

An overview of corpus linguistics and its application to form-meaning relationship in Indonesian voice-morphological constructions

Gede Primahadi Wijaya Rajeg

Bachelor of English, Faculty of Humanities, Universitas Udayana, Indonesia

Keynote presentation at the Linguistics Master's Program of Universitas Sebelas Maret, Indonesia

Wednesday, 25 August 2021

 <https://orcid.org/0000-0002-2047-8621>

 @PrimahadiWijaya

Outlines

- Defining “corpus” and other key concepts
- Main analytical tools in corpus linguistics
 - concordance/keyword in context (KWIC)
 - frequency list
 - collocation
- Applications
 - Form-meaning relationship in Indonesian voice-morphological constructions – case study with *kena-i* & *kena-kan*

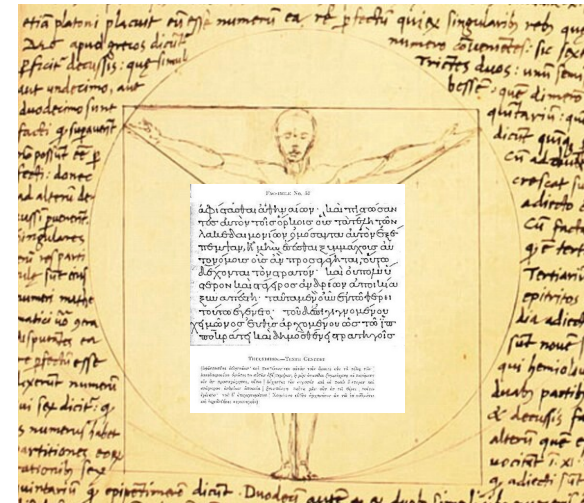
Defining “corpus”

What is a “corpus”?

- Latin word for ‘body’ (Baker 2010: 93)
 - The plural is **corpora**
 - A body of texts

https://en.wikipedia.org/wiki/Manuscript#/media/File:Thucydides_Manuscript.jpg

https://en.wikipedia.org/wiki/Vitruvian_Man#/media/File:Vitruvian_Man_by_Giacomo_Andrea.jpg



What is a “corpus”?

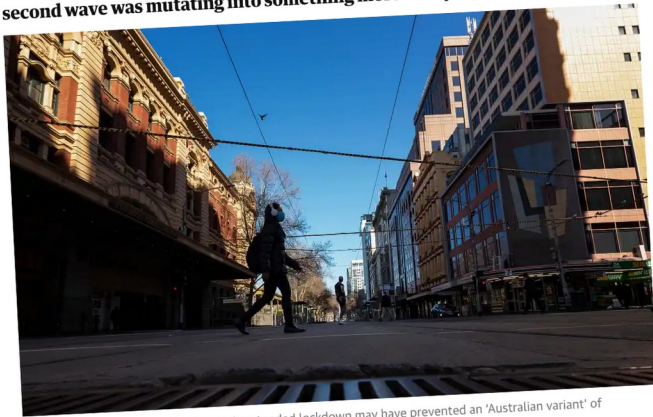
- Latin word for ‘body’ (Baker 2010: 93)
 - The plural is **corpora**
 - A body of texts
- In corpus linguistics: “a set of **texts** in **computer-readable** form” (Wray & Bloomer 2006: 196)

Baker, Paul. 2010. Corpus Methods in Linguistics. In Lia Litosseliti (ed.), *Research methods in linguistics (Research Methods in Linguistics)*, 93–113. London & New York: Continuum.

Wray, Alison & Aileen Bloomer. 2006. *Projects in Linguistics: A Practical Guide to Researching Language*. 2nd ed. London & New York: Hodder Arnold & Distributed in the United States of America by Oxford University Press.

'Dodged a bullet': Melbourne lockdown may have prevented more deadly Covid-19 variant

Researchers say the variant that swept Victoria during last year's second wave was mutating into something more worrying



▲ A leading virologist says Melbourne's extended lockdown may have prevented an 'Australian variant' of coronavirus. Photograph: Daniel Pockett/Getty Images

A variant of Covid-19 similar to the one that spread rampantly in the UK would likely have developed in Victoria during last year's second wave had **Melbourne** not gone into an extended lockdown, a leading virologist says.

Associate Prof Stuart Turville from the Kirby Institute at the University of New South Wales said when his laboratory examined samples **from patients** as part of a study called "ADAPT" in Sydney, they started to see key differences in those infected with the virus during the second wave.

