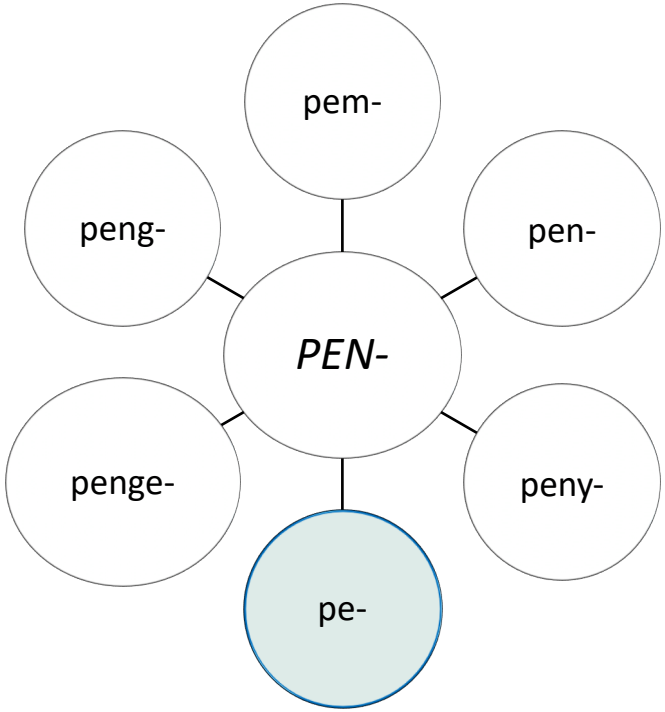
# Affix substitution in Indonesian and its impact for discriminative learning

Karlina Denistia and R. Harald Baayen
karlinadenistia@gmail.com

# Presentation overview

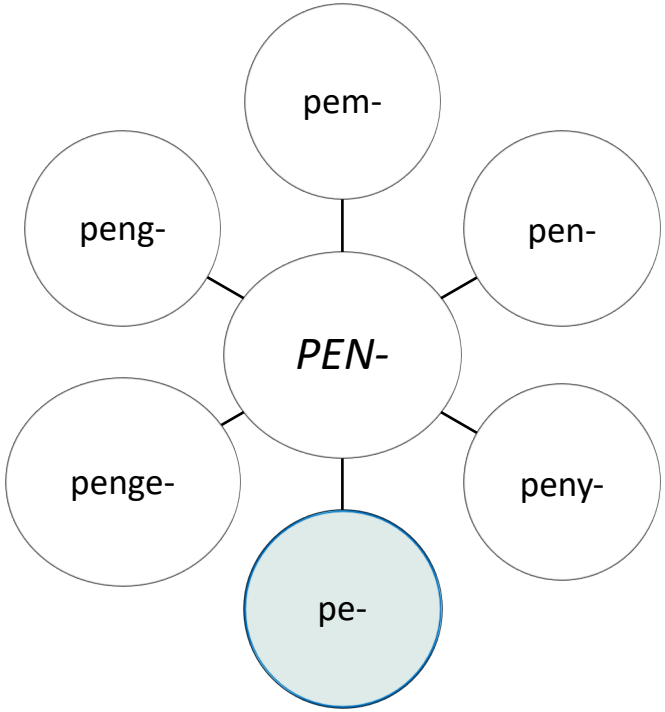
- Background of the study
- Affix substitution in Indonesian morphology
- Methods
- Results
- Discussion

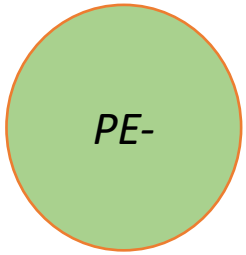
# Form and meaning similarity




/pə ˈneɪz(ə)l/
*PE-N*asal, the *N* stands for ’nasal’

# Form and meaning similarity




/pə ˈneɪz(ə)l/
*PE-N*asal, the *N* stands for 'nasal'

*/pə/*

# Examples

| | | |
|---|---|---|
| *PEN* + lukis | **pe**lukis | ‘painter’ |
| *PEN* + warta | **pe**warta | ‘broadcaster’ |
| *PEN* + murni | **pe**murni | ‘purifier’ |

| | | |
|---|---|---|
| *PE* + lari | **pe**lari | ‘runner’ |
| *PE* + wisata | **pe**wisata | ‘traveler’ |
| *PE* + mukim | **pe**mukim | ‘citizen’ |

# Presentation overview

- Background of the study
- Affix substitution in Indonesian morphology
- Methods
- Results
- Discussion

