hat eine lange Geschichte in der englischen Sprachforschung (O'Keeffe and McCarthy, 2010; Biber and Randi, 2015), jedoch zählen die Kapitel 3 und 4 meines Wissens nach zu den allerersten korpusbasierten Studien zu Aspekten des indonesischen Lexikons, die auf einem großen Korpus der indonesischen Sprache basierten[15]. In meiner Arbeit nutzte ich die Leipziger Korpus-Sammlung (`http://corpora2.informatik.uni-leipzig.de/download.html`), die zahlreiche indonesischen Korpora umfasst (Goldhahn et al., 2012). Mithilfe des `MorphInd`-Parsers (Larasati et al., 2011) konnten alle Wörter im Korpus auf ihre morphologische Funktion hin analysiert werden. Dies ermöglichte auch den Aufbau einer Datenbanken von Wörtern mit den von mir untersuchte Präfixen *PE-* und *PEN-* im Korpus. Diese Datenbanken sind verfügbar unter `http://bit.ly/PePeNProductivity`, `http://bit.ly/PePeNSemVector`,`https://bit.ly/PePeNwithLDL`.

Durch eine systematische Untersuchung der Verwendung von *PE-* und *PEN-* in indonesischen Texten sind neue Verwendungsformen dieser Präfixe, die in der qualitativen Literatur bisher unbemerkt waren, zum Vorschein gekommen. Beispielsweise bezeichnet *PE-* nicht nur die semantischen Rollen von Agenten und Patient, sondern es kann auch Substantive bilden, die Instrumente bezeichnen (z.B. *kasih* "Liebe"-*pekasih* "Liebestrank"). Weit-

[15]Rajeg (2019) ist die andere Studie in der theoretischen Linguistik, die korpusbasiert ist.

erhin markiert *PEN-* nicht nur Agenten und Instrumente, sondern auch Substantive, die den Verursacher (z.B. *sakit* 'krank'-*penyakit* 'Krankheit') und den Ort (z.B. *hujung* 'das Ende'-*penghujung* "das allerletzte Ende') spezifizieren. Darüberhinaus verwendete ich das Korpus zur Untersuchung einiger Negationsformen auf Indonesisch (Rajeg et al., 2018). Es hat sich gezeigt, dass *tidak* 'nicht' häufig für verbale Negation verwendet wird, während *tak* 'nicht' bei Verben mit dem Zirkumfix *ter-/-kan* vorkommt (z.B. *tak terhindarkan* 'unvermeidlich'). Da sich diese Studie nicht auf *PE-* und *PEN-* konzentriert, wird sie in dieser Dissertation nicht berücksichtigt.

Dank der von Computerlinguisten für Indonesisch entwickelten Ressourcen war es mir möglich, die Produktivität von *PE-* und *PEN-* mittels der von Baayen (1993) entwickelten quantitativen Methoden zu untersuchen. Ein wichtiges Ergebnis dieser Untersuchung ist, dass die Produktivität der Allomorphe von *PEN-* stark mit der Produktivität der Allomorphe von *MEN-* verbunden ist. Es zeigte sich, dass *PE-* und sein Verbstamm mit *BER-* Ausreißer sind. Das bedeutet, dass die Ausreißer nicht zu der paradigmatischen Proportionalität, die für *PEN-* und *MEN-* bezeichnend ist, gehören. Weiterhin korreliert die Frequenz der Verwendung *BER-*Verben nicht mit der Frequenz der *PE-* Nominalisierung. Dies ist ein überraschendes Ergebnis, da die Literatur *PE-* in der gleichen paradigmatischen Beziehung wie *BER-* beschreibt, wie sie für *PEN-* und *MEN-* vorliegt (Chaer, 2008; Ramlan, 2009; Ermanto, 2016; Sneddon et al., 2010; Putrayasa, 2008; Dardjowidjojo, 1983; Benjamin, 2009).

