The morphology of Indonesian: Data and quantitative modeling

Karlina Denistia[1] and R. Harald Baayen[2]
[1]Vocational School - Sebelas Maret University,
[2]Quantitative Linguistics - Eberhard Karls Universitaet Tübingen

Abstract

This chapter first provides an overview of Indonesian morphology. It then proceeds to show how the morphology of Indonesian can be modeled using a computational error-driven model for lexical processing in the mental lexicon. The chapter concludes with a case study of two rival prefixes and illustrates how methods from corpus linguistics and computational modeling contribute to our understanding of affixal rivalry and morphological productivity.

Keywords: Indonesian morphology, reduplication, computational modeling, functional load, linear discriminative learning, productivity

1 Introduction

The standard Indonesian language (*Bahasa Indonesia*) is the official and national language of Indonesia. Indonesian serves as a lingua franca for more than 300 ethnic groups speaking around 700 local languages. Indonesian belongs to the Austronesian (Malayo-Polynesian) language family. Austronesian languages are characterised by rich morphology, with prefixation, suffixation, circumfixation, affix substitution, reduplication, compounding, and cliticisation. Indonesian is no exception.

This chapter has three parts. The first part provides an overview of the morphology of Indonesian. The second part shows how a recent computational model of the mental lexicon proposed by Baayen et al. (2019) can be used to understand strengths and weaknesses in Indonesian word formation for comprehension and production. This model integrates insights from distributional semantics and Word and Paradigm morphology, and by doing so makes it possible to consider the full morphology of Indonesian as a system. The third part of our chapter presents a case study of the derivational prefixes *pe-* and *peN-*, which realize agent and instrument nouns, and which stand in a close paradigmatic relation with the verb-forming prefixes *ber-* and *meN-*. (The *N* in *peN-* and *meN-* denotes that the prefix undergoes nasal assimilation). We illustrate how methods from corpus linguistics, computational linguistics, distributional semantics, and computational modeling can be used to probe the the dynamics of these rival prefixes.

2 An overview of Indonesian morphology

In Indonesian, word formation makes use of prefixation, suffixation, infixation, circumfixation, reduplication, and compounding (Sneddon et al., 2010; Ermanto, 2016; Sugerman, 2016; Chaer, 2008; Putrayasa, 2008). Examples of word formation processes in Indonesian are presented in Table 1.

2.1 Prefixation

An interesting feature of Indonesian prefixes is the nasal allomorphy that characterizes the verb-forming prefix *meN-* and the noun-forming prefix *peN-*. These prefixes have five nasal allomorphs: *peng-* and *meng-*, *pen-* and *men-*, *pem-* and *mem-*, *peny-* and *meny-*, *penge-* and *menge-*. This allomorphy is conditioned by the place of articulation of the initial segment of the stem. There is also one pair of allomorphs without a final nasal: *pe-* and *me-*, and one additional prefix, *pe-*, that does not participate in the nasal allomorphy. Evidence for distinguishing *pe-* from the homophonous allomorph of *peN-* will be discussed in section 3.

The prefixes *meN-* and *peN-* stand in a paradigmatic relation that can be analysed as involving affix substitution (van Marle, 1984). The same holds for the prefixes pe- and ber-. For instance, the base word *pukul*, 'a punch', underlies both the verb *memukul* 'to hit' and the nominalization *pemukul* 'someone who hits' or 'something to hit'. The initial consonants of the derived verb and the derived noun differ only in their manner of articulation: nasal versus stop. Nouns with peN-, such as *penari* 'dancer', are generally understood as being derived from the corresponding verbs with the prefix *meN-*, such as *menari*, 'to dance'

Word formation	Base word	Base translation	Derived word	Word translation	Semantics
			prefixation		
ber-	temu	to meet	bertemu	to meet each other	reciprocative
di-	bawa	to carry	dibawa	to be carried	passive
meN-	bantu	to help	membantu	to help	active-transitive
ter-	tidur	to sleep	tertidur	to sleep unintentionally	accidental
peN-	bantu	to help	pembantu	maid	agent
pe-	tembak	to shoot	petembak	shooter	athelete
meN-per-	mudah	easy	mempermudah	to make something easier	active-transitive-causative
di-per-	luas	wide	diperluas	to be made wider	passive-causative
			suffixation		
-kan	kata	to say	katakan	to make someone say something	causative
-i	temu	to meet	temui	to make someone meet someone else	causative
-an	baca	to read	bacaan	reading	nominaliser
			infixation		
-em-	jari	finger	jemari	finger	-
-el-	gembung	bloated	gelembung	bubble	-
-er-	gigi	teeth	gerigi	serration	-
			circumfixation		
ber-/-kan	dasar	base	berdasarkan	to have the base	having X
meN-/-kan	ajar	to teach	mengajarkan	to teach something to someone	active-transitive-beneficiary
meN-/-i	pukul	to hit	memukuli	to hit repeteadly	active-transitive-iterative
di-/-kan	temu	to find	ditemukan	to be found	passive
di-/-i	setuju	to agree	disetujui	to be agreed	passive
ter-/-kan	pisah	to be separated	terpisahkan	can be separated	intransitive
ber-/-an	anggap	to assume	beranggapan	to have assumption	having X
per-/-an	alat	tools	peralatan	many kinds of tools	many kinds of X
ter-/-i	tanding	to beat	tertandingi	can be defeated	intransitive
ter-/-an	tawa	laughter	tertawaan	to laugh intensively at something	intransitive
ke-/-an	lihat	to look	kelihatan	visible	adversative
se-/-nya	wajar	normal	sewajarnya	not too much, not too less	-
ber-ke-/-an	lanjut	to continue	berkelanjutan	to countinuously continue	intransitive
ber-per-/-an	alat	tools	berperalatan	having many kinds of tools	having many kinds of X
meN-per-/-kan	timbang	to weight	mempertimbangkan	to consider	active-transitive-causative
meN-per-/-i	ingat	to remember	memperingati	to commemorate	active-transitive-causative
meN-ber-/-kan	henti	stop	memberhentikan	to make someone stop	active-transitive-causative
di-per-/-kan	juang	to fight	diperjuangkan	to be fought for	passive-causative
			reduplication		
full	buku	book	buku-buku	books	plurality
	marah	angry	marah-marah	to be angry intensively	intensity
	dua	two	dua-dua	two by two	by group of X
imitative	sayur	vegetable	sayur-mayur	vegetables	many kinds of X
	gerak	movement	gerak-gerik	various movement at the same time	many kinds of X
	kaya	rich	kaya-raya	extraordinarily rich	intensity
partial	laki	man	lelaki	man/men	plurality
	daun	leaf	dedaunan	foliage	plurality
	pohon	tree	pepohonan	trees	plurality
affixed	lihat	to see	melihat-lihat	to see around	casual action
	kuda	horse	kuda-kudaan	horse toy	imitation
	bantu	to help	bantu-membantu	to help each other	reciprocal
	kecil	small	mengecil-ngecilkan	to make something very small	intensity
			clitics		
ku-	ambil	to take something	kuambil	I take	first singular subject pronoun
kau-	ambil	to take something	kauambil	you take	second singular subject pronoun
-ku	mencintai	to love someone	mencintaiku	to love me	first singular object pronoun
	buku	book	bukuku	my book	first singular possesive pronoun
-mu	mencintai	to love someone	mencintaimu	to love you	second singular object pronoun
	buku	book	bukumu	my book	second singular possesive pronoun
-nya	buku	book	bukunya	my book	third singular possesive pronoun
	buku	book	bukunya	my book	third singular possesive pronoun
	buku	book	bukunya	the book	definite determiner
-lah	buku	book	bukulah	the book	emphasis
-kah	dia	he/she	diakah	is he/she?	question
			compounding		
	C1, C2	C1, C2 translation	Compound	Compound translation	
adjective compounding	panjang,tangan	long, hand	panjang tangan	someone who likes stealing something	
noun compounding	orang, tue	human, old	orang tua	parents	
	daun, telinga	leaf, ear	daun telinga	outer ear	
	kepala, sekolah	head, school	kepala sekolah	head master	
verb compounding	gigit, jari	to bite, finger	gigit jari	feeling disappointed	
	makan, besar	to eat, big	makan besar	to eat meals	

Table 1: Examples of word formation processes in Indonesian

(Dardjowidjojo, 1983; Nomoto, 2006, 2017; Putrayasa, 2008; Benjamin, 2009; Ramlan, 2009; Sneddon et al., 2010; Ermanto, 2016). In section 4, we will discuss evidence that quantitatively, the productivity of the allomorphs of the nouns and the productivity of the allomorphs of the verbs are tightly connected, an observation that fits well with a paradigmatic structuring of the words with *meN-* and *peN-* in the lexicon of Indonesian.

The prefixes of Indonesian typically change the argument structure of the verb on the dimensions of voice (active/passive) and transitivity (active/transitive/causative/reciprocative). But they may also realize aspect and agency. The prefixes *meN-* and *di-* are voice markers signaling whether the verb is active or passive. The passive can be realized as well by *ter-*, but this exponent also can express perfective aspect. What sets *ter-* apart from *di-* is that, unlike *di-*, *ter-* expresses that the action happened accidentally. With regards to the verb-forming prefixes *meN-* and *ber-*, the former is found in both transitive and intransitive verbs, whereas *ber-* occurs mostly in intransitive verbs, often with a slight change in meaning. Examples illustrating the subtle semantics that Indonesian prefixes can express, often in combination with suffixes, are given in (1) for the base verb *jatuh*, 'to fall'.

(1) a. Dia jatuh.
he/she fall
'He/She falls.'

b. Dia menjatuhkan bukunya.
he/she drop book-his/her
'He/She drops his/her book.'

c. Buku dijatuhkan oleh dia.
book drop-passive by him/her
'The book is dropped by him/her.'

d. Bukuku terjatuh.
book-my has fallen (passive-perfective)
'My book has fallen accidentally.'

e. Dia kejatuhan buku.
he/she fall-passive book
'A book happened to fall on him/her.'

f. Buku-buku berjatuhan.
books fall
'Books are falling down.'

