Affix substitution in Indonesian: A computational modeling approach

Abstract: Indonesian has two noun-forming prefixes, \textit{PE-} and \textit{PEN-}, that often stand in a paradigmatic relation to verbal base words with the prefixes \textit{BER-} and \textit{MEN-}. The central question addressed in the present study is whether the form similarities between \textit{PEN-} and \textit{MEN-} make \textit{PEN-} easier to learn compared to \textit{PE-}. To address this question, we made use of a computational model, the ‘discriminative lexicon’ (DL) model. We trained this model on 2,517 word forms that were inflected or derived variants of 99 different base words. Of these word forms, 109 were nouns with \textit{PE-} and 221 words were nouns with \textit{PEN-}. Both the production and the comprehension networks of the model performed with high accuracy for both prefixes. However, the model was able to provide more precise predictions for \textit{PE-} as compared to \textit{PEN-}, implying that \textit{PE-} should have a processing advantage compared to \textit{PEN-}. There are two reasons for why \textit{PE-} is learned more robustly than \textit{PEN-}. First, \textit{PE-} words tend to be longer and hence have more discriminative triphones. Second, due to cue competition with \textit{MEN-}, the prefixal triphones of \textit{PEN-} are less effective cues than those of \textit{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, role of junctural phonotactics, and (dis)functionality of affix substitution.

Keywords: affix substitution; computational modeling; junctural phonotactics; linear discriminative learning; paradigmatic relations

1 Introduction

In Indonesian, there are two nominalizing prefixes: \textit{PE-} and \textit{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 indistinguishable in form from the second prefix investigated in this study, *PE*- (Denistia 2018). Qualitative studies (Ramlan 2009; Sneddon et al. 2010) argue that *PE*- and *PEN*- are independent prefixes. On the other hand, Dardjowidjojo (1983) and Kridalaksana (2007) take them to be allomorphs.

Many nouns with *PEN*- are derived by affix substitution1 from verbs with a prefix *MEN*- that is characterized by a similar set of allomorphs as *PEN*- (Benjamin 2009; Dardjowidjojo 1983; Ermanno 2016; Nomoto 2006, 2017; Putrayasa 2008; Ramlan 2009; Sneddon et al. 2010). For example, the word *penari* ‘dancer’ corresponds to the verb *menari* ‘to dance’; these two derivations have *tari* ‘dance’ as the base word. A recent corpus study (Denistia and Baayen 2019) revealed 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 (nomen agentis) and the prefix *pa*- is the question marker for instruments (nomen instrumenti) (Dempwolff 1934). Affix pairs that differ with respect to the initial consonant (stop versus corresponding nasal) are widespread in Austronesian languages (Blust 2004; Halle and Clements 1983; Pater 1999, 2001). 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,

---

1 In what follows, we use the term ‘affix substitution’ as a descriptive term, for theoretical discussion of affix substitution, see, e.g., van Marle (2016 [1984]).
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 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 (Oh et al. 2015; Wedel et al. 2013). 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, in which the word is the smallest unit of analysis (Baayen et al. 2018; Blevins 2003, 2006, 2016; Chuang et al. 2020a; Matthews 1974, 1991).

Given the substantial prevalence of affix substitution in Indonesian morphology (see, e.g., Table 1), and the general importance of paradigmatic relations for the theory of morphology (for the more general importance of paradigmatic relations, see also Hathout and Namer 2019; van Marle 2016 [1984]; Štekauer 2014), the present study addresses the question of whether LDL, a computational theory of morphology that does not have units for stems or exponents, is useful as a tool for understanding the Indonesian lexicon (for overview of Indonesian morphology, see Denistia and Baayen 2022).

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.

2 Linear discriminative learning

Linear discriminative learning provides a computational framework for setting up mappings between numeric vectors representing words’ forms and numeric
Table 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* - .