Eine andere Methodologie, die ich geeignet fand, um die Verwendung von *PE-* und *PEN-*, ist das Toolkit der Verteilungssemantik. Mit der Software `Word2Vec` (Mikolov et al., 2013) konnte ich aus dem Leipziger Korpus sogenannte Worteinbettungen ableiten, das heißt Vektoren von realen Zahlen, die die Bedeutungen von Wörtern darstellen. In der *PEN-* Wortgruppe waren die verwendeten Wörter, die Instrumente ausdrücken, semantisch kohärent, während die Wörter in der *PE-* Wortgruppe, welche Athleten oder Berufsspieler bezeichnen, starke semantische Ähnlichkeit zeigten. Dies stimmt mit der Beobachtung von Chaer (2008) überein, dass *PE-* auf solche semantischen Aspekte spezialisiert ist. Die Verteilungssemantik hat sich auch in einer anderen Studie über die indonesische Morphologie von Rajeg et al. (2019) als

nützlich gezeigt. Diese Studie ist ebenfalls nicht Teil dieser Dissertation, da sie sich nicht auf *PE-* und *PEN-* bezieht.

Die korpusbasierten Untersuchungen und die quantitativen Analysen der Produktivität sowie die semantische Ähnlichkeit, die ich für *PE-* und *PEN-* durchführte, zeigten, dass *PE-* nicht nur ein Allomorph von *PEN-* ist. Im Gegenteil. *PE-* ist ein unabhängiges Präfix mit seinen eigenen spezifischen semantischen und quantitativen Eigenschaften. Eine letzte Frage, die ich mithilfe der Computermodellierung untersuchte, ist, ob die Ähnlichkeit der Wörter mit *PEN-* und ihren Basisverben mit *MEN-*, die die gleiche nasale Allomorphie aufweisen, dazu führt, dass *PEN-* einfacher als *PE-* zu lernen ist. Nasale Substitution ([m] in *MEN-* wird durch ein [p] in *PEN-* ersetzt) findet sich auch in anderen austronesischen Sprachen, siehe z.B. (Dempwolff, 1934) zur [m] / [p] - Substitution. Es ist vorstellbar, dass diese paradigmatischen Ersatzmuster für eine Systematik in der Grammatik vorliegt, die das Lernen erleichtert.

Um zu untersuchen, ob *PEN-* überhaupt leichter zu lernen als *PE-* ist, verwendete ich das *linear discriminative learning* (Baayen et al., 2019)). Es wurde ein Modell von 2500 Wörtern erstellt. Dieses Modell wurde aufgebaut, wobei 99 einfache Wörter genommen und alle flektierten und abgeleiteten Formen dieser Wörter, die im Leipziger Korpus vorkommen, hinzugefügt wurden. Das Modell lernte diese Wörter mit hoher Genauigkeit (94%) zu verstehen und zu produzieren. Danach analysierte ich die Wörter mit *PEN-* und *PE-* tiefergehend und stellte fest, dass das Modell beim Verstehen die semantischen Vektoren von Wörtern mit *PE-* exakter vorhersagen konnte als bei *PEN-*. Es zeigte sich, dass dies eine direkte Folge der *Cue-Competition* ist: Phone (oder genauer gesagt, Triphone), die von *PEN-* und *MEN-* gemeinsam genutzt werden, konkurrieren für nominale versus verbale Semantik. Obwohl die Ähnlichkeitsform von *PEN-* und *MEN-* für das Lernen auf höheren kognitiven Verarbeitungsebenen nützlich sein kann, führte sie doch zu einer etwas geringeren Genauigkeit beim subliminalen impliziten Lernen. Ein neues Verfahren für die Funktionsbelastung von Triphonen ermöglichte es, genauer zu untersuchen, wie sich die diskriminierende Belastung auf die Triphone von *PEN-*, *MEN-* und *PE-* verteilt. Für zukünftige Studie wäre es von Interesse, die durch das Modell simulierten semantischen Vektoren mit welchen von einem Korpus abgeleiteten semantischen Vektoren zu ersetzen.

Durch diese Arbeit konnte ich hoffentlich zeigen, dass Korpuslinguistik, Verteilungssemantik und Computermodellierung neue spannende Möglichkeiten zum Verständnis der indonesischen Morphologie bieten können.