2.2 Suffixation

The suffixes of Indonesian mark valency changes in argument structure. The suffix *-kan* requires patients or benefactives to be realized as indirect objects, indicated by the prepositions *kepada* 'to' or *untuk* 'for' (Chung, 1976). The suffix *-i* can replace *-kan*, in which case it highlights the recipient in what in English would be a double object construction (Sukarno, 2017; Arka et al., 2009; Chung, 1976), see (2). There is some fluidity in this system, for instance, *-kan* can be followed immediately by an indirect object in ditransitive clauses (see (3), cf. Sneddon et al., 2010), and it may also happen that there is no valency change with *-i* (see (4), cf. Arka et al., 2009).

(2) a. Aku mengirim surat.
I send letter
'I send a letter.'

b. Aku mengirimkan surat untuk kamu.
I send letter to you
'I send a letter to you.'

c. Aku mengirimi kamu surat.
I send you letter
'I send you a letter.'

(3) a. Aku membuat kopi.
I make coffee
'I make a cup of coffee.'
b. Aku membuatkan kamu kopi.
I make you coffee
'I make a cup of coffee for you.'

(4) a. Dia memukul meja.
he/she hit table
'He/She hits the table.'
b. Dia memukuli meja.
he/she hit-repetitive table
'He/She hits the table repeatedly.'

2.3 Infixation

Unlike in other Austronesian languages, such as Tagalog, infixation is no longer productive in Indonesian (Chaer, 2008; Nomoto et al., 2018). As can be seen in Table 1, only three interfixes are in use, *-em-*, *-er-*, and *-el-*, and it is unclear what the semantic contribution of these interfixes is to their carrier words (Sneddon et al., 2010). It should be noted that *-em-* occurs only in words with reduplication, e.g., *tali* 'rope' - *tali-temali* 'all sorts of rope', *turun* 'descend' - *turun-temurun* 'hereditary'.

2.4 Circumfixation

Circumfixation is defined as the simultaneous addition of both a prefix and suffix. However, the theoretical status of circumfixation, which is productive in Indonesian, is controversial. The problem is illustrated by the verb *kelihatan* 'visible', which is derived from the base verb *lihat* 'to look'. As neither *ke-lihat* and *lihat-an* are attested in *Kamus Besar Bahasa Indonesia*, a comprehensive Indonesian dictionary (Alwi, 2012), *ke-/-an* can be argued to be an independent exponent Nomoto et al. (2018); Putrayasa (2008). Alternatively, circumfixation can be understood as a straightforward combination of prefixation and suffixation (Samuel, 2009; Kridalaksana, 2012), as prefix and suffix contribute semantics that are equivalent to that of the interfix. In what follows, we use the term circumfixation descriptively. In section 3, we will clarify that within a Word an Paradigm approach, discussions about the proper bracketing of circumfixed words is perhaps not that fruitful.

As an example of circumfixation, consider *ke-/-an*. Verbs with this circumfix are similar in meaning to verbs with the prefix *ter-*, in both cases, passive voice and perfective aspect is expressed. In contrast to the passive prefix *di-*, *ke-/-an* adds a goal/patient to its argument structure. The examples (5) are from Hidajat (2014).

(5) a. Gudang itu kebakaran.
Warehouse that KE-burn-AN
'That warehouse was on fire.'
b. Joni kejatuhan mangga.
Joni KE-fall-AN mango
'Joni was fallen on by a mango.'

Both examples illustrate perfective aspect. (5)(b) is an example of the addition of a goal/patient argument, 'Joni', the endpoint of the falling event.

The circumfix *ke-/-an* functions as a nominalizer when it attaches to adjectives (e.g., *aman* 'safe' - *keamanan* 'safety'). In this case, the circumfix *ke-/-an* can express either an abstract noun as in *tinggi* 'high' - *ketinggian* 'height', or excessive degree, as in *tinggi* 'high' - *ketinggian* 'too highTherefore, it is possible that one word (e.g., *ketinggian*) has two readings:

(6) a. Ketinggian air mencapai satu meter.
height water reach one meter

'The water level is up to one meter.'

b. Nadanya ketinggian. Aku tidak bisa menyanyikannya.
the-note too-high. I can not sing-it
'The note is too high. I cannot sing the song'

When *ke-/-an* attaches to nouns, it expresses 'having some property', e.g., *ibu* 'mother' - *keibuan* 'mother-like'.

2.5 Reduplication

Reduplication is a particularly interesting word formation process in Indonesian. Four kinds of reduplication are distinguished: full reduplication, imitative reduplication, partial reduplication, and affixed reduplication (Sugerman, 2016; Chaer, 2008; Rafferty, 2002; Mistica et al., 2009; Dalrymple and Mofu, 2011; Lander, 2003).

Full reduplication can be applied to nouns, verbs, adjectives, adverbs, connectors, and pronouns. Reduplicated adjectives express intensity (e.g., *marah-marah*, very angry) or contrast (e.g. *kecil* 'small' is used contrastively in the sentence *kecil-kecil bisa punya pacar*, 'although still very young, she/he has a girlfriend/boyfriend'). Reduplicated ordinal numerals indicate groupings (e.g., *dua-dua*, two by two).

For nouns, reduplication can be used to express plurality (e.g., *buku-buku*, books). However, explicit marking of plurality with reduplication is optional. Indonesian does not make systematic distinctions between singular and plural nouns, nor between countable and uncountable nouns (Corbett, 2000; Greenberg, 1972). Reduplicated nouns and their simple counterparts are often described as having no difference in meaning (see, e.g., Dalrymple and Mofu, 2011). However, as will become clear in section 4, some differences in the semantics of simple and reduplicated nouns can be seen using distributional semantics.

As expected for a word formation process, reduplicated words may vary substantially with respect to their semantic transparency. Whereas the plural reading of *buku-buku*, 'books', is quite transparent (but see section 4), many words with reduplication are semi transparent (e.g., *langit* 'sky' - *langit-langit* 'attic') or even opaque (*tiba* 'to arrive' - *tiba-tiba* 'suddenly').

Imitative reduplication is found for nouns, verbs and adjectives. Unlike full reduplication, it is not productive. For adjectives, imitative reduplication realizes intensity or contrast, just as full reduplication (e.g., *kaya-raya*, extremely rich). When applied to nouns and verbs, imitative reduplication expresses that many kinds of objects, or many kinds of activities, are intended (e.g., *sayur-mayur*, vegetables; *gerak-gerik*, various coincident movements).

As illustrated by the above examples, imitative reduplication effects a change in one of the reduplicants, involving either a consonant or a vowel. These changes are not well predictable, which may explain the lack of productivity of imitative reduplication.

In a majority of cases, the initial reduplicant is an existing word, and the second reduplicant does not exist by itself. But there are also cases where the second reduplicant is the base word (e.g., *coret* in *corat-coret* 'to make random drawing'), or neither reduplicant can occur alone (e.g., *mondar-mandir* 'back and forth', *teka-teki* 'riddle'). Occasionally, both reduplicants are bound forms (e.g., *mondar-mandir* 'to go back and forth' and *teka-teki* 'ridlle', in which neither *mondar, mandir, teka* nor *teki* is attested on the dictionary). Compounding can also lead to forms that fit the pattern of imitative reduplication (e.g., *cerai* 'to get divorced' and *berai* 'to be not unified' - *cerai-berai* 'to be scattered everywhere').

Partial reduplication is also not productive in Indonesian (Sneddon et al., 2010). Examples of partial reduplication, in which the initial reduplicant is reduced to a CV syllable, are *laki* 'man' - *lelaki* 'man/men', *daun* 'leaf' - *dedaunan* 'foliage', and *pohon* 'tree' *pepohonan* 'trees'. Some words with partial reduplication also carry the suffix *-an*.

Affixed reduplication describes cases where reduplication combines with affixation. Affixed reduplication is fully productive. Whereas complex nouns allow full reduplication (e.g., *bacaan* 'reading' - *bacaan-bacaan* 'readings') , verbs allow only the stem to be reduplicated (Sato and McDonnell, 2013). The range of possibilities this creates for verbs is illustrated in (7) for the verb *pukul*, 'to hit'.

(7)

pukulan-pukulan	plural of nominalization
pukul-memukul	reciprocal
memukul-mukul	iteration, active voice
memukul-mukuli	iteration with object focus
memukul-mukulkan	iteration with instrument focus
dipukul-pukul	iteration, passive voice

Cranberry stems may serve as reduplicants. For example, *menembak-nembak* 'to shoot repeatedly' is constructed from the verb *menembak* 'to shoot' and a stem, *nembak*, that is not listed in dictionaries. However, the form *nembak* has been observed in informal language use as a shortening of *menembak* 'to shoot' (Benjamin, 2009; Muhadjir, 1981).

2.6 Compounding

Indonesian also makes extensive use of compounding, creating new onomasiological units by juxtaposition of other words (Sugerman, 2016; Alisjahbana, 1954). As in other languages, some compounds have transparent meanings (e.g., *meja tulis* 'table write', i.e., 'writing table'), others have semi transparent meanings (e.g., *kepala sekolah* 'head school', i.e., 'headmaster'), whereas yet others have idiosyncratic, idiomatic meanings (e.g., *panjang tangan* 'long hand', i.e., 'thief').

It is often not clear for Indonesian where to draw the boundary between compounds and fixed phrases (Muslich, 2009; Chaer, 2008). For instance, a phrase such as *kakek nenek*, 'grand mother grand father' can be understood as a coordinative compound, and a phrase such as *baju baru* 'clothes new', i.e., 'new clothes') can be analysed as a subordinative compound (compare *blackbird* in English).