# Verb-noun paradigmatic relation

MEN-

/mə ˈneɪz(ə)l/

| | |
|---|---|
| **me**lukis | ‘to paint’ |
| **men**cari | ‘to search’ |
| **meng**gambar | ‘to draw’ |

*PEN-*

/pə ˈneɪz(ə)l/

| | |
|---|---|
| **pe**lukis | ‘painter’ |
| **pen**cari | ‘search engine’ |
| **peng**gambar | ‘drawer’ |

# Verb-noun paradigmatic relation

MEN-

/mə ˈneɪz(ə)l/

| | |
|---|---|
| **me**lukis | ‘to paint’ |
| **men**cari | ‘to search’ |
| **meng**gambar | ‘to draw’ |

*PEN-*

/pə ˈneɪz(ə)l/

| | |
|---|---|
| **pe**lukis | ‘painter’ |
| **pen**cari | ‘search engine’ |
| **peng**gambar | ‘drawer’ |

BER-

/bər/

| | |
|---|---|
| **ber**lari | ‘to run’ |
| **ber**wisata | ‘to travel’ |
| **ber**mukim | ‘to stay’ |

*PE-*

/pə/

| | |
|---|---|
| **pe**lari | ‘runner’ |
| **pe**wisata | ‘traveler’ |
| **pe**mukim | ‘citizen’ |

Sneddon et al. (2010)

# Affix substitution in Indonesian

| Noun | English Noun | Verb | English Verb |
|---|---|---|---|
| **pen**cinta | who loves something | **men**cinta | to love |
| **pen**inju | who punches | **men**inju | to punch |
| **penge**cek | checker | **menge**cek | to check |
| **pe**lukis | painter | **me**lukis | to paint |
| **peng**ajar | teacher | **meng**ajar | to teach |

# Affix substitution in Indonesian

| Noun | English Noun | Verb | English Verb |
|---|---|---|---|
| **pen**cinta | who loves something | **men**cinta | to love |
| **pen**inju | who punches | **men**inju | to punch |
| **penge**cek | checker | **menge**cek | to check |
| **pe**lukis | painter | **me**lukis | to paint |
| **peng**ajar | teacher | **meng**ajar | to teach |

*PEN-* and *MEN-* are discriminable in one phoneme: [p] and [m]

# Affix substitution in Indonesian

| Noun | English Noun | Verb | English Verb |
|---|---|---|---|
| **pen**cinta | who loves something | **men**cinta | to love |
| **pen**inju | who punches | **men**inju | to punch |
| **penge**cek | checker | **menge**cek | to check |
| **pe**lukis | painter | **me**lukis | to paint |
| **peng**ajar | teacher | **meng**ajar | to teach |

*PEN-* and *MEN-* are discriminable in one phoneme: [p] and [m]

Does the form similarity between *PEN-* (and its allomorphs) and *MEN-* (and its allomorphs) facilitate learning?

Regularity facilitates prediction (Blevins et al., 2017)

# Affix substitution in Indonesian

| Noun | English Noun | Verb | English Verb |
|---|---|---|---|
| **pen**cinta | who loves something | **men**cinta | to love |
| **pen**inju | who punches | **men**inju | to punch |
| **penge**cek | checker | **menge**cek | to check |
| **pe**lukis | painter | **me**lukis | to paint |
| **peng**ajar | teacher | **meng**ajar | to teach |

*PEN-* and *MEN-* are discriminable in one phoneme: [p] and [m]

Does the form similarity between *PEN-* (and its allomorphs) and *MEN-* (and its allomorphs) facilitate learning?

Regularity facilitates prediction (Blevins et al., 2017)

# Presentation overview

- Background of the study
- Affix substitution in Indonesian morphology
- Methods
- Results
- Discussion

# Database

- Leipzig Corpora Collection (36M word tokens) (Goldhahn et al., 2012)

# Database

- Leipzig Corpora Collection (36M word tokens) (Goldhahn et al., 2012)

- 99 mono-morphemic words (adjectives, verbs, nouns, and adverbs)
  - the highest counts of derived words are attested
  - at least one derived word with *PE-* or *PEN-* is attested
  - expanded for the derived and inflected formations

# Database

- Leipzig Corpora Collection (36M word tokens) (Goldhahn et al., 2012)

- 99 mono-morphemic words (adjectives, verbs, nouns, and adverbs)
  - the highest counts of derived words are attested
  - at least one derived word with *PE-* or *PEN-* is attested
  - expanded for the derived and inflected formations

- Size of database: 2517 non-reduplicated words
  - 109 words with *PE-*
  - 221 words with *PEN-*