Ringkasan dan kesimpulan

Sejauh ini, penelitian dalam bidang morfologi Bahasa Indonesia masih menggunakan pendekatan kualitatif, yang mengandalkan kemampuan dan intuisi dari penutur asli, untuk menganalisis dan menjelaskan fenomena kebahasaan. Kontribusi yang signifikan dari penelitian saya terhadap linguistik Indonesia adalah saya mampu menunjukkan bahwa penelitian kuantitatif berbasis korpus dapat menjadi pelengkap bagi penelitian kualitatif yang sudah ada. Linguistik korpus telah memiliki sejarah panjang dalam linguistik Bahasa Inggris (O'Keeffe and McCarthy, 2010; Biber and Randi, 2015). Akan tetapi, sejauh yang saya ketahui, Bab 3 dan 4 dapat dikategorikan sebagai pionir penelitian korpus[16] untuk meneliti leksikon bahasa Indonesia dengan menggunakan bank data yang terdiri dari jutaan kalimat. Leipzig Corpora Collection (`http://corpora2.informatik.uni-leipzig.de/download.html`, Goldhahn et al. (2012)) dan `MorphInd` (Larasati et al., 2011) sangat membantu penelitian saya. Ketersediaan data yang melimpah dan perangkat lunak untuk memecah imbuhan dan kata dasar ini membantu saya untuk membuat bank data dari semua kata berimbuhan *PE-* dan *PEN-* yang terdapat di dalam korpus. Bank data ini tersedia dan dapat diunduh secara gratis melalui link `http://bit.ly/PePeNProductivity`, `http://bit.ly/PePeNSemVector`, `https://bit.ly/PePeNwithLDL`.

Melalui metode penelitian yang sistematis mengenai penggunaan *PE-* dan *PEN-* dalam bahasa Indonesia, saya menemukan beberapa makna baru yang ada pada kedua imbuhan ini. *PE-* tidak hanya mengungkapkan peran semantik sebagai agen dan penderita, melainkan juga berfungsi sebagai pembentuk instrumen (contoh, *kasih - pekasih*). Selain itu, *PEN-* bukan

[16] Ada penelitian yang sebelumnya dilakukan oleh Rajeg (2019), yang juga merupakan pionir penelitian korpus

hanya berfungsi sebagai pembentuk agen dan instrumen, tetapi juga dapat mengekspresikan penyebab (contoh, *sakit - penyakit*) dan lokasi (contoh, *hujung - penghujung*). Makna semantik yang saya temukan ini belum pernah dibahas dalam literatur kualitatif. Selain temuan yang saya paparkan dalam disertasi ini, saya juga melakukan penelitian lain menggunakan data korpus untuk mempelajari bentuk negasi dalam bahasa Indonesia (Rajeg et al., 2018). Dari penelitian ini, kami menemukan bahwa kata 'tidak' lazim digunakan dalam negasi verba, sedangkan kata 'tak' memiliki kecenderungan kolokasi dengan kata berimbuhan *ter-/-kan* (contoh, tak terhindarkan).

Berkat perkembangan ilmu linguistik komputasional dan metode kuantitatif, dua imbuhan *PE-* dan *PEN-* dapat dianalisis dari segi produktifitasnya. Analisis produktifitas ini dikembangkan oleh Baayen (1993). Salah satu temuan penting yang ada dalam disertasi ini adalah bahwa produktivitas alomorf *PEN-* sangat berkorelasi dengan produktifitas alomorf *MEN-*. Akan tetapi, frekuensi verba berawalan *BER-* tidak berkorelasi dengan nominalisasi yang tercipta dari imbuhan *PE-*. Temuan ini cukup mengejutkan sebab teori kualitatif menyebutkan bahwa awalan *PE-* berkorelasi dengan awalan *BER-*; sama halnya seperti awalan *PEN-* yang berkorelasi dengan awalan *MEN-* (Chaer, 2008; Ramlan, 2009; Ermanto, 2016; Sneddon et al., 2010; Putrayasa, 2008; Dardjowidjojo, 1983; Benjamin, 2009).