2.7 Clitics

Indonesian makes extensive use of pronominal clitics, reduced and bound forms of the free pronouns (Himmelmann, 2005; Sneddon et al., 2010). Some are proclitics, such as the clitic for the first person singular subject *ku-* and the second person singular subject *kau-*. Others are enclitics, such as the object and possessive forms *-ku, -mu*, and *-nya* (Kridalaksana, 2008; Mistica et al., 2012). These enclitics realize first, second, or third person objects when attaching to verbs. When attaching to nouns, they realize first, second, or third person possessives. The pronoun *-nya* is special, as it has four possible meanings, depending on the word category of its host and the context of use (see Pastika (2012); Sedeng (2015) and Grangé (2015) for further details). When attaching to a noun, *-nya* expresses either definiteness or the third person singular possessive. Furthermore, *-nya* realizes third person objects when attaching to active verbs with the prefix *meN-*, but the third person subject pronoun when it attaches to passive verbs with the prefix *di-*. Finally, there are two clitics (or perhaps particles) that have several interaction-related functions. For instance, *-lah* is used to express polite imperatives, or to provide emphasis, and *-kah* is used as a marker for questions (Himmelmann, 2005; Sugerman, 2016; Sneddon et al., 2010). Table 2 lists examples of these clitics. When a noun is followed by an adjectival modifier, the clitic attaches to the adjective (e.g., *bantuan danaku* 'my financial help').

2.8 Inflection?

Whether or not Indonesian actually has real inflection is not straightforward to decide (Levin and Polinsky, 2021; Kridalaksana, 2007). The pronominal clitics as used with verbs (e.g., *kaubantu* 'you help') perhaps come closest to inflection for person and number in Indo-European languages. However, these are clitics rather than affixes, and they can be separated from their hosts: *bantuan dana* 'a financial help' - *bantuan danaku* 'my financial help', *baju kuning* 'yellow dress' - *baju kuningku* 'my yellow dress'.

The exponents *ter-, ke-/-an*, and *se-*, which can realize superlatives and excessives (e.g., *penting* 'important', *terpenting* 'the most important'; *jauh* 'far', *kejauhan*, 'too far'), are semantically fully transparent, and do not change word class, and therefore could be classified as instances of inflection (Chaer, 2008; Putrayasa, 2008; Ermanto, 2016; Levin and Polinsky, 2021; Kridalaksana, 2007), if so desired.

Clitic	Base word	Base translation	Non reduced form	Clitised word	Word translation	Semantics
ku-	bantu	to help	aku bantu	kubantu	I help	first person singular subject
kau-	bantu	to help	engkau bantu	kaubantu	you help	second person singular subject
-ku	bantu	to help	membantu aku	membantuku	to help me	first person singular object
	bantuan	a help	-	bantuanku	my help	first person singular possesion
	bantuan dana	a financial help	-	bantuan danaku	my financial help	first person singular possesion
	baju kuning	a yellow dress	-	baju kuningku	my yellow dress	first person singular possession
-mu	bantu	to help	membantu kamu	membantumu	to help you	second person singular object
	bantuan	a help	-	bantuanmu	your help	second person singular possesion
	bantuan dana	a financial help	-	bantuan danamu	your financial help	second person singular possesion
	baju kuning	a yellow dress	-	baju kuningmu	your yellow dress	second person singular possession
-nya	bantu	to help	membantu dia	membantunya	to help him/her/it	third person singular object
	bantuan	a help	-	bantuannya	her/his help	third person singular possesion
	bantuan dana	a financial help	-	bantuan dananya	his/her financial help	third person singular possession
	baju kuning	a yellow dress	-	baju kuningnya	his/her yellow dress	third person singular possession
	bantuan	a help	-	bantuannya	the help	-nya determiner
-lah	bantu	to help	-	bantulah	please, help	polite imperative
	bantuan	a help	-	bantuanlah	the help	emphasis
	bantuan	a help	-	bantuanlah	the help	emphasis
	baju kuning	a yellow dress	-	baju kuninglah	the yellow dress	emphasis
-kah	bisa	be able to	-	bisakah	asking for an ability	question for a request

Table 2: Indonesian pronominal clitics

2.9 Affixal homophony and polysemy

Many exponents of Indonesian morphology depend for their interpretation on the properties of the word they attach to, or on the broader context of use (Pastika, 2012; Rajeg, 2013; Sukarno, 2017; Denistia, 2018; Rajeg et al., 2019; Sneddon et al., 2010; Samuel, 2009; Ermanto, 2016; Soekarno, 2010). (8) provides an overview of the kind of homophony and polysemy found in Indonesian, using in part examples presented earlier.

(8) the prefix *ter-*, which realizes passive-perfective, volition, or ability:

- *tabrak* 'to crash', *tertabrak* 'to be crashed',
  dia tertabrak mobil 'he/she was crashed by a car'
- *beli* 'to buy', *terbeli* 'accidentally being bought',
  baju itu tidak sengaja terbeli oleh dia 'she/he accidentally bought the clothes'
- *beli* 'to buy', *terbeli* 'be able to be bought',
  akhirnya, mobil itu terbeli oleh dia 'she/he finally managed to buy the car'

the suffix *-i*, which realizes causative, iteration or application:

- *datang* 'to come', *datangi* 'to come to X',
  datangi rumahnya 'come to his/her place'
- *tembak* 'to shoot', *tembaki* 'to shoot repeatedly',
  kenapa kamu tembaki pohon itu? 'why do you shoot that tree over and over again'
- *panas* 'hot', *panasi* 'to apply heat on something',
  panasi sticker dengan pengering rambut 'heat the sticker with a hairdryer'

the suffix *-kan*, which realizes causatives or benefactives:

- *panas* 'hot', *panaskan* 'to make something hot',
  panaskan air 'boil the water'
- *tulis* 'to write', *tuliskan* 'write on behalf of someone',
  tuliskan kalimat dengan menggunakan kata 'cinta' 'write (on behalf of the teacher) a sentence using the word 'love"

the circumfix *ke-/-an*, which when applied to adjectives realizes either nominalization or excessive degree:

- *tinggi* 'high', *ketinggian* 'height',
  dia mengukur ketinggian letusan debu vulkanis 'she/he measures the height of ash due to the volcano eruption'

–*kecil* 'small', *kekecilan* 'too small',
pastikan ukuran sepatunya tidak kekecilan. 'make sure that the shoe size is not too small'

Not only can affixes realize multiple meanings, but the same (or similar) meanings can be realized by different affixes. For instance, a causative reading can be expressed by either the prefix *per-* or the suffix *-kan*, compare *besar* 'big' - *perbesar* and *besarkan* 'to make something big' (Benjamin, 2009; Ogloblin, 1998; Sneddon et al., 2010). Likewise, the nominalizing prefixes *pe-* and *peN-* both appear in agent nouns: *lari* 'to run' - *pelari* 'runner'; *lompat* 'to jump' - *pelompat* 'jumper' (Ramlan, 2009; Sneddon et al., 2010; Denistia, 2018). *Pe-* and *peN-*, however, are distinguished by the absence vs presence of nasal allomorphy. For example, *peN-* is realized as *pen-* when this prefix attaches to a base word with initial /t/ (*tembak* 'to shoot', *penembak* 'someone who shoots'), whereas *pe-* attached to *tembak* results in *petembak* ('shooter (athlete)').

3 A computational model for the Indonesian mental lexicon

The preceding section presented a description of Indonesian morphology that was phrased in terms of stems, prefixes, and suffixes. However, the usefulness of the theoretical construct of the 'morpheme' as a minimal sign combining form and meaning has been questioned, and it plays no role in realizational theories of morphology (see, e.g., Stump, 2001; Blevins, 2016). This raises the question of how to model the morphology of Indonesian without morphemes. One possibility is to follow the realizational approach Stump (2001), which provides a formalism for realizing bundles of semantic features in phonological form (see Karttunen, 2003, for a finite-state implementation).

In this study, we take a different approach, and show how a computational model based on error-driven learning, the 'Discriminative Lexicon' (DL) model introduced by Baayen et al. (2019), can be set up to approximate the Indonesian mental lexicon. A methodological discussion of this approach is given by Heitmeier et al. (2021), and an overview of applications of this model is given by Chuang and Baayen (2021). In what follows, we first introduce the model, then we present the dataset that was given to the model for learning, and finally we report model performance and the theoretical implications of the results.

3.1 Representing form and meaning in the DL

The DL model sets up numeric representations for words' forms on the one hand, and for their meanings on the other hand. These representations (technically, numeric vectors) are brought together as the rows of two tables (technically, matrices), one for the forms ($\boldsymbol{C}$) and one for the meanings ($\boldsymbol{S}$). For comprehension, the model learns a mapping from the forms in $\boldsymbol{C}$ to the meanings in $\boldsymbol{S}$. For production, it learns the reverse mapping, from $\boldsymbol{S}$ to $\boldsymbol{C}$. To illustrate how the model works, consider the toy lexicon in Table 3, which lists three words with the lexeme *pukul* ('to hit'), and selected semantic features.

Lexeme	Word	Voice	Transitivity	Object Semantic Role	Mood
pukul	pukul	ACTIVE	TRANSITIVE	PATIENT	IMPERATIVE
pukul	memukul	ACTIVE	TRANSITIVE	PATIENT	
pukul	dipukul	PASSIVE	TRANSITIVE	PATIENT	

Table 3: An example lexicon with three words (*pukul, memukul, dipukul*) and selected semantic features.

As a first step, we consider how to represent the words' forms numerically. Here, there are many options (see Heitmeier et al., 2021, for detailed discussion), in what follows, we make use vectors that specify which syllable pairs, henceforth disyllables, jointly define a word's form. The presence of a disyllable is indicated by 1, its absence by 0. The form matrix $\boldsymbol{C}$ can now be set up as follows, with words on rows and disyllables on columns:

$$\boldsymbol{C} = \begin{array}{c} \\ memukul \\ dipukul \\ pukul \end{array} \begin{array}{c} \begin{array}{cccccccc} \#me & memu & mukul & kul\# & \#di & dipu & pukul & \#pu \end{array} \\ \begin{pmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 & 1 \end{pmatrix} \end{array}, \tag{1}$$

The # symbol is used to denote word boundaries, thus, the disyllable *#me* denotes the word-initial syllable *me*.

Words' meanings are represented by real-valued vectors, as in distributional semantics (see, e.g., Landauer and Dumais, 1997; Mikolov et al., 2013). The central idea behind distributional semantics goes back to Firth (1957), who argued that "You shall know a word by the company it keeps" (p. 11) An important first implementation of this idea was developed by Landauer and Dumais (1997): they showed how words can be represented as points in a high-dimensional semantic space. Distributional semantics is now widely used in natural language processing.