<https://www.theguardian.com/world/2021/jan/29/dodged-a-bullet-melbourne-lockdown-may-have-prevented-more-deadly-covid-19-variant>



<https://twitter.com/AcademicsSay/status/98234145334623>

Oliver Twist by Charles Dickens



Download This eBook

Format	Size
Read this book online: HTML	1.1
EPUB (no images)	47
Kindle (no images)	1.1
Plain Text UTF-8	93
More Files...	

Similar Books

🔍 Readers also downloaded...

<https://www.gutenberg.org/ebooks/730>

Captain America: The First Avenger



Previous transcript:

[Thor](#)

Next transcript:

[The Avengers](#)

Previous transcript:

[Captain America: The Winter](#)

Next transcript:

[first lines; in the Arctic]

Search Team Leader: Are you the guys from Washington?

SHIELD Tech: You get many other visitors out here?

SHIELD Lieutenant: How long have you been on site?

Search Team Leader: Since this morning. A Russian oil team called it in about 18 hours

SHIELD Lieutenant: How come nobody spotted it before?

https://transcripts.fandom.com/wiki/Captain_America:_The_First_Avenger

What is a “corpus”?

- **Large-scale** textual data
- Difficult to read, search, and manipulate by hand and eye which guarantees no errors

What is a “corpus”?

- **Large-scale** textual data
- Difficult to read, search, and manipulate by hand and eye which guarantees no errors
- **Exploited with computer tools** for rapid & reliable search through the corpus

What is a “corpus”?

- Representing language produced in any mode:
 - corpora of (transcribed) **spoken** language
 - corpora of **written** language
 - **Audiovisual** corpora that also capture paralinguistic features:
 - gestures
 - corpora of **signed** language

Defining “corpus linguistics”

Corpus Linguistics

- “[T]he **analysis** of (usually) very **large** collections of electronically stored **texts**, aided by **computer software**” (Baker 2010: 93)
 - Characterised as a “**methodology**” (McEnery & Wilson 2001: 1)
 - Not a traditional branch of linguistics (e.g. semantics, grammar, phonetics, or sociolinguistics)

- Baker, Paul. 2010. Corpus Methods in Linguistics. In Lia Litosseliti (ed.), *Research methods in linguistics (Research Methods in Linguistics)*, 93–113. London & New York: Continuum.
- McEnery, Tony & Andrew Wilson. 2001. *Corpus linguistics: An introduction (Edinburgh Textbooks in Empirical Linguistics)*. 2. ed., repr. Edinburgh: Edinburgh Univ. Press.

Corpus Linguistics

- Empirical (i.e. data-based), inductive form of analysis
- Relying on **real-world** instances of **language use**
 - Can act **as control/yardstick to** model of language that rely on **artificial linguistic data** (usually via introspection)
- Deriving rules, or exploring trends, about how people **actually** produce and use language

Corpus Linguistics

<< providing **access to quantitative data** >>



<https://boostlabs.com/wp-content/uploads/2019/09/10-types-of-data-visualization-1.jpg>

Corpus Linguistics

Enables researchers to test hypotheses/theories about language from a new perspective

Allows researchers to raise new questions and theories about language impossible otherwise

Outlines

- ~~Defining “corpus” and other key concepts~~
- Main analytical tools in corpus linguistics
 - concordance/keyword in context (KWIC)
 - frequency list
 - collocation
- Applications
 - Form-meaning relationship in Indonesian voice-morphological constructions – case study with *kena-i* & *kena-kan*

Main analytical tools in corpus linguistics

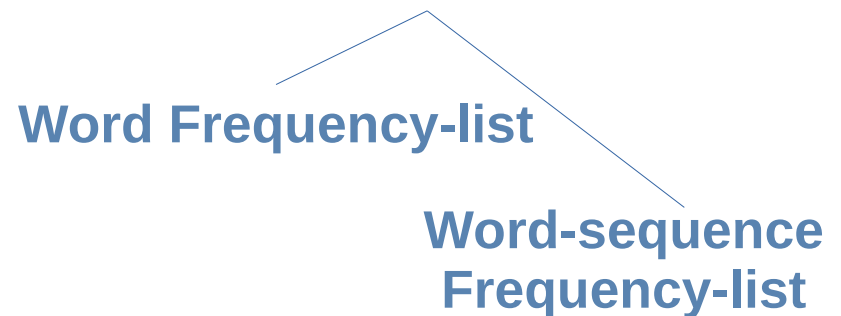
**Concordances/Key
word in context
(KWIC)**