Metode lain yang menurut saya berguna untuk memahami bagaimana *PE-* dan *PEN-* digunakan dalam konteks kalimat adalah metode distribusi semantik. Menggunakan perangkat lunak `Word2Vec` (Mikolov et al., 2013), saya memperoleh distribusi vektor dari setiap penggunaan kata yang berawalan *PE-* dan *PEN-*. Konteks penggunaan kata-kata tersebut diambil dari kalimat yang ada di korpus leipzig. Setelah `Word2Vec` diaplikasikan ke dalam data, perhitungan statistik semantik membuktikan bahwa makna instrumen untuk kata-kata berawalan *PEN-* lebih transparan dibanding makna *PEN-* sebagai agen. Di sisi lain, himpunan kata-kata berawalan *PE-* yang menunjukkan makna semantik atlet menunjukkan kemiripan semantik yang lebih tinggi dibandingkan dengan makna *PE-* sebagai pembentuk agen secara umum. Hal ini sejalan dengan pengamatan Chaer (2008) yang menyatakan bahwa awalan *PE-* memang dikhususkan sebagai pembentuk atlet. Selain itu, metode distribusi semantik juga terbukti berguna dalam studi lain, Rajeg et al. (2019), yang tidak terkait dengan awalan *PE-* dan *PEN-*.

Penelitian berbasis korpus dan analisis kuantitatif dari segi produktifitas serta kemiripan semantik yang saya lakukan untuk imbuhan *PE-* dan *PEN-* membuktikan bahwa *PE-* bukan merupakan alomorf *PEN-*. *PE-* adalah imbuhan yang memiliki karakter semantik dan karakter kuantitasnya sendiri.

Dalam pembahasan terakhir di disertasi ini, saya menggunakan pemodelan komputasi (*computational modelling*) untuk menjelaskan apakah kesamaan fonologi yang terdapat pada kata-kata berawalan *PEN-* dan *MEN-* membuat *PEN-* menjadi lebih mudah dipelajari daripada *PE-*. Pada dasarnya, saya melatih komputer untuk dapat membedakan imbuhan *PE-* dan *PEN-*. Subsitusi awalan yang terjadi antara nasal [m] pada *MEN-* dan [p] pada *PEN-* juga ditemukan dalam rumpun Bahasa Austronesia lainnya (baca Dempwolff (1934) untuk substitusi [m] dan [p]). Bagaimana dampak substitusi afiks dalam pemodelan komputasi? Bisa jadi, pola substitusi yang bersifat paradigmatik dan memberikan sebuah regularitas dalam tata bahasa justru memfasilitasi proses pembelajaran bagi pemodelan komputasi.

Untuk membuktikan hipotesa yang telah dijelaskan di paragraf sebelum ini, saya menggunakan pemodelan komputasi *linear discriminative learning* (Baayen et al., 2019). Pemodelan komputasi ini melibatkan lebih dari 2500 kata yang diturunkan dari 99 kata dasar. Kata-kata yang ada dalam penelitian ini bisa merupakan kata yang mengandung afiks derivasi, afiks infleksi, maupun kata monomorfemik. Seluruh kata yang diujikan terdapat dalam korpus Leipzig. Perhitungan kuantitatif dari proses pembelajaran komputer untuk awalan *PE-* dan *PEN-* menunjukkan bahwa model dapat mempelajari *PE-* dan *PEN-* dengan level akurasi 94%. Kemudian, saya memfokuskan penelitian pada perbandingan pembelajaran untuk kedua awalan ini. Data menunjukkan bahwa model mampu memprediksi vektor semantik kata-kata berawalan *PE-* dan *PEN-*. Namun demikian, prediksi vektor semantik untuk kata berawalan *PE-* lebih tepat dibanding prediksi untuk kata berawalan *PEN-*. Hal ini merupakan konsekuensi dari kemiripan fonologi antara *PEN-* dan *MEN-*. Lebih tepatnya, eksistensi kemiripan *triphone* yang sama-sama digunakan oleh kata benda berawalan *PEN-* dan kata kerja berawalan *MEN-* mengakibatkan terjadinya kompetisi semantik dalam proses pembelajaran nomina dan verba yang terjadi secara bersamaan. Meskipun awalnya ada asumsi bahwa kemiripan bentuk antara *PEN-* dan *MEN-* membawa keuntungan bagi proses pembelajaran pada level kognitif, kemiri-