Numeric representations for words' meanings (also known as word embeddings) are typically derived from large corpora. In what follows, we make use of simulated semantic vectors that are carefully constructed to reflect words' semantic features (see, e.g., Baayen et al., 2018; Chuang et al., 2020b). In section 4, we will illustrate the use of corpus-based semantic vectors.

For the lexicon shown in Table 3, as a first step, the model generates numeric vectors consisting of random numbers sampled from a normal distribution for the lexeme and for all feature values (ACTIVE, PASSIVE, TRANSITIVE, PATIENT OBJECT). Then, the semantic vectors for the words are constructed by adding the semantic vectors of the lexeme and the words' semantic feature values. Thus, for *memukul*, the semantic vector is obtained as follows:

$$\overrightarrow{\text{MEMUKUL}} = \overrightarrow{\text{PUKUL}} + \overrightarrow{\text{ACTIVE}} + \overrightarrow{\text{TRANSITIVE}} + \overrightarrow{\text{PATIENT}} \tag{2}$$

The semantic vectors for the three words are brought together in the semantic matrix $\boldsymbol{S}$, which for this example has 8 dimensions.

$$\boldsymbol{S} = \begin{array}{c} \\ \textit{pukul} \\ \textit{memukul} \\ \textit{dipukul} \end{array} \begin{array}{c} \begin{array}{cccccccc} \text{S1} & \text{S2} & \text{S3} & \text{S4} & \text{S5} & \text{S6} & \text{S7} & \text{S8} \end{array} \\ \begin{pmatrix} 4.991 & 8.156 & 3.846 & 0.694 & 0.565 & 2.617 & 3.068 & 3.013 \\ 5.640 & 2.467 & 2.313 & 0.938 & 0.296 & 2.068 & 1.809 & 3.064 \\ 3.836 & 4.675 & 4.242 & 2.639 & 0.914 & 5.352 & 5.226 & 3.175 \end{pmatrix} \end{array} \tag{3}$$

Now that the form matrix $\boldsymbol{C}$ matrix and the semantic matrix $\boldsymbol{S}$ matrix have been defined, we can estimate a mapping $\boldsymbol{F}$ from form to meaning by solving

$$\boldsymbol{CF} = \boldsymbol{S}, \tag{4}$$

where $\boldsymbol{CF}$ denotes the matrix multiplication of $\boldsymbol{C}$ and $\boldsymbol{F}$, and likewise a mapping $\boldsymbol{G}$ from meaning to form by solving

$$\boldsymbol{SG} = \boldsymbol{C}. \tag{5}$$

Formally, these mappings can be understood as simple networks with connections from form units (here, disyllables) to semantic dimensions, or, equivalently, as the beta weights of a multivariate multiple regression model. Mathematically, of all possible mappings between form and meaning, the linear mappings $\boldsymbol{F}$ and $\boldsymbol{G}$ are the simplest.

Once the comprehension mapping $\boldsymbol{F}$ and the production mapping $\boldsymbol{G}$ have been estimated, we can use these mappings to predict meaning vectors from words' forms, and form vectors from words' meanings:

$$\begin{aligned} \boldsymbol{CF} &= \hat{\boldsymbol{S}} \\ \boldsymbol{SG} &= \hat{\boldsymbol{C}} \end{aligned} \tag{6}$$

Here, we borrow notation from statistics: $\hat{\boldsymbol{S}}$ denotes the matrix of semantic vectors that the model predicts. The predicted vectors typically are not identical to the gold standard semantic vectors in matrix $\boldsymbol{S}$, however, if the model is reasonably accurate, then they should show strong correlations with the gold standard vectors. To assess comprehension accuracy for a given word ω_i, we calculate the correlations of its predicted semantic vector $\hat{\boldsymbol{s}}_i$ (the i-th row of matrix $\hat{\boldsymbol{S}}$) and correlate it with all gold standard vectors, resulting in a vector of correlations $\boldsymbol{r}_i$. We take the model to correctly understand a word if the correlation $r(\hat{\boldsymbol{s}}_i, \boldsymbol{s}_i)$ is the highest of all correlations in $\boldsymbol{r}_i$.

For the modeling of speech production, an extra step is needed. The reason is that for word i, the predicted form vector $\hat{\boldsymbol{c}}_i$ (the i-th row vector of $\hat{\boldsymbol{C}}$) only specifies the amount of support that disyllables

Semantic feature	Values
Concreteness	CONCRETE; ABSTRACT
Animacy	ANIMATE; INANIMATE
Voice	ACTIVE; PASSIVE
Transitivity	TRANSITIVE; INTRANSITIVE
SubjectSemanticRole	AGENT; INSTRUMENT; AGENT-INSTRUMENT; PATIENT; PROFESSIONAL; LOCATION; CAUSER
ObjectSemanticRole	GOAL; PATIENT OBJECT; PLACE; RECIPIENT; RECIPIENT, PLACE; TOOL; THEME; THEME, BENEFICIARY
Manner	ACTION; APPLICATIVE; CAUSATIVE; DISTRIBUTIVE MANNER; INTENSITY; ITERATIVE; LOCATIVE; RANDOM ACTION; RECIPROCAL; REFLECTIVE; REPETITIVE
Aspect	PERFECTIVE; IMPERFECTIVE; CONDITION; PROCESS; RESULT
Volition	ABILITATIVE; UNINTENTIONAL
State	POSSESSION; SHARED POSSESSION; REGULARITY; STATIVE
Degree	COMPARATIVE; INTENSIVE DEGREE; SUPERLATIVE
Gradation	GRADUAL; NON GRADUAL
ChangeOfObject	CHANGE OF FORM; CHANGE OF INSTRUMENT USED CHANGE OF LOCATION; CHANGE OF STATE
Simplified redup meaning	PROXIMITY; PLURALITY; DISTRIBUTIVE; CASUAL; IMITATIVE; INDEFINITE SPECIFITY; RECIPROCATIVE; INTENSIVE; REPETITIVE REDUP MEANING
Mood	EMPHATIC; IMPERATIVE; POLITE IMPERATIVE; QUESTION
BaseRelationship	TO GIVE X; TO HAVE CHARACTER TRAIT X; TO PRODUCE X; TO USE X
PronounPerson	FIRST; SECOND; THIRD
PronounFunction	SUBJECT; OBJECT; POSSESSIVE
NyaFunction	NYADEFINITEDETERMINER; NYASUBJECT; NYAOBJECT; NYAPOSSESSIVE

Table 4: Semantic features and feature values in the modeling dataset.

receive from the semantics, but leaves the ordering of the diphones unspecified. The model therefore first uses an algorithm to generate candidate sequences of disyllables, and then ranks the resulting candidate word forms by how well their meanings (calculated with the comprehension mapping $\boldsymbol{F}$) correspond to the semantic vector $\boldsymbol{s}_i$ (the i-the row vector of $\boldsymbol{S}$) that is targeted for production. The model is judged to correctly produce a word if the candidate it selects corresponds to a word's actual form. The model is implemented in Julia, and available as a Julia package under the name **JudiLing** (Luo et al., 2021), documentation is available at `https://megamindhenry.github.io/JudiLing.jl/stable/`.

3.2 Materials

The materials for this case study comprise 3102 complex words, for 99 different lexemes. The number of complex words for a lexeme varied from 6 to 78, with a median equal to 29. Compounds were not included.

The words were selected from an Indonesian corpus that is part of the Leipzig Corpora Collection, available at `http://corpora2.informatik.uni-leipzig.de/download.html`(accessed on August 2016), which consists of 7,964,109 different word types and 1,206,281,985 word tokens (Goldhahn et al., 2012). Initially, we selected 2517 Indonesian non-reduplicated words which were enriched with semantic information as described in **?**. We then expanded their dataset with 585 forms with reduplication (13 adjectives, 37 adverbs, 376 nouns, and 159 verbs). The reduplications are derived from 95 mono-morphemic words (31 adjectives, 1 adverb, 33 nouns, and 30 verbs). Reduplication was restricted to full reduplication and affix-reduplication.

The words in this dataset were annotated for the semantic features and feature values listed in Table 4. All annotations were checked by the first author against examples of use in the corpus. For any given word form, only a subset of the features is relevant. For instance, the prefix *meN-* generally creates active-transitive verbs. Thus, the verb *menembak* 'to shoot' is specified for the content lexome *tembak* and for the feature values `active`, `transitive`, and `theme`. The prefix *di-* indicates the passive. So, the word *ditembak* 'to be

shot' is specified as having the lexomes `passive`, `transitive`, and `theme`. Here, we used the term `theme` to describe the direct object that is not undergoing a change of state, following Soekarno (2010). By contrast, the suffix *-i* takes a `patient` as its object.

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). Examples are given in (9).

(9) CHANGE OF STATE

−*besar* 'big', *besarkan* 'to make something bigger',
dia membesarkan baju itu 'she/he makes that shirt bigger (imperative)'

−*jelas* 'clear', *jelaskan* 'to make something clearer',
jelaskan maksudmu 'make your intention clearer (imperative)'

CHANGE OF INSTRUMENT USED

−*pukul* 'to hit (by hand)',
pukulkan tongkat itu ke meja 'hit that table (by stick)'

CHANGE OF FORM

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

−*masalah* 'problem',
dia mempermasalahkan hak asuh anak 'he/she makes the custody into a problem'

CHANGE OF LOCATION

−*turun* 'go down',
turunkan meja itu 'get that table down (imperative)'

−*datang* 'to come',
datangkan Bapak Presiden Jokowi 'make Mr. President Jokowi come (imperative)'

The values of the feature for `Simplified redup meaning` approximate the semantics of reduplication. Following Sugerman (2016); Chaer (2008); Rafferty (2002); Dalrymple and Mofu (2011), we distinguished the feature values listed in (10).