<table>
<thead>
<tr>
<th>Noun</th>
<th>English noun</th>
<th>Verb</th>
<th>English verb</th>
<th>Noun</th>
<th>English noun</th>
<th>Verb</th>
<th>English verb</th>
</tr>
</thead>
<tbody>
<tr>
<td>pencinta</td>
<td>who is very enthusiastic about something</td>
<td>mencinta</td>
<td>to love</td>
<td>pecinta</td>
<td>lover</td>
<td>bercinta</td>
<td>to make love</td>
</tr>
<tr>
<td>peninju</td>
<td>who punches</td>
<td>meninju</td>
<td>to punch</td>
<td>petinju</td>
<td>boxer</td>
<td>bertinju</td>
<td>to do boxing</td>
</tr>
<tr>
<td>pengecek</td>
<td>checker</td>
<td>mengecek</td>
<td>to check</td>
<td>petani</td>
<td>rice farmer</td>
<td>bertani</td>
<td>to do rice farming</td>
</tr>
<tr>
<td>pelukis</td>
<td>painter</td>
<td>melukis</td>
<td>to paint</td>
<td>pelari</td>
<td>runner</td>
<td>berlari</td>
<td>to run</td>
</tr>
<tr>
<td>pengajar</td>
<td>teacher</td>
<td>mengajar</td>
<td>to teach</td>
<td>pekasih</td>
<td>love potion</td>
<td>kasih</td>
<td>love</td>
</tr>
<tr>
<td>penyumbang</td>
<td>donator</td>
<td>menyumbang</td>
<td>to donate</td>
<td>pesuruh</td>
<td>who is commanded</td>
<td>suruh</td>
<td>order</td>
</tr>
<tr>
<td>pembaca</td>
<td>reader</td>
<td>membaca</td>
<td>to read</td>
<td>pegolf</td>
<td>golf player</td>
<td>golf</td>
<td></td>
</tr>
</tbody>
</table>
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) and German (Baayen and Smolka 2020). It has also been successfully used to study the lexical processing of auditory nonwords (Chuang et al. 2020b) and to model a double dissociation in aphasia (Heitmeier and Baayen 2020). A study addressing the productivity of LDL networks is (Chuang et al. 2020a), which addresses inflection for case and number in Estonian.

We will use the toy lexicon in Table 2 to illustrate how LDL works. 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 *ajaran* /əˈjarən/ ‘lesson’, we obtain the triphones #aj, aja, jar, ar#. Here, the # symbol denotes a word boundary. Equation (1) shows the form matrix $C$,

\[
\begin{bmatrix}
#pe & pet & eta & tan & ani & ni# & pen & eng & nga & gaj & aja & jar & ar# & #aj \\
1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1
\end{bmatrix}
\]

for the lexicon shown in Table 2. The $i$-th row of $C$ 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 Baayen et al. (2018); Chuang et al. (2020a); Chuang et al. (2020b), we make use of simulated semantic vectors. For studies using vectors derived from corpora, see Baayen et al. (2019), and among the very first exploration

<table>
<thead>
<tr>
<th>Lexeme</th>
<th>Word</th>
<th>Phonetic Transcription</th>
<th>Animacy</th>
<th>Concreteness</th>
<th>SemanticRole</th>
</tr>
</thead>
<tbody>
<tr>
<td>ajar</td>
<td>ajar</td>
<td>ḍar</td>
<td>inanimate</td>
<td>abstract</td>
<td></td>
</tr>
<tr>
<td>ajar</td>
<td>pengajar</td>
<td>pəŋɡəjar</td>
<td>animate</td>
<td>concrete</td>
<td>agent</td>
</tr>
<tr>
<td>tani</td>
<td>petani</td>
<td>petani</td>
<td>animate</td>
<td>concrete</td>
<td>agent</td>
</tr>
</tbody>
</table>
for Indonesian, see Denistia et al. (2022); Rajeg et al. (2019). For the present toy example, the dimension of the semantics vector is 14. These vectors are constructed as follows. First, every elementary semantic feature in Table 2, henceforth referred to as lexomes, is coupled with a vector of random numbers that follow a normal distribution. For the lexomes in Table 2, these randomly generated vectors can look like those in matrix $A$.