pan ini ternyata menyebabkan penurunan presisi dalam pembelajaran kata. Selanjutnya, saya menganalisis lebih lanjut dengan menggunakan perhitungan *functional load* untuk memeriksa lebih secara lebih detail bagaimana *functional load* pada awalan *PE-, PEN-,* dan *MEN-* tersebar dalam distribusi yang berbeda. Untuk penelitian lanjutan mengenai hal ini, saya akan mensimulasikan data yang ada ke dalam vektor semantik yang sesungguhnya, yang ada di dalam korpus.

Harapan saya, disertasi ini menawarkan perspektif yang baru dan menarik terkait dengan kegunaan linguistik korpus, semantik distribusi, dan pemodelan komputasi untuk memahami morfologi bahasa Indonesia.

Bibliography

Baayen, R. H. (1993). On frequency, transparency, and productivity. In Booij, G. E. and van Marle, J., editors, *Yearbook of Morphology 1992*, pages 181–208. Kluwer Academic Publishers, Dordrecht.

Baayen, R. H., Chuang, Y.-Y., Shafaei-Bajestan, E., and Blevins, J. P. (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*, pages 1–39.

Benjamin, G. (2009). Affixes, Austronesian and iconicity in Malay. *Bijdragen tot de Taal-, Land- en Volkenkunde*, 165(2–3):291–323.

Biber, D. and Randi, R. (2015). *The Cambridge Handbook of English Corpus Linguistics*. Cambridge Handbooks in Language and Linguistics. Cambridge University Press, Cambridge.

Chaer, A. (2008). *Morfologi Bahasa Indonesia (Pendekatan Proses)*. PT Rineka Cipta, Jakarta.

Dardjowidjojo, S. (1983). *Some Aspects of Indonesian Linguistics*. Djambatan, Jakarta.

Dempwolff, O. (1934). *Vergleichende Lautlehre des austronesischen Wortschatzes*. Number 19 in Vergleichende lautlehre des austronesischen wortschatzes. D. Reimer.

Ermanto (2016). *Morfologi Afiksasi Bahasa Indonesia Masa Kini: Tinjauan dari Morfologi Derivasi dan Infleksi*. Kencana, Jakarta.

Goldhahn, D., Eckart, T., and Quasthoff, U. (2012). Building large monolingual dictionaries at the Leipzig Corpora Collection: From 100 to 200 languages. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation*, pages 1799–1802.

Larasati, S., Kuboň, V., and Zeman, D. (2011). Indonesian morphology tool MorphInd: Towards an Indonesian corpus. In C., M. and M., P., editors, *Systems and Frameworks for Computational Morphology*, volume 100, pages 119–129. Springer.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

O'Keeffe, A. and McCarthy, M. (2010). *The Routledge Handbook of Corpus Linguistics*. Routledge.

Putrayasa, I. B. (2008). *Kajian Morfologi: Bentuk Derivasional dan Infleksional*. PT Refika Aditama, Bandung.

Rajeg, G. P. W. (2019). *Metaphorical profiles and near-synonyms: A corpus-based study of Indonesian words for happiness*. PhD thesis, Monash University, Australia, Clayton, VIC.

Rajeg, G. P. W., Denistia, K., and Musgrave, S. (2019). Vector space models and the usage patterns of Indonesian denominal verbs: A case study of verbs with men-, men-/-kan, and men-/-i affixes. *NUSA: Linguistic studies of languages in and around Indonesia*, 67(1):35–76.

Rajeg, G. P. W., Denistia, K., and Rajeg, I. M. (2018). Working with a linguistic corpus using R: An introductory note with Indonesian negating construction. *Linguistik Indonesia*, 36(1):1–36.

Ramlan, M. (2009). *Morfologi: Suatu Tinjauan Deskriptif*. CV Karyono, Yogyakarta.

Sneddon, J. N., Adelaar, A., Djenar, D. N., and Ewing, M. C. (2010). *Indonesian: A Comprehensive Grammar*. Routledge, New York, second edition.