(10) PLURAL (*petembak* 'shooter (athlete)' - *petembak-petembak* 'shooters')

PROXIMITY (*terakhir* 'the end' - *terakhir-terakhir* 'almost the end')

DISTRIBUTIVE (*berat* 'heavy' - *sulit-sulit* 'things are equally heavy')

CASUAL (*jalan* 'road' - *berjalan-jalan* 'to walk casually')

IMITATIVE (*kuat* 'strong' - *menguat-nguatkan* 'pretending to be strong')

INDEFINITE SPECIFITY (*lari* 'to run' - *berlari-larian* 'to run in a random direction')

RECIPROCATIVE (*pukul* 'to hit' - *berpukul-pukulan* 'to hit each other')

INTENSIVE (*senang* 'happy' - *sesenang-senangnya* 'as happy as possible')

REPETITIVE (*tembak* 'to shoot' - *menembak-nembak* 'to shoot repeatedly')

The feature `BaseRelationship` specifies how the meaning of the base word functions within the derived word, examples for the possible values are listed in (11).

(11) to GIVE THE OBJECT designated by the base word
korban 'sacrifice' - *berkorban* 'to give a sacrifice'

to HAVE A CHARACTERISTIC PROPERTY expressed by the base word
keras 'hard' - *berkeras* "to have a strong belief about something'

to PRODUCE THE OBJECT denoted by the base
musik 'music' - *bermusik* 'to produce music', *suara* 'voice' - *bersuara* 'to make a sound'

Word	English Translation	Animacy	Concreteness	Voice	Transitivity	Object Semantic Role
menembak	to shoot			ACTIVE	TRANSITIVE	THEME
ditembak	to be shot			PASSIVE	TRANSITIVE	THEME
penembak	shooter	ANIMATE,INANIMATE	CONCRETE			
petembak	shooter (athlete)	ANIMATE	CONCRETE			
petembakmu	your shooter (athlete)	ANIMATE	CONCRETE			
nembak	(cranberry form)			ACTIVE	TRANSITIVE	THEME
nembakkan	(cranberry form)			ACTIVE	TRANSITIVE	TOOL
tembak	to shoot			ACTIVE	TRANSITIVE	THEME
penembak-penembak	shooters	ANIMATE,INANIMATE	CONCRETE			
petembak-petembakmu	your shooters (athlete)	ANIMATE	CONCRETE			
menembak-nembak	to shoot repeatedly			ACTIVE	TRANSITIVE	THEME
menembak-nembakkan	to shoot something repeatedly			ACTIVE	TRANSITIVE	TOOL
ditembak-tembak	to be shot repeatedly			PASSIVE	TRANSITIVE	THEME

Word	Semantic Role	Change Of Object	Pronoun Person	Pronoun Function	Number	Simplified Redup Meaning
menembak						
ditembak						
penembak	AGENT-INSTRUMENT					
petembak	AGENT					
petembakmu	AGENT		SECOND	POSSESSIVE		
nembak						REPETITIVE REDUP MEANING
nembakkan		INSTRUMENT USED				REPETITIVE REDUP MEANING
tembak						
penembak-penembak	AGENT-INSTRUMENT				PLURAL	PLURALITY
petembak-petembakmu	AGENT		SECOND	POSSESSIVE	PLURAL	PLURALITY
menembak-nembak						REPETITIVE REDUP MEANING
menembak-nembakkan		INSTRUMENT USED				REPETITIVE REDUP MEANING
ditembak-tembak						REPETITIVE REDUP MEANING

Table 5: Words sharing the base lexeme *tembak* 'to shoot', with their semantic features.

to USE THE OBJECT expressed by the base word
layar 'sail' - *berlayar* 'to sail'

Table 5 provides, by way of example, the annotation in our dataset for words which have *tembak*, 'to shoot' as base lexeme. The total number of distinct feature values in the dataset is 92. As explained above, the semantic vector for a given word is obtained by summation of the vectors of all feature values and the vector of the word's lexeme. The way in which the semantic vectors of feature values and lexemes are generated ensures that they are independent (i.e., orthogonal in semantic space). As a consequence, the model is setting up a semantic space that is maximally differentiated with respect to the semantic feature values. In all likelihood, this is a simplification of the true complexity of the semantic system of Indonesian, in which features values may actually be correlated rather than independent. This is an area of active research, and beyond the scope of this contribution.

On the basis of the above semantic features, we constructed for the 2,517 words in our dataset, a 2517×852 form matrix $\boldsymbol{C}$ and a 2517×852 semantic matrix $\boldsymbol{S}$.

3.3 Model performance

How accurate is the comprehension and production performance of the model?[1]. To answer this question, we first evaluated model performance on the full dataset. We constructed the mappings between form and meaning for all words, and then examined how well the model had learned. For comprehension, accuracy was at 98.6%. Production accuracy was slightly lower, at 96.3%. Of the 112 production errors, 71 concerned forms with reduplication. Clearly, for the model, reduplication is the most difficult to learn. Given the many different semantic functions that are realized through reduplication (see (10)), this is perhaps unsurprising.

Training and evaluating the model on the same data is informative about how well the model performs as a memory. What we do not know is how productive this memory is. To assess the productivity of the model, we split the data in two parts, one part comprising 90% of the data that we used for training, and one part (10% of the data) that we set aside for the evaluation of how well the model understands and produces unseen, novel words. The held-out data was constructed such that the words always contained lexemes, semantic feature values, and disyllables that had been encountered already in the training data. For comprehension, accuracy on the training data was 98.2%, and accuracy on the test data was 97.0%. For production, accuracy on the training data was also good, at 91.1%, but accuracy on the held-out data was lower, at 72.7%. Words with reduplication were the most difficult to learn: only 61% of the words with reduplication were produced correctly for the training data, and only 43% for the held-out data. By contrast, the corresponding percentages for words without reduplication were 99% and 79%.

Examples of the errors made by the model for the training and held-out data are given in Table 6. For the training data, the model mostly fails to reduplicate, and for one non-reduplicative form it omits the clitic. For the held-out data, affix and clitic errors are more common, and in this particular set of examples, reduplication is absent for only one form.

Table 6: Examples of production errors for training and held out data.

training data		held-out data	
observed	predicted	observed	predicted
a-da-kan-lah	a-da-kan	ak-hi-ran	ak-hi-ran-nya
a-ja-ran-a-ja-ran	a-ja-ran	ba-kar	mem-ba-kar
a-ja-ran-a-ja-ran-ku	a-ja-ran-ku	be-nar-be-nar	be-nar
a-ja-ran-a-ja-ran-mu	a-ja-ran-mu	be-nar-nya	be-nar
a-ja-ran-a-ja-ran-nya	a-ja-ran-nya	be-re-nang-nya	be-re-nang-ku
a-jar-kan	dia-jar-kan	be-ru-sa-ha-lah	be-ru-sa-ha

In summary, the model is an excellent memory for both comprehension and production, and it shows good productivity for comprehension, but reduced productivity for production, and specifically for words with reduplication.

[1]Code for setting up and running the model, as well as the dataset, is available in the supplementary materials at `https://osf.io/bqzgc/`

3.4 Discussion

The model's reduced production performance for words with reduplication raises several questions. First, is our dataset too small and too sparse, so that the model does not have enough experience during training with the wide variety of meanings that are realized through reduplication? Given that training on the full dataset, without holding back data for testing, yields much better results, also for words with reduplication, this is a real possibility. But it should be kept in mind that a sample of only 3000 words comes with inherent limitations for computational modeling: results can either fall short due to data sparsity, or they can be too flattering.

Second, is the way we coded the semantics possibly too simplistic? In the next section, we will show that when corpus-based semantic vectors are used, words with reduplication cluster in a way that is not straightforwardly predicted from our semantic features. Although we took great care to construct linguistically informed semantic vectors, the current implementation of the model is set up such that the feature values ANIMATE, INANIMATE, AGENT, PATIENT are all completely unrelated, whereas one may expect that in reality, the vectors or ANIMATE and INANIMATE are more similar to each other, and similarly AGENT and PATIENT. Furthermore, the latter feature values might have to be more similar to ANIMATE than to INANIMATE.

Third, could it be the case that reduplication in Indonesian is actually less productive than the other kinds of word formation? The model may be telling us that the mapping from meaning to form is the most difficult to learn specifically for reduplication. Given that typologically, reduplication is a common word formation mechanism (Robino, 2013), it is unlikely that the process of reduplication itself is out of reach of the linear mappings of the DL model. Instead, the diversity of the semantics realized with reduplication are the most likely reason for a reduction in productivity for production.

The model shows an asymmetry between production and comprehension accuracy that is similar to the asymmetry observed for human language learning, where comprehension also has an advantage over production (see, e.g., Chuang et al., 2020a). For instance, receptive vocabularies tend to be much larger than the sets of words individual speakers use themselves. The present study suggests that likewise, productivity in comprehension can be ahead of productivity in production. Given that words have been found to have their own specific usage patterns (Bybee, 2001; Sinclair, 1991; Hawkins, 2003), it is conceivable that the words that we use are slowly learned over multiple exposures (Bybee, 2010), first by listening, and later also in production. The increase in production accuracy from held-out data (72.7%) to training data (91.1%), and then to the full dataset (96.3%) is compatible with this view.

This perspective on the mental lexicon is very different from a realizational perspective working with bundles of semantic features that are realized by sequences of stems and exponents (see, e.g., Stump, 2001). This type of realizational model can be formalized using finite-state morphology (Karttunen, 1993; Beesley and Karttunen, 2000; Beesly and Kartunnen, 2003; Mistica et al., 2009), but getting this approach to work for affix-reduplication requires careful hand-crafting of lexical representations with mark-up specifying boundaries for for reduplication. For affix-reduplication, Mistica et al. (2009) introduce a further feature specifying how the affix should be reduplicated. No such linguistic engineering is required for the DL model.

In this case study, we did not include imitative reduplication, as it is said to be unproductive. To test the DL model more stringently, words with imitative reduplication can also be added into the dataset. The prediction here is that, due to the unpredictability of the form changes in imitative reduplication, the model will learn these words well, but will show very little productivity on held-out data (see Heitmeier et al., 2021, for modeling semi-productive morphology).

An important feature of this 'discriminative lexicon' for Indonesian is that the full morphology (to the extent that it is represented in our dataset) is captured by the production and comprehension mappings. There are no specific rules (or mappings) for individual affixes, and the issue of affixes being multiply ambiguous does not arise. Furthermore, the question of whether circumfixation is a process of its own, or instead just repeated affixation, does not arise, and the same holds for whether an exponent is inflectional or derivational.