Collocates Tables

Frequency Tables

Word Frequency-list

**Word-sequence
Frequency-list**



Main analytical tools in corpus linguistics

**Concordances/Key
word in context
(KWIC)**

Collocates Tables

Frequency Tables

Word Frequency-list

Word-sequence
Frequency-list

Concordance/Key Word in Context (KWIC)

- Examples of **word/phrases** and **their contexts**
 - Possible to sort them (e.g., alphabetically)
- Difficult to do by hand, unless with corpus-linguistic softwares (e.g., *AntConc* [Anthony 2019])
 - *AntConc* tutorial in Indonesian (Rajeg 2020)

AntConc 3.5.8 (Macintosh OS X) 2019

Corpus Files

- ACE_A.TXT
- ACE_B.TXT
- ACE_C.TXT
- ACE_D.TXT
- ACE_E.TXT
- ACE_F.TXT
- ACE_G.TXT
- ACE_H.TXT
- ACE_J.TXT
- ACE_K.TXT
- ACE_L.TXT
- ACE_M.TXT
- ACE_N.TXT
- ACE_P.TXT
- ACE_R.TXT
- ACE_S.TXT
- ACE_W.TXT

Total No. 17
Files Processed

Concordance Concordance Plot File View Clusters/N-Grams Collocates Word List Keyword List

Concordance Hits 146

Hit	KWIC	File
1	that someone who is just doing a job would not have the same commitment to	ACE_A.TXT
2	office, where Les Jones had found a job of sorts, to the barbed wire beyond	ACE_S.TXT
3	starving because a man cannot get a job. <bl> MARY HILL Lakemba, NSW</bl> <h>	ACE_B.TXT
4	an incentive to go and get a job, especially when the children reach school age.)	ACE_B.TXT
5	the prospect of having to get a job. Mr Watson was becoming worried by this	ACE_K.TXT
6	White. Modernism meets modernity...and gets a job. In June 1984 a painting by Charles Sheeler,	ACE_J.TXT
7	, from the initial routine of getting a job in Pacific Data Central to the final	ACE_M.TXT
8	us left. Everyone else had got a job. I said, "What else is there?" "Well,	ACE_S.TXT
9	. In London soon afterwards, I had a job finding the old Fred, but eventually met	ACE_R.TXT
10	'd like me to find her a job. That's what I'm here for -	ACE_L.TXT
11	, and you will be out of a job.' Grandfather was not an unbeliever. In his	ACE_D.TXT
12	end of Stage 3 he was offered a job by Phil Drioni, who was at that	ACE_E.TXT
13	ried through, no longer automatically provide a job for life in which the deadheads can	ACE_B.TXT
14	the road who has just seen a job advertised for a nurse. She would love	ACE_G.TXT
15	we felt was open to us. A job in a small country town, sadly six	ACE_F.TXT
16	announcer after seven years". "I want a job where I get home at 4 o'clock	ACE_A.TXT
17	Alice Springs. Linda makes a believ+able job of playing widow Kate Hannon, the woman	ACE_C.TXT
18	ng to achieve economic recovery and abundant job op+ortunities is linked to the increasing	ACE_G.TXT
19	sent. She was transferred to an administrative job in the Education Department's regional office.	ACE_A.TXT
20	success of his films drew attention and job offers. They taught him the pricelessness of	ACE_C.TXT

Search Term Words Case Regex Search Window Size 50

job Advanced

Start Stop Sort Show Every Nth Row 1

Kwic Sort

Level 1 1L Level 2 2L Level 3 1R

Clone Results

Data from the *Australian Corpus of English*; concordance display via *AntConc* software

<< Identify usage-pattern for the node word >>

Concordance/Key Word in Context (KWIC)

that someone who is just **doing a job** would not have the same commitment to office, where Les Jones had **found a job** of sorts, to the barbed wire beyond starving because a man cannot **get a job**. <bl> MARY HILL Lakemba, NSW</bl> <h> an incentive to go and **get a job**, especially when the children reach school age.) the prospect of having to **get a job**. Mr Watson was bec -White. Modernism meets modernity...and **gets a job**. In June 1984 a paint , from the initial routine of **getting a job** in Pacific Data Centra us left. Everyone else had **got a job**. I said, "What else is t . In London soon afterwards, I **had a job** finding the old Fred, b 'd like me to find **her a job**. That's what I'm here , and you will be out **of a job**.' Grandfather was not an unbeliever. In his end of Stage 3 he was **offered a job** by Phil Drioni, who was at that arried through, no longer automatically **provide a job** for life in which the deadheads can the road who has **just seen a job** advertised for a nurse. She would love we felt was open to **us. A job** in a small country town, sadly six announcer after seven years" "I **want a job** where I get home at 4 o'clock