# Word and Paradigm

concentrating on the word form rather than segments of the word

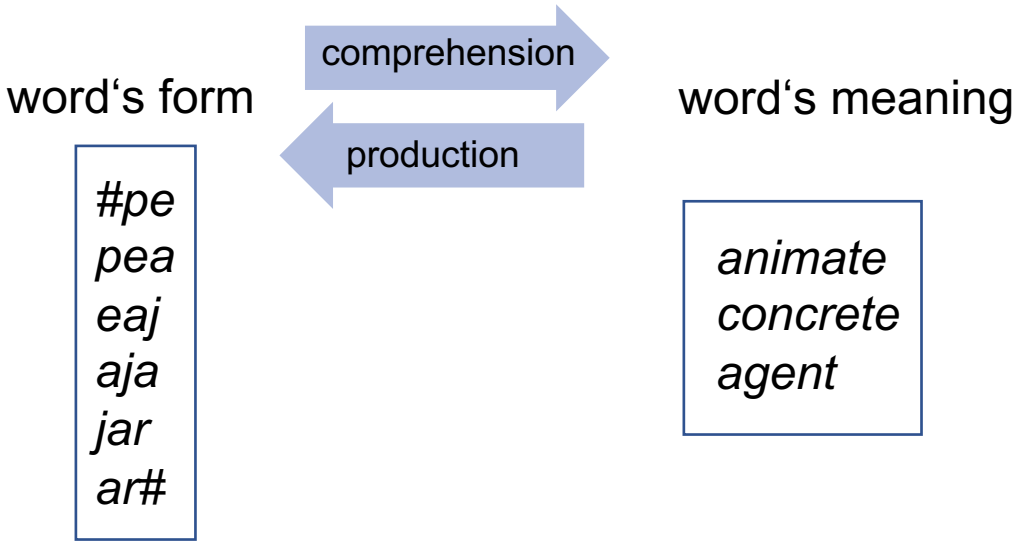
(Matthews, 1974; Blevins, 2016)

| Lexeme | Word | Animacy | Concreteness | SemanticRole |
|---|---|---|---|---|
| ajar | ajar | inanimate | abstract | |
| ajar | pengajar | animate | concrete | agent |

# Computational modelling

| Lexeme | Word | Animacy | Concreteness | SemanticRole |
|---|---|---|---|---|
| ajar | ajar | inanimate | abstract | |
| ajar | pengajar | animate | concrete | agent |

- The 'discriminative lexicon' (DL) model (Baayen et al., 2019)



# Presentation overview

- Background of the study
- Affix substitution in Indonesian morphology
- Methods
- Results
- Discussion

# Model performance

- 94% accuracy for both comprehension and production
- Comprehension dataset

| Lexeme | form | target | rank | correct | r | Animacy | Concreteness | SemanticRole |
|---|---|---|---|---|---|---|---|---|
| ajar | pelajar | pelajar | 1 | TRUE | 0.8845222 | animate | concrete | patient |
| ajar | pengajar | pengajar | 1 | TRUE | 0.8548545 | animate | concrete | agent |
| dagang | pedagang | pedagang | 1 | TRUE | 0.8758514 | animate | concrete | agent |
| dagang | pedagang | pendagang | 2 | FALSE | 0.7691482 | animate | concrete | agent |
| data | pendata | pendata | 1 | TRUE | 0.8205853 | animate | concrete | agent |
| data | pendata | pendatanya | 2 | FALSE | 0.7150659 | animate | concrete | agent |

# Model performance

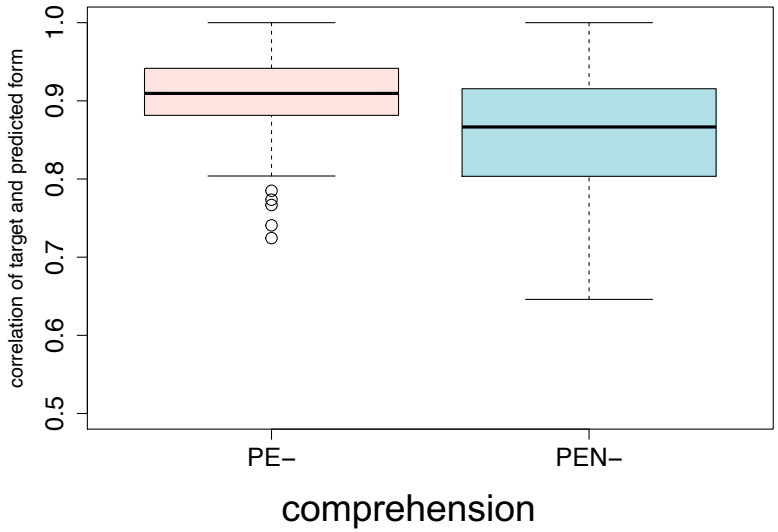
- 94% accuracy for both comprehension and production
- Comprehension dataset