However, it is possible to set up the model differently, for instance with one set of mappings for nouns, and another set of mappings for verbs. Such a modeling strategy would receive some support from evidence that verbs and nouns are subserved by different brain areas (see, e.g., Damasio and Tranel, 1993; Laudanna and Voghera, 2002). Furthermore, mappings can also be set up separately for non-reduplicated words on

the one hand, and for reduplicated words on the other hand (see, e.g., Denistia, 2020, for DL simulation on Indonesian non-reduplicated words).

In summary, the complex derivational morphology of Indonesian, including reduplication, appears to be within the scope of what the simple linear mappings of the DL model can handle. Obviously, the model developed above is explorative in nature, and it remains an open question whether the model will properly scale up to a realistically sized lexicon. Furthermore, the DL model aims to capture subliminal learning, and is blind to conscious and deliberate learning strategies that may also be part of our lexical knowledge. Nevertheless, the model seems promising as a 'statistical' tool that, by integrating distributional semantics with morphology, makes it possible to detect which parts of the morphology system are especially frail.

In this section, we have studied the Indonesian lexicon holistically, as a system. In the next section, we zoom in on two similar and yet distinct prefixes, *pe-* and peN-.

4 A case study on the Indonesian *pe-* and *peN-*

The similarity in form of the nominalizing prefixes *pe-* and *peN-*, especially for the cases where both are realized with the exponent *pe-* (see Table 7 for examples), raises the question of whether these prefixes are allomorphs or independent prefixes in their own right. Sneddon et al. (2010) and Ramlan (2009) argue that they are allomorphs, whereas Dardjowidjojo (1983) and Kridalaksana (2007) conclude that there are too many formal and semantic differences between the two to analyse them as allomorphs (see also Baayen et al., 2013, for a similar argument for Russian prefixes).

Noun	Prefix	Noun Translation	Base Word	Base Translation	Base Word Class	Semantic Role
pewawancara	PEN-	interviewer	wawancara	interview	n	agent
perintis	PEN-	pioneer	rintis	pioneer	n	agent
peminta	PEN-	demander	minta	to ask for	v	agent
pelukis	PEN-	painter	lukis	to paint	v	agent
pewisata	PE-	traveller	wisata	to travel	v	agent
perunding	PE-	who are in discussion	runding	discussion	n	agent
pemusik	PE-	musician	musik	music	n	agent
pelari	PE-	runner	lari	to run	v	agent

Table 7: Examples of *pe-* and *peN-* realized with the same exponent *pe-*.

Denistia and Baayen (2019) investigated the status of *pe-* and *peN-* using methods from corpus linguistics. They collected all word types with these prefixes that occur in the Indonesian corpus in the Leipzig corpora collection (Goldhahn et al., 2012), using the MorphInd parser (Larasati et al., 2011), followed by manual checking against the online version of the *Kamus Besar Bahasa Indonesia* dictionary (`http://kbbi.kemdikbud.go.id`, accessed on June 2016) (Alwi, 2012). These words are included in the dataset that was used in the previous section to model part of the Indonesian mental lexicon.

The DL model has no real problems with these prefixes. For the full dataset, it gets two words (out of 329) with *peN-* wrong (one of which involves reduplication), and three words (out of 182) with *pe-* are produced incorrectly (all three with reduplication). Within this approach, the question of whether the two prefixes are allomorphs does not arise. However, the question concerning the status of *pe-* as a potential allomorph of *peN-* can be restated as a question concerning in what way *pe-* differs from *peN-*.

4.1 Differences in form

First consider the way their forms differ. *peN-* shows nasal allomorphy, *pe-* does not. For example, when *peN-* attaches to the base word *suruh* 'to give command', the resulting derived word is *penyuruh* 'who gives command', but when *pe-* attaches to the same base word, the nominalization is *pesuruh* 'who is commanded'. Interestingly, it is possible that even in contexts where *peN-* is realized with the same segments as *pe-*, the two prefixes have different phonetic properties. Recent research on English word-final /s/ indicates that the acoustic duration of the /s/ varies with the semantic features that it realizes (Plag et al., 2017). A follow-up computational modeling study (Tomaschek et al., 2019) suggests that the observed durational differences reflect differences in the learnability of form-meaning mappings. If *pe-* and *peN-* realize different semantics, then it is conceivable that this has consequences for the way in which they are actually pronounced.

4.2 Differences in meaning

Do *peN-* and *pe-* differ with respect to the meanings that they realize? The corpus-based survey of Denistia and Baayen (2019) indicates that *peN-* is more often used to realize instruments than is the case for *PE-*. In fact, using a quantitative measure for productivity that estimates the probability of observing new forms (using the ratio $\mathcal{P}$ of hapax-legomena and the number of tokens Baayen, 2009), it can be shown that *PE-* is not productive at all for instruments, and primarily realizes agents. Conversely, *peN-* is reasonably productive for instruments. Within the set of words realizing agents, *PE-* is specialized for professional sports persons, an observation made by Chaer (2008) and confirmed, using distributional semantics, by Denistia et al. (2021).

The power of corpus-based semantic vectors for understanding semantic differences between *PE-* and *peN-* is illustrated in Figure 1. The semantic vectors used here were created with *fasttext* (Bojanowski et al., 2017) and were downloaded from `https://dl.fbaipublicfiles.com/fasttext/vectors-crawl/cc.id.300.vec.gz`. The words that we analyzed here are nominalizations with *PE-* and *peN-* as well as complex words in which these nominalizations occur, available in a database compiled by the first author and available at `https://osf.io/bqzgc/`. From this database, which contains 2938 words, we selected the 1672 words for which a *fasttext* semantic vector was available. These words were all bare nouns, without clitics, but their reduplicated counterparts were included.

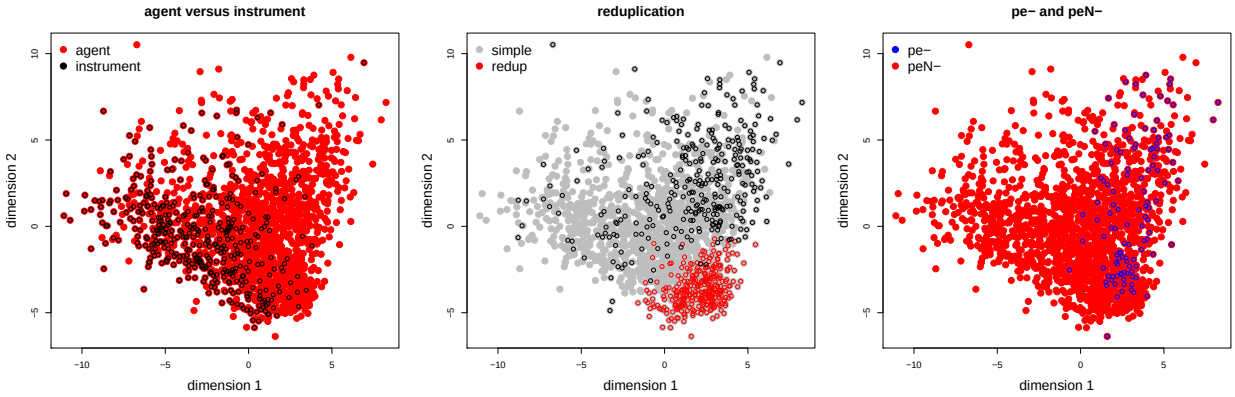

Figure 1: Projection of 300-dimensional *fasttext* vectors for words with *pe-* and *peN-* onto a two-dimensional plane, using principal components analysis. In the center panel, the black circles represent the non-reduplicated counterparts of the forms with reduplication, shown in red.

Each of the three panels of Figure 1 presents the 1672 words in the same two-dimensional plane, obtained by using principal components analysis to visualize 300-dimensional semantic vectors. Each panel highlights different properties of the words. The left panel highlights agents and instruments; the center panel highlights simple words (gray), words with reduplication (red), and their non-reduplicated counterparts (black); and the right panel presents words with *pe-* in blue and words with *peN-* in red.

The left panel shows that the first dimension distinguishes between instruments, which tend to have large negative values, from agents, which tend to have positive values. The third panel clarifies that *pe-* and *peN-* differentiate on the same dimension: words with *pe-* appear at the far right.

The center panel of Figure 1 shows that the second dimension separates words with reduplication (large negative values) from words without reduplication (which tend to have less negative, and predominantly large positive values). The non-reduplicated counterparts of the words with reduplication are highlighted in black. Apparently, there are marked differences in how reduplicated and non-reduplicated words are used, which argues against the claim that there is no distinction in meaning between the two, as argued by Dalrymple and Mofu (2011). A comparison of the first two panels reveals that reduplicated nouns are less likely to be instruments: the reduplicated words are positioned more to the right on the first dimension. However, a challenge for further research is how to understand the second dimension. This dimension may be profiling the commonalities of the many different semantic features realized by reduplication (see Table 4), but to some extent it is also differentiating between *pe-* words and *peN-* words: the latter occur more often with

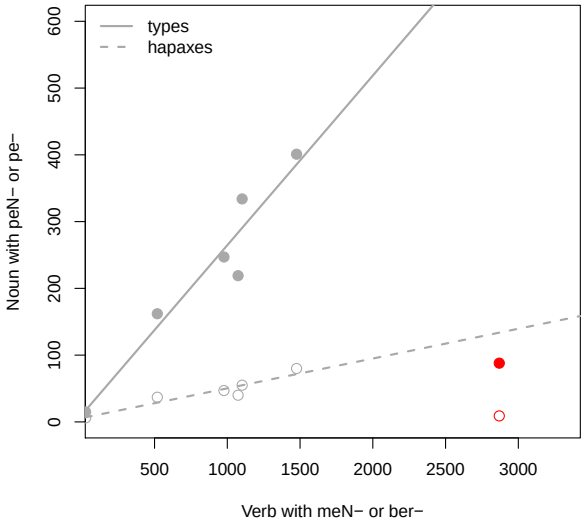

Figure 2: Counts of types for verbs with *meN-* and *ber-* (horizontal axis) and counts of types and hapaxes for *pe-* and *peN-* allomorphs (vertical axis). *pe-* (red points) is an outlier with respect to the regression lines to the *peN-* allomorphs (grey points) for both counts of types (filled circles) and counts of hapax legomena (open circles).

extreme negative values on the second dimension.