VERB a job
|
GET a job

Main analytical tools in corpus linguistics

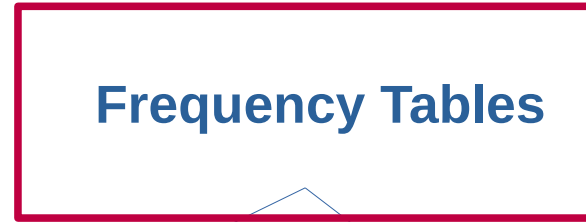
Concordances/Key
word in context
(KWIC)

Collocates Tables

Frequency Tables

Word Frequency-list

Word-sequence
Frequency-list



The role of **frequency** in language – a brief overview

- An important concept in (**usage-based, cognitive**) **linguistics** (cf. Bybee 2010)
- A major factor in **language change**
 - Irregular verbs such as *speak-spoke-spoken* **resists regularisation** (of past tense verb) due to their high token-frequency (Lindquist 2009)
 - **Grammaticalisation**: semantic bleaching or generalisation (Bybee 2010)
 - *BE going to V* and *GET to V* acquires more grammatical meanings of ‘future’ and ‘obligation’ respectively
- May have impact on the **strength of cognitive representation and productivity** of constructions (Bybee 2010)
 - **Entrenchment** of *that drives me crazy* in AmE (due to **high token-frequency**)
 - **Productivity** of *that drive me ADJ* cxn (**high type-frequency** of the ADJ fillers)

Word frequency list

AntConc 3.5.8 (Macintosh OS X) 2019

Corpus Files

- ACE_A.TXT
- ACE_B.TXT
- ACE_C.TXT
- ACE_D.TXT
- ACE_E.TXT
- ACE_F.TXT
- ACE_G.TXT
- ACE_H.TXT
- ACE_J.TXT
- ACE_K.TXT
- ACE_L.TXT
- ACE_M.TXT
- ACE_N.TXT
- ACE_P.TXT
- ACE_R.TXT
- ACE_S.TXT
- ACE_W.TXT

Concordance Concordance Plot File View Clusters/N-Grams Collocates Word List Keyword List

Word Types: 39795 Word Tokens: 786783 Search Hits: 0

Rank	Freq	Word
1	50758	the
2	25693	of
3	21877	and
4	19595	to
5	18305	a
6	16035	in
7	7893	is
8	7404	for
9	7356	that
10	6611	it
11	6518	was
12	5397	with
13	5390	on
14	5308	as
15	5302	s
16	5281	i
17	4643	be
18	4602	he
19	4502	by
20	3865	at

Search Term Words Case Regex

Hit Location Advanced Search Only 0

Lemma List Loaded

Word List Loaded

Sort by Invert Order

Sort by Freq

Total No. Files Processed: 17

- **Word types** and their **token-frequencies** (i.e., how many times a given word-type occur) in the corpus
 - “**the**” occurs **50,758** times in *ACE*

Data from the *Australian Corpus of English*; frequency-list display via *AntConc* software

Word frequency list

AntConc 3.5.8 (Macintosh OS X) 2019

Corpus Files

- ACE_A.TXT
- ACE_B.TXT
- ACE_C.TXT
- ACE_D.TXT
- ACE_E.TXT
- ACE_F.TXT
- ACE_G.TXT
- ACE_H.TXT
- ACE_J.TXT
- ACE_K.TXT
- ACE_L.TXT
- ACE_M.TXT
- ACE_N.TXT
- ACE_P.TXT
- ACE_R.TXT
- ACE_S.TXT
- ACE_W.TXT

Word Types: 39795 Word Tokens: 786783 Search Hits: 0

Rank	Freq	Word
1	50758	the
2	25693	of
3	21877	and
4	19595	to
5	18305	a
6	16035	in
7	7893	is
8	7404	for
9	7356	that
10	6611	it
11	6518	was
12	5397	with
13	5390	on
14	5308	as
15	5302	s
16	5281	i
17	4643	be
18	4602	he
19	4502	by
20	3865	at

Word Tokens: 786,783

Size of the *ACE* corpus (measured in word-tokens)

Total Sum of the **Freq** column

Search Term Words Case Regex

Hit Location Search Only

Lemma List Loaded

Word List Loaded

Sort by Invert Order

Sort by Freq

Clone Results

Total No. 17 Files Processed

Data from the