\[
A = \begin{bmatrix}
\text{animate} & -2.548 & -0.427 & -0.421 & -0.729 & -2.106 & 1.993 & 0.286 & 1.101 & 1.531 & -1.125 & -0.692 & 1.388 & -1.598 & 0.203 \\
\text{abstract} & 0.511 & 0.297 & 0.186 & 0.308 & 0.400 & 1.302 & -0.525 & 2.306 & 2.557 & -0.569 & 0.224 & -0.999 & -1.144 & -0.678 \\
\text{agent} & 1.628 & 0.688 & 0.006 & 0.030 & 1.523 & 1.181 & 0.360 & 0.957 & -1.240 & -1.043 & 1.117 & 2.229 & 0.624 & 1.429 \\
\text{animate} & 2.038 & 1.124 & 1.564 & 1.173 & 1.865 & 1.508 & 0.892 & 0.248 & 1.526 & 1.655 & 1.963 & 0.672 & 2.146 & 0.301 \\
\text{ajar} & 1.511 & 2.015 & 0.411 & 1.131 & 1.304 & 0.577 & 2.242 & -0.218 & -0.022 & 1.178 & 0.557 & 2.379 & 2.784 & 0.144 \\
\text{ajar} & -0.486 & 0.123 & -2.523 & -0.876 & 0.248 & -3.041 & -2.960 & 1.025 & -0.777 & -0.389 & 0.553 & -1.853 & -1.281 & -0.557
\end{bmatrix}
\]

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

\[
S = \begin{bmatrix}
\text{abstract} & 4.671 & 1.127 & -1.194 & -0.114 & 0.408 & 1.762 & -2.209 & 4.680 & 4.835 & -0.457 & 2.057 & -0.792 & -1.877 & 0.099 \\
\text{ajar} & 2.274 & 1.403 & -1.691 & 0.739 & 3.085 & -0.851 & -0.904 & 4.023 & -0.542 & 0.297 & 1.498 & 2.003 & 2.859 & 2.719
\end{bmatrix}
\]

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

\[
CF = S.
\]

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

\[
SG = C.
\]

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

The matrices $F$ and $G$ 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) $F$ are approximate. Following notational conventions in statistics, we denote the predicted, and necessarily approximate, semantic vectors by $\hat{\mathbf{s}}$:

$$\mathbf{CF} = \mathbf{\hat{s}}$$

(6)

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

$$\mathbf{SG} = \mathbf{\hat{C}}$$

(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 $\mathbf{S}$ (see Figure 1). This idea is formalized by constructing the correlation matrix $R_s$ that specifies for each row vector of the predicted semantic matrix $\mathbf{\hat{S}}$ how well it correlates with the semantic vectors of $\mathbf{S}$. The word the semantic vector $\mathbf{s}$ of which has the highest correlation with the predicted semantic vector $\hat{s}$ is then chosen as the predicted meaning.$^2$

For production, the evaluation process is more complex because a predicted form vector $\hat{\mathbf{c}}$ 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 amount of semantic support less than a given threshold $\theta$. In a second step, the algorithm constructs all possible sequences of triphones that

![Figure 1: A sample lexicon with four words. Simulated semantic vectors are marked in blue, and predicted semantic vector is represented by the red dashed arrow. As the predicted semantic vector is the closest to the vector of pengajar, the predicted meaning is pengajar.](image)

$^2$ Note that in equations (6) and (7), the elements of the predicted matrices are obtained by simple summation. In network terminology, the activations received from incoming connections are summed and are not subjected to further modification by a squashing function, as is usually the case in multi-layer networks.
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, \text{ABC} and \text{BCD} properly overlap, but \text{ABC} and \text{PCD} do not. Finally, the algorithm calculates for each path the corresponding semantic vector using equation (6) and selects that path for articulation for which the predicted semantic vector is closest to the semantic vector targeted for production.

LDL does not make any claims about how actual neurons work together in the brain to enable lexical processing, the complexity of which exceeds by many orders of magnitude the complexity of the simple two-layer networks that LDL makes use of. What LDL does provide is a high-level functional characterization of the problem of learning mappings between form and meaning, when form and meaning are represented by high-dimensional vectors. Since LDL gives the mathematically simplest solution for this learning problem, the model must be too simple. But this makes it possible to use the model as a tool for tracing what aspects of morphological systems are the most challenging to learn. In what follows, we show how it can be used to formalize functional load. First, however, we introduce the dataset that we have compiled and studied.