This analysis with *fasttext* semantic vectors illustrates the potential of distributional semantics for the study of word structure. The semantic vectors constructed through simulation in the modeling study are useful as a first step, but these vectors miss out on the true complexity and richness of words' actual meanings. For modeling studies using the DL approach and building on corpus-based semantic vectors, see Baayen et al. (2019), Heitmeier and Baayen (2020), and Shafaei-Bajestan et al. (2021).

The model laid out in the previous section has a further disadvantage that has not been mentioned yet, namely, that it is not sensitive to frequency of occurrence. The reason is that the way the mappings are estimated provides the best solution for infinite training on the data. The model can therefore be thought of as providing a window on the 'endstate' of learning, and with infinite experience, differences in frequency vanish (see Heitmeier et al., 2021, for detailed discussion). It is possible to learn the mappings step-by-step, using the update rules of Rescorla and Wagner (1972) and Widrow and Hoff (1960), but doing this in a computationally efficient way is a topic of ongoing research. A complication that arises in the context of incremental learning is that we, as we learn, not only learn new words, but also refine the meanings of the words we already know. Therefore, in principle, the semantic vectors of words also need to be updated as experience unfolds.

4.3 Differences in productivity

Although incremental learning of the Indonesian lexicon is expected to yield useful insights for morphological productivity, in what follows, we use corpus-based productivity measures to highlight yet another difference between *pe-* and *peN-*. This difference is illustrated in Figure 2.

Both axes of Figure 2 denote the number of types (a measure of the extent to which an affix is used) and the number of hapax legomena (the words that are used once, these counts are proportional to the contribution of an affix to the growth rate of the vocabulary). The horizontal axis presents these counts for the base verbs of *peN-* and *pe-*, which carry the prefixes *meN-* and *ber-*. The gray dots represent the allomorphs of *meN-* and *peN-*, the red dots represent *ber-* and *pe-*. The solid and dashed lines are obtained with linear regression models fitted to the *meN-/peN-* allomorphs. The points for *ber-/pe-* are far removed

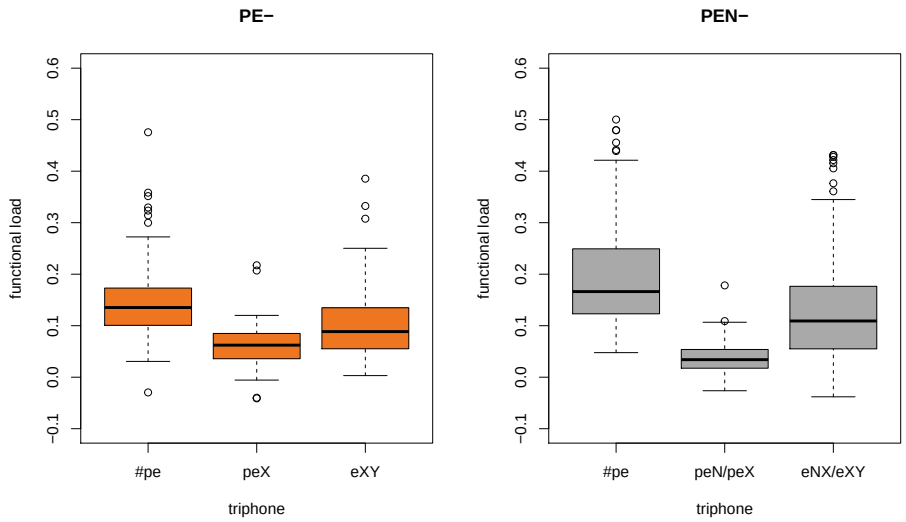

Figure 3: The distributions of functional load of the first three triphones of words with *pe-* (left panel) and *peN-* (right panel).

from these regression lines. This is clear evidence that *pe-* is not simply an allomorph of *peN-*, and that *ber-* is not a variant of *meN-*. Importantly Figure 2 shows that the productivity of complex words' allomorphs reflects the productivity of the allomorphs of their complex base words. Whether this is a property specific to affix substitution is an open question.

Returning to the question of whether *pe-* is an allomorph of *peN-*, our conlusion is that the two prefixes differ in too many ways for it to be theoretically profitable to speak of allomorphy. At the same time, the question of whether allomorphy is involved is a theoretically useful question in the sense that it forces the analyst to carefully consider whether the only difference between the prefixes is one that concerns a predictable alternation in form only.

4.4 Nasal allomorphy and functional load

What are the consequences of the allomorphy of *peN-* and *meN-*, in combination with a separate exponent *pe-*, for learning? Nasal allomorphy facilitates articulation, but does this come at a cost for comprehension? How is it possible that *pe-*, which is less productive that *peN-*, has not become unproductive?

These kinds of questions can be addressed within the framework of the DL model, but for the representation of words' forms, we need more granularity than provided by di-syllables. One possibility, explored by Denistia and Baayen (2021), is to use triphones. The importance of a triphone for understanding a word can be assessed by removing the triphone and calculating to what extent the correlation decreases of the predicted semantic vector and the gold standard semantic vector. The greater this decrease is, the greater the 'functional load' of the triphone is.

Figure 3 presents the distributions of functional loads for the first three triphones of words with *pe-* (left panel), and for the first three triphones of words with *peN-*. For words with *pe-*, the first triphone has a word boundary (*#pe*), the second triphone covers the phones of the exponent and the first phone of the stem (*peX*), and the third triphone consists of the second phone of the exponent and the first two stem phones (*eXY*). For words with *peN-* (right panel), the first triphone is identical to that of *pe-*, the second triphone consists of the nasal allomorph (or, in case of the *pe-* allomorph, of *pe* and the first phone of the stem), and the final triphone consists of the vowel of the exponent, followed by the nasal and the initial phone of the stem (*eNX*), or, for the *pe* allomorph, the two stem-initial phones (*eXY*).

The distributions of functional loads are different for the two prefixes in a theoretically interesting way. For both prefixes, the second triphone has the lowest functional load. For *peN-*, it is, on average, lower than for *pe-*. The first triphone, again for both prefixes, has the highest functional load, but now, on average,

the functional load is higher for *peN-* as compared to *pe-*. The functional load of the third triphone is intermediate, and similar for both prefixes.

To understand this pattern, it is useful to know that there is no difference between the two prefixes with respect to the sum of the functional loads of the first and second triphone. Apparently, the two prefixes achieve a different balance of the same total functional load. Next, it is important to realize that a triphone *peX* is more discriminative than a triphone *peN*. There are many more triphones *peX*, as X can be any stem-initial phone. By contrast, the triphones *peN* have a third segment that is restricted to a small set of nasals. Because *peX* triphones are more specific to individual words, they acquire higher functional loads than *peN* triphones, which are shared by more words. Apparently, the advantage of nasal co-articulation for speech production is offset by the disadvantage of a reduced functional load of the *peN* triphone. Fortunately, this disadvantage is offset by a higher functional load that for *peN-* accrues to the *#pe*. For the less productive 'rival' prefix *pe-*, it is the functional load of the initial triphone that suffers.

We can now address the general questions raised at the beginning of this subsection. It appears that one factor that may be contributing to the continuing productivity of *#pe-* is exactly the nasal allomorphy of its rival prefix, which ensures that the *peX* triphone of *#pe-* remains a reliable cue for words' meanings. With respect to the potential disfunctionality of nasal allomorphy for comprehension, the present analysis suggests that the decrement in the functional load of the *peN* triphone is easily offset by an increased functional load for the initial triphone, even in the presence of a rival suffix.

5 Concluding remarks

Compared to languages such as English or Dutch, the derivational morphology of Indonesian is amazingly rich. Indonesian has large morphological families with on average around 30 derived words, which stand in stark contrast with the handful of derived words in the morphological families of English or Dutch.

In this chapter, we first provided an overview of word formation in Indonesian, using traditional descriptive constructs such as stems and affixes. However, Indonesian is no exception to the polyfunctionality of word formation devices. English [s] realizes third person singular on verbs, plurals for nouns, and in addition genitives and reduced forms of auxilaries. Indonesian reduplication realizes no less than 9 different semantic functions. The many-to-many relations between form and meaning cast doubt on the usefulness of the theoretical construct of the morpheme.

The second part of this study therefore explored the usefulness for understanding Indonesian morphology of a recent computational model for the mental lexicon. This model, which is a computational formalization of Word and Paradigm morphology (Blevins, 2016), does not make use of stems, affixes, and morphemes, but instead considers how numerical representations of words' forms and numerical representations of words' meanings can be mapped onto each other. Trained on a set of some 3,000 words, this model appears to be able to understand and produce familiar words with high accuracy, but it struggles somewhat for production with reduplication. As the model was constructed using very simple representations of form and meaning, there is considerable room for improvement. An important property of our modeling approach is that it is not necessary to hand-engineer lexical representations. Once the analyst has decided on how to represent form and meaning in general, the model is defined and its performance can be tested. For Indonesian, the model suggests that reduplication suffers from frailty in production, which we think may be due to the large number of different meanings that are realized with reduplication. We have discussed several modeling strategies that may improve on the exploratory analysis that we presented. We have also shown how this model can be used to study functional load with a precision that goes beyond that afforded by methods based on minimal pairs (Martinet, 1995; Wedel et al., 2013).

The computational model captures all the word formation devices of Indonesian as one single system. The third part of this chapter zooms in on two rival prefixes that are similar in form, but at the same time show various subtle differences that can be brought to light using methods from corpus linguistics, distributional semantics, and computational modeling. The analyses in this part of our chapter are presented in the hope that they will be useful for the analyses of other Asian languages, and specifically, Austronesian languages.[2].