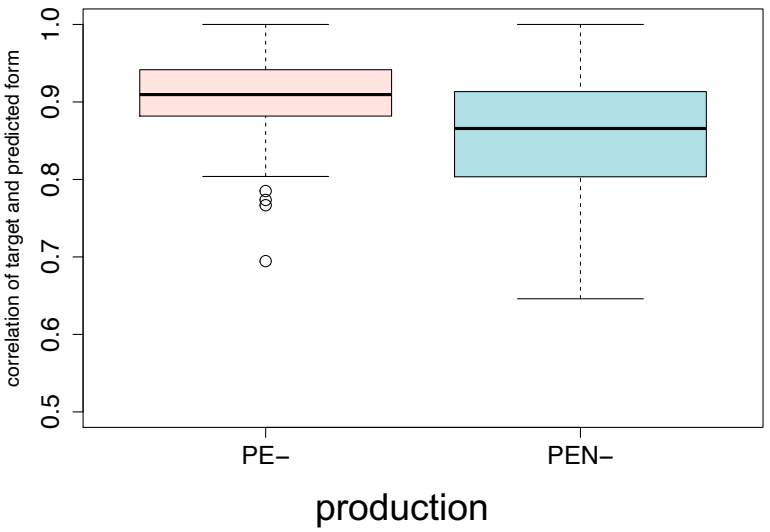
| Lexeme | form | target | rank | correct | r | Animacy | Concreteness | SemanticRole |
|---|---|---|---|---|---|---|---|---|
| ajar | pelajar | pelajar | 1 | TRUE | 0.8845222 | animate | concrete | patient |
| ajar | pengajar | pengajar | 1 | TRUE | 0.8548545 | animate | concrete | agent |
| dagang | pedagang | pedagang | 1 | TRUE | 0.8758514 | animate | concrete | agent |
| dagang | pedagang | pendagang | 2 | FALSE | 0.7691482 | animate | concrete | agent |
| data | pendata | pendata | 1 | TRUE | 0.8205853 | animate | concrete | agent |
| data | pendata | pendatanya | 2 | FALSE | 0.7150659 | animate | concrete | agent |

- Production dataset

| target | predicted | cors | correct |
|---|---|---|---|
| pelajar | pelajar | 0.8845222 | TRUE |
| pendata | pendata | 0.8205853 | TRUE |
| pendatanya | pendata | 0.7550630 | FALSE |
| pedagang | pedagang | 0.8758514 | TRUE |
| pendagang | pendagang | 0.7576054 | TRUE |
| pengajar | pengajar | 0.8548545 | TRUE |

# Difference in correlation strengths




Correlations are higher for *PE*- than for *PEN*-

$p < 0.0001$ for all comparison

# Reason of difference in correlation strengths

1. Cue competition between *PEN-* and *MEN-*

| 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# |

| Triphones | Verb paradigm with *MEN-* | | | Noun paradigm with *PEN-* | | |
|---|---|---|---|---|---|---|
| eng, nga, gaj, aja, jar, ar# | | | | animate | concrete | agent |
| | active | transitive | theme | | | |
| enc, nci, cin, int, nta, ta# | | | | animate | concrete | agent |
| | active | intransitive | | | | |

| 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# |

the more discriminative cues a word have = the better the model will learn the word

# Reason of difference in correlation strengths

2. *PE-* tends to be longer in form than *PEN-*

| Comparison | *PE-* | *PEN-* | p value |
|---|---|---|---|
| **Words in dataset** | 109 | 211 | |
| **Mean length in characters** | 7.4 | 6.6 | $p < 0.0005$ |
| **Probability of being inflected** | 71% | 66% | $p = 0.53$ |

the longer the words = the more discriminative cues a word have
= the better the model will learn the word

# Presentation overview

- Background of the study
- Affix substitution in Indonesian morphology
- Methods
- Results
- Discussion

# Discussions

- Computational modeling with linear discriminative learning revealed the predicted form and meaning showed stronger correlations for *PE-* as compared to *PEN-*

# Discussions

- Computational modeling with linear discriminative learning revealed the predicted form and meaning showed stronger correlations for *PE-* as compared to *PEN-*

- 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).

# References


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.

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.

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.

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.

Matthews, P. H. (1974). *Morphology. An Introduction to the Theory of Word Structure.* Cambridge University Press, Cambridge.

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.