2.1 Dataset

The initial data was retrieved from Leipzig Corpora Collection available at https://wortschatz.uni-leipzig.de/en/download/Indonesian#ind_mixed_2013, 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 \text{PE}- or \text{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 \text{PEN\_penge-} and \text{MEN\_menge-} were not present in our dataset. We then added the derived words with \text{PEN\_penge-} and \text{MEN\_menge-} to our dataset, and also included inflected forms (e.g., -\text{ku}, -\text{mu}, and -\text{nya} for first, second, and third person singular possessives or objects, \text{ku-} and \text{kau-} for first and second person subjects, as well as the marker of emphasis -\text{lah} and the question marker -\text{kah} (Kridalaksana 2007; 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 \text{MEN-} were attested in our dataset. For most of these verbs, the corresponding word with \text{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 2012), 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-.

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 $C$ 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 $S$, we simulated numeric vectors of length 852. These vectors were constructed by adding the vectors of a word’s content lexeme and its inflectional and derivational lexemes. 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 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., $ter/-kan$, menN/-kan, meN/-i, ber/-an). Whether Indonesian has ‘inflection’ is under debate – in Austronesian languages, distinguishing between derivational and inflectional affixes, as well as clitics, is not always straightforward (Levin and Polinsky 2021). In what follows, we will use the term ‘inflection’ descriptively, to refer to the expression of object person (-$ku$, -$mu$, -$nya$) and mood (-lah, -kah).

Table 3 lists the semantic features and their lexemic 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 lexeme $ajar$ and for the function lexemes 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 4.

Derived words can be ambiguous. For instance, *berpukulan* can have either a possesive reading, \([\text{ber} + [\text{pukul}]_{\text{N}} + \text{an}]_{\text{V}}\) ‘to have the ability to deliver a real punch’ or a reciprocal reading \([\text{ber} + [\text{pukul}]_{\text{N}} + \text{an}]_{\text{V}}\) ‘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 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
Table 4: Examples of Indonesian derived words for the base word *ajar*