[2]For a study of productivity using corpus linguistics and distributional semantics, applied to Mandarin Chinese, see Shen and Baayen (2021)

The approach that we pursued for understanding the Indonesian lexicon is inspired by cognitive linguistics, corpus linguistics, usage-based linguistics, and computational modeling. Models, as well as quantitative analyses, necessarily require simplified representations and mappings between these representations, and the analyses in this chapter are no exception to this rule. The analyses and models that we presented are no more than explorations of what should be possible once we have richer resources. On our wish list for Indonesian are:

1. Corpora of both spoken and written varieties of Indonesian, as the ways in which morphology is put to use across registers varies substantially (Plag et al., 1999; Biber, 1988);
2. Multimodal corpora of Indonesian, combining video, audio, gesture annotation, and transcriptions (see, for English, Uhrig, 2017);
3. Grounded embeddings that merge information gleaned from images into corpus-based semantic vectors (see, e.g., Shahmohammadi et al., 2021); and, of course,
4. More sophisticated computational models that, unlike the black boxes of many deep learning algorithms, remain transparent to interpretation. The very simple models that we introduced in this chapter are similar to statistical models: once linguistically motivated representations for form and meaning have been decided on, the mappings are basically multivariate multiple regression models that cannot be tweaked by the analyst. For a tool that hopefully is useful for exploring how sublexical form and distributional meaning interact, this may well be a useful constraining feature.

In other words, for understanding the morphology of Indonesian, of Asian languages, and languages in general, we have only just started to harvest the benefits of the language resources that are now becoming available.

Acknowledgements

This study was funded by the Indonesia Endowment Fund for Education (*Lembaga Pengelola Dana Pendidikan*) to the first author (No. PRJ-1610/LPDP/2015) and by ERC advanced grant 742545 to the second author. We are indebted to Elnaz Shafaei-Bajestan, Maria Heitmeier, and Yu-Ying Chuang for their feedback on earlier versions of this chapter.

References

Alisjahbana, S. T. (1954). *Tata bahasa baru bahasa Indonesia*. Pustaka Rakyat, Jakarta.

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., 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. (2009). Corpus linguistics in morphology: morphological productivity. In Kytö, M. and Lüdeling, A., editors, *Corpus Linguistics. An international handbook.*, pages 900–919. Mouton de Gruyter, Berlin.

Baayen, R. H., Chuang, 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. (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*.

Beesley, K. R. and Karttunen, L. (2000). Finite-state non-concatenative morphotactics. *arXiv preprint cs/0006044*.

Beesly, K. and Kartunnen, L. (2003). Finite state morphology. cslipublications.

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

Biber, D. (1988). *Variation across speech and writing*. Cambridge University Press, Cambridge.

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

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Bybee, J. (2010). *Language, usage and cognition*. Cambridge University Press, Cambridge.

Bybee, J. L. (2001). *Phonology and language use*. Cambridge University Press, Cambridge.

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

Chuang, Y. Y. and Baayen, R. H. (2021). Discriminative learning and the lexicon: NDL and LDL. In Aronoff, M., editor, *Oxford Research Encyclopedia of Linguistics*. Oxford University Press. (in press).

Chuang, Y.-Y., Bell, M., Banke, I., and Baayen, R. H. (2020a). Bilingual and multilingual mental lexicon: A modeling study with linear discriminative learning. *Language Learning*, 71(S1):219–292.

Chuang, Y.-Y., Lõo, K., Blevins, J. P., and Baayen, R. H. (2020b). *Estonian Case Inflection Made Simple: A Case Study in Word and Paradigm Morphology with Linear Discriminative Learning*, page 119–141. Cambridge University Press.

Chung, S. (1976). An object-creating rule in bahasa Indonesia. *Linguistic: Inquiry*, 7:41–87.

Corbett, G. G. (2000). *Number*. Cambridge University Press, Cambridge.

Dalrymple, M. and Mofu, S. (2011). Plural semantics, reduplication, and numeral modification in Indonesian. *Journal of Semantics*, 29(2):229–260.

Damasio, A. R. and Tranel, D. (1993). Nouns and verbs are retrieved with differently distributed neural systems. *Proceedings of the National Academy of Sciences*, 90(11):4957–4960.

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. (2020). *Quantitative studies on the Indonesian prefixes PE-and PEN*. PhD thesis, Universität Tübingen.

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

Denistia, K. and Baayen, R. H. (2021). Affix substitution in Indonesian: A computational modeling approach. To appear in *Linguistics*.

Denistia, K., Shafaei-Bajestan, E., and Baayen, R. H. (2021). Exploring semantic differences between the Indonesian prefixes pe- and pen- using a vector space model. *Corpus Linguistics and Linguistic Theory*. doi:10.1515/cllt-2020-0023.

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

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.

Grangé, P. (2015). The indonesian verbal suffix ânya. *Wahana*, 16(1):133–166.

Greenberg, J. H. (1972). Numeral classifiers and substantival number: Problems in the genesis of a linguistic type. *Working Papers on Language Universals*, 9:1–39.

Hawkins, S. (2003). Roles and representations of systematic fine phonetic detail in speech understanding. *Journal of Phonetics*, 31:373–405.

Heitmeier, M. and Baayen, R. H. (2020). Simulating phonological and semantic impairment of English tense inflection with Linear Discriminative Learning. *The Mental Lexicon*, 15:385–421.

Heitmeier, M., Chuang, Y.-Y., and Baayen, R. H. (2021). Modeling morphology with linear discriminative learning: considerations and design choices. *arXiv preprint arXiv:2106.07936*.

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

Himmelmann, N. P. (2005). *The Austronesian Languages of Asia and Madagascar*, chapter The Austronesian Languages of Asia and Madagascar: Typological Characteristics. Routledge, London and Newyork.

Karttunen, L. (1993). Finite-state constraints. pages 173–194.

Karttunen, L. (2003). Computing with realizational morphology. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 203–214. Springer.

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.

Kridalaksana, H. (2012). *Pembentukan kata dalam bahasa Indonesia*. Gramedia Pustaka Utama, Jakarta.

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.

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.

Lander, Y. A. (2003). Nominal reduplication in Indonesian challenging the theory of grammatical change. In *The 7th International Symposium on Malay/Indonesian Linguistics: The Book of Papers*, Nijmegen. IIAS.

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.

Laudanna, A. and Voghera, M. (2002). Nouns and verbs as grammatical classes in the lexicon. *Italian journal of linguistics*, 14:9–26.

Levin, T. and Polinsky, M. (2021). *The Oxford encyclopedia of morphology*, chapter Morphology in Austronesian Languages. Oxford University Press, Oxford.

Luo, X., Chuang, Y. Y., and Baayen, R. H. (2021). Judiling: an implementation in Julia of Linear Discriminative Learning algorithms for language modeling.

Martinet, A. (1995). *Économie des changements phonétiques*. Francke, Berne.

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

Mistica, M., Andrews, A., Arka, I. W., and Baldwin, T. (2009). Double double, morphology and trouble: Looking into reduplication in indonesian. In *Proceedings of the 2009 Australasian Language Technology Association Workshop*, pages 44—52, Sydney, Australia.

Mistica, M., Andrews, A., Arka, I. W., and Baldwin, T. (2012). Handling Indonesian clitics: A dataset comparison for an Indonesian-English statistical machine translation system. In *26th Pacific Asia Conference on Language,Information and Computation*, pages 146–152.

Muhadjir (1981). *Morphology of Jakarta dialect, affixation, and reduplication*. Badan Penyelenggara Seri Nusa, Jakarta.

Muslich, M. (2009). *Tata Bentuk Bahasa Indonesia; Kajian ke arah deskriptif*. Bumi Aksara, Jakarta Timur.

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.

Nomoto, H., Choi, H.-G., Moeljadi, D., and Bond, F. (2018). MALINDO morph: Morphological dictionary and analyser for Malay/Indonesian.

Ogloblin, A. K. (1998). *Typology of Verbal Categories*, chapter From inert to actional causative, pages 235–256. De Gruyter, Berlin, Boston.

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

Plag, I., Dalton-Puffer, C., and Baayen, R. H. (1999). Morphological productivity across speech and writing. *English Language and Linguistics*, 3(2):209–228.

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.

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.

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.

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

Rescorla, R. A. and Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In Black, A. H. and Prokasy, W. F., editors, *Classical conditioning II: Current research and theory*, pages 64–99. Appleton Century Crofts, New York.

Robino, C. (2013). Reduplication. In Dryer, M. and Haspelmath, M., editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology.

Samuel, J. (2009). Potensialitas dan keterbatasan inovasi morfologis dalam bahasa Indonesia. *Wacana*, 11(2):294–318.

Sato, Y. and McDonnell, B. (2013). Reduplication in Indonesian and the lexicalist hypothesis. In *Proceedings of the Thirty-Third Annual Meeting of the Berkeley Linguistics Society*, pages 365–372, University of California, Berkeley. Department of Linguistics [with Bradley McDonnell].

Sedeng, I. N. (2015). Syntactical marker -nya in Indonesian. *RETORIKA: Jurnal Ilmu Bahasa*, 1(2):258–278.

Shafaei-Bajestan, E., Tari, M. M., P., U., and Baayen, R. H. (2021). LDL-AURIS: Error-driven learning in modeling spoken word recognition. *Language, Cognition and Neuroscience*. `https://www.tandfonline.com/doi/full/10.1080/23273798.2021.1954207`.

Shahmohammadi, H., Lensch, H., and Baayen, R. H. (2021). Learning zero-shot multifaceted visually grounded word embeddings via multi-task training. *arXiv preprint arXiv:2104.07500*.

Shen, T. and Baayen, R. H. (2021). Adjective-noun compounds in Mandarin: A study on productivity. *Corpus Linguistics and Linguistic Theory*. `https://www.degruyter.com/document/doi/10.1515/cllt-2020-0059/html`.

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford University Press.

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.

Stump, G. (2001). *Inflectional Morphology: A Theory of Paradigm Structure*. Cambridge University Press.

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.

Uhrig, P. (2017). Newsscape and the Distributed Little Red Hen Lab – a digital infrastructure for the large-scale analysis of TV broadcasts. *Anne-Julia Zwierlein, Jochen Petzold, KB and Decker, M., editors, Anglistentag*, pages 99–114.

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(2):179–186.

Widrow, B. and Hoff, M. E. (1960). Adaptive switching circuits. *1960 WESCON Convention Record Part IV*, pages 96–104.