<table>
<thead>
<tr>
<th>Word</th>
<th>Animacy</th>
<th>Concreteness</th>
<th>Aspect</th>
<th>Manner</th>
<th>SemanticRole</th>
<th>Voice</th>
<th>Transitivity</th>
<th>ObjectSemanticRole</th>
<th>Volition</th>
</tr>
</thead>
<tbody>
<tr>
<td>terajar</td>
<td></td>
<td></td>
<td></td>
<td>causative</td>
<td>passive transitive</td>
<td></td>
<td></td>
<td>abilitative</td>
<td></td>
</tr>
<tr>
<td>terajarkan</td>
<td></td>
<td></td>
<td></td>
<td>causative</td>
<td>passive transitive</td>
<td></td>
<td></td>
<td>abilitative</td>
<td></td>
</tr>
<tr>
<td>berpelajaran</td>
<td>inanimate</td>
<td>abstract</td>
<td>result</td>
<td>action</td>
<td>active intransitive</td>
<td></td>
<td></td>
<td>theme</td>
<td></td>
</tr>
<tr>
<td>mengajar</td>
<td></td>
<td></td>
<td></td>
<td>active</td>
<td>active transitive</td>
<td>theme</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mengajarimu</td>
<td></td>
<td></td>
<td></td>
<td>locative</td>
<td>active transitive</td>
<td>theme</td>
<td></td>
<td>patient object</td>
<td></td>
</tr>
<tr>
<td>diajar</td>
<td></td>
<td></td>
<td></td>
<td>active</td>
<td>passive transitive</td>
<td>theme</td>
<td></td>
<td>theme, beneficiary</td>
<td></td>
</tr>
<tr>
<td>diajarkan</td>
<td></td>
<td></td>
<td></td>
<td>passive</td>
<td>passive transitive</td>
<td>theme</td>
<td></td>
<td>theme, beneficiary</td>
<td></td>
</tr>
<tr>
<td>diajarkannya</td>
<td></td>
<td></td>
<td></td>
<td>passive</td>
<td>passive transitive</td>
<td>theme</td>
<td></td>
<td>theme, beneficiary</td>
<td></td>
</tr>
<tr>
<td>pelajar</td>
<td>animate</td>
<td>concrete</td>
<td></td>
<td></td>
<td>patient</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>pelajarku</td>
<td>animate</td>
<td>concrete</td>
<td></td>
<td></td>
<td>patient</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ajarannya</td>
<td>inanimate</td>
<td>abstract</td>
<td>result</td>
<td></td>
<td>result action</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>pelajaran</td>
<td>inanimate</td>
<td>abstract</td>
<td>result</td>
<td>action</td>
<td>animate concrete</td>
<td>agent</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>pembelajaran</td>
<td>inanimate</td>
<td>abstract</td>
<td>result</td>
<td>action</td>
<td>animate abstract</td>
<td>agent</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Word</th>
<th>Manner</th>
<th>Aspect</th>
<th>State</th>
<th>ChangeOfObject</th>
<th>PronounPerson</th>
<th>PronounFunction</th>
<th>NyaFunction</th>
<th>Mood</th>
</tr>
</thead>
<tbody>
<tr>
<td>terajar</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>state</td>
</tr>
<tr>
<td>terajarkan</td>
<td>causative</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>state</td>
</tr>
<tr>
<td>berpelajaran</td>
<td>action</td>
<td>result</td>
<td>possession</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>state</td>
</tr>
<tr>
<td>mengajar</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>second</td>
</tr>
<tr>
<td>mengajarimu</td>
<td>locative</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>second</td>
</tr>
<tr>
<td>diajar</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>state</td>
</tr>
<tr>
<td>diajarkan</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>state</td>
</tr>
<tr>
<td>Word</td>
<td>Manner</td>
<td>Aspect</td>
<td>State</td>
<td>ChangeOfObject</td>
<td>PronounPerson</td>
<td>PronounFunction</td>
<td>NyaFunction</td>
<td>Mood</td>
</tr>
<tr>
<td>------------------</td>
<td>--------</td>
<td>--------</td>
<td>----------</td>
<td>----------------</td>
<td>---------------</td>
<td>----------------</td>
<td>-------------</td>
<td>------------</td>
</tr>
<tr>
<td>diajarkannya</td>
<td></td>
<td></td>
<td>state</td>
<td></td>
<td>third</td>
<td>subject</td>
<td>NyaSubject</td>
<td></td>
</tr>
<tr>
<td>pelajar</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>pelajarku</td>
<td></td>
<td></td>
<td>first</td>
<td></td>
<td>third</td>
<td>possessive</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ajarannya</td>
<td></td>
<td>result</td>
<td></td>
<td></td>
<td>third</td>
<td>possessive</td>
<td>NyaPossessive</td>
<td>emphasize</td>
</tr>
<tr>
<td>pelajaran</td>
<td>action</td>
<td></td>
<td>result</td>
<td></td>
<td>third</td>
<td>possessive</td>
<td></td>
<td></td>
</tr>
<tr>
<td>pembelajaranmu</td>
<td>action</td>
<td>process</td>
<td>regularity</td>
<td></td>
<td>second</td>
<td>possessive</td>
<td></td>
<td></td>
</tr>
<tr>
<td>pengajar</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>pengajarlah</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ajaran</td>
<td></td>
<td></td>
<td>state</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>imperative</td>
</tr>
</tbody>
</table>

Table 4: (continued)
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; Kroeger 2007; Sneddon et al. 2010; Sutanto 2002; Tomasowa 2007). 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 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.

2.3 Accuracy

For the 2,517 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%).

With respect to production accuracy, the lexeme was predicted 100% correctly by the model. The same 100% accuracy also holds for Animacy, Voice, Manner, Transitivity, Volition, Aspect, State, Gradation, ChangeOfObject, Base-Relationship, 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. In actual lexical processing, the context in which words and morpheme occur can further constrain the mappings between form and meaning. Since the current version of LDL is a ‘local’ model of morphology, such contextual constrains cannot be taken into account.
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. 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).

The error made for *pekasih*, incorrectly understood as *kekasih*, is an interesting one. It has been observed (Chaer 2008; Ernanto 2016; Ramlan 2009; Sneddon et al. 2010; 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 (Benjamin 2009; Dardjowidjojo 1983). 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*.-.

Production accuracy for the *PE*- and *PEN*- words was at 100% (109 words) and 96% (211 out of 221 words) respectively. Table 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 #po, pen, ena, nap, apa, pan, ana, na#. 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
Table 5: Comprehension errors involving PE- and PEN-, including omissions and intrusions.

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

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

3 Results

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 $\hat{c}$ and the gold standard ($c$, production), or the strength of the correlation between the predicted semantic vector $\hat{s}$ and the gold standard semantic vector ($s$), can vary considerably. Figure 2 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

![Figure 2](image-url)

(a) comprehension (b) production.

**Figure 2**: 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 uninflated words. (a) comprehension (b) production.
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 (see the upper left panel of Figure 2), a Wilcoxon test
clarified that the mean correlation between target and predicted form is higher for
PE- (0.902) than for PEN- (0.859, \( W = 17,458, p < 0.0001 \)). When we subset PE- and
PEN- into those words that have an inflectional exponent and those that do not, as
shown in the upper right panel of Figure 2, the same pattern that PE- is recognised
more accurately than PEN- is also observed. For inflected PEN- (0.881) and PE-
(0.917), \( W = 7,941, p < 0.0001 \), and for uninflected PEN- (0.817) and PE- (0.867),
\( W = 1,964, p < 0.0001 \). For production (see the lower panels of Figure 2), the pattern
that PE- is produced more accurately than PEN- is virtually the same (PE- (0.91)
versus PEN- (0.858): \( W = 17,430, p < 0.0001 \), inflected PE- (0.917) and PEN- (0.879):
\( W = 1,983, p < 0.0001 \), uninflected PE- (0.872) and PEN- (0.818): \( W = 7,875, p < 0.0001 \)).

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 2 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 (\( W = 13,480, p = 0.0782 \) for comprehension, and
\( W = 13,721, p = 0.04 \) 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.

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 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 discrimi-
native 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-.
Table 7: Examples of distinct and shared triphones for PE- and PEN-, and their corresponding verbal prefixes BER- and MEN-

<table>
<thead>
<tr>
<th>Base word</th>
<th>English Noun</th>
<th>Prefix</th>
<th>English Verb</th>
<th>Distinct triphones</th>
<th>Shared triphone</th>
</tr>
</thead>
<tbody>
<tr>
<td>ajar</td>
<td>lesson</td>
<td>pengajar</td>
<td>to teach a lesson</td>
<td>#pe, pen, #me, men</td>
<td>eng, nga, gaj, lar, ar#</td>
</tr>
<tr>
<td>cinta</td>
<td>love</td>
<td>pencinta</td>
<td>who is very enthusiastic</td>
<td>mencinta to love</td>
<td>#pe, pen, #me, men</td>
</tr>
<tr>
<td>cinta</td>
<td>love</td>
<td>pecinta</td>
<td>who makes love</td>
<td>bercinta to make love</td>
<td>#pe, pec, ecl, #be, ber, etc, rcl</td>
</tr>
<tr>
<td>suruh</td>
<td>order</td>
<td>penyuruh</td>
<td>to give an order</td>
<td>menyuruh to give an order</td>
<td>#pe, pen, #me, men</td>
</tr>
<tr>
<td>suruh</td>
<td>order</td>
<td>pesuruh</td>
<td>who is commanded</td>
<td>pesuruh to command</td>
<td>#pe, pes, #esu, #pe, pel, #be, ber, el, rja</td>
</tr>
<tr>
<td>suruh</td>
<td>order</td>
<td>pejalan</td>
<td>pedestrian</td>
<td>berjalan to walk</td>
<td>#pe, pes, #esa, #sa, sak, aki, kit, it#</td>
</tr>
<tr>
<td>jalan</td>
<td>street</td>
<td>pejalan</td>
<td>pedestrian</td>
<td>berjalan to walk</td>
<td>#pe, pes, #esa, #sa, sak, aki, kit, it#</td>
</tr>
<tr>
<td>sakit</td>
<td>ill</td>
<td>pesakit</td>
<td>person</td>
<td>sick</td>
<td>#pe, pes, #esa, #sa, sak, aki, kit, it#</td>
</tr>
</tbody>
</table>
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 ($W = 14,974, p < 0.0005$). 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).

### 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 is very enthusiastic about 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 $\hat{s}$. Let $c_\tau$ denote a form vector for which the weights from triphone $\tau$ have been set to zero. In what follows, we refer to the semantic vector that is predicted by $c_\tau$ as $\hat{s}_\tau$. The functional load $L_\tau$ of triphone $\tau$ can now be assessed as the difference between the correlation of the original estimated vector $\hat{s}$ with the gold standard vector $s$ and the correlation of the gold standard vector $s$ with the vector $\hat{s}_\tau$ predicted by $c_\tau$:

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

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 $r(s, \hat{s}_\tau)$, and as a consequence, its functional load $L_\tau$ will be large.
The upper panel of Figure 3 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- ($W = 8,467, p < 0.0001$) and that the second triphone of PE- has a higher functional load than the second triphone of PEN- ($W = 17,438, p < 0.0001$). There is no significant difference between the third triphones ($W = 10,529, p < 0.0631$). The lower panel of Figure 3 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-.

**Figure 3:** Summaries of the distribution of $L_\tau$, using boxplots. Upper panel: functional load for the first three triphones of words with PE- (red) and PEN- (blue). Lower 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.
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 ($W = 10,849, p < 0.1426$). 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 4 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.

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- ($W = 10,355, p < 0.0001$). 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 4 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 4 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 4, the different allomorphs show the same general pattern, but also exhibit considerable variation. The pattern for the PEpe- 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, 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.

Figure 5: Functional load of triphones (ordered by position in the word) for word triplets with PEN-, MEN-, and PE- that share the same base word.
4 General discussion

In this study, we addressed the question whether the prefix \textit{PEN}- is easier to learn than its rival prefix \textit{PE}-, thanks to \textit{PEN}- showing a systematic relation with base words with the prefix \textit{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 \textit{PE}-. Importantly, the predicted form and meaning vectors showed stronger correlations with the targeted gold standard vectors for \textit{PE}- as compared to \textit{PEN}-. The presence of a difference as such dovetails well with several studies reporting qualitative and quantitative differences between these prefixes (Denistia et al. 2022; Denistia and Baayen 2019; Ramlan 2009; Sneddon et al. 2010). However, the present finding suggests, surprisingly, that the paradigmatic relation between \textit{PEN}- and \textit{MEN}- may come with a small learning disadvantage, instead of a learning advantage. This in turn predicts that \textit{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. 2020b).

One reason that \textit{PE}- is learned more robustly is that \textit{PE}- has more inflected variants, which help make words with this prefix more discriminable. Denistia and Baayen (2019) observed that although \textit{PE}- is less productive than \textit{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 \textit{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 \textit{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 \textit{PE}- is that the triphones shared by \textit{PEN}- and \textit{MEN}- are in competition. For instance, the \textit{enX} triphone cue (with \textit{X} representing the first phone of the base word) has to compromise between the verbal and nominal meanings associated with \textit{PEN}- and \textit{MEN}-. Furthermore, due to the formal similarity of \textit{PEN}- and \textit{MEN}-, words with \textit{PEN}- have fewer distinctive cues compared to words with \textit{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 \textit{PEN}- and \textit{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 precise 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: \( L_\tau \). This measure gauges to what extent the similarity between the predicted semantic vector and the targeted semantic vector decreases when a triphone \( \tau \) 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. Furthermore, 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 (Hay 2003; Hay and Baayen 2003; Seidenberg 1987). 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 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-. We note here that the present study has followed Indonesian orthography, and that it will be fruitful to conduct further simulations that are strictly phone-based (using triphones rather than trigrams) in order to obtain more precision with respect to where in the speech signal the allomorphs pen-, peng-, and peny-, are discriminated.

In the literature, studies on the nasal/plosive alternation in Austronesian languages have focused on the initial segment (see, e.g. Blust 2004; Halle and Clements 1983; Pater 1999; Ramlan 2009; Sugerman 2016; Sukarno 2017), 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 van Marle 2016 [1984], 1986: for an extended discussion of affix substitution). 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 (Blust 2004; Dempwolff 1934).

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., Arndt-Lappe 2011; Eddington 2002; Daelemans and van den Bosch 2005; Krott 2001), 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., Chuang and Baayen 2021; Milin et al. 2017).

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. 2021). 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.

Data availability statement

The dataset generated and analyzed during this study are available online at https://bit.ly/PePeNwithLDL.

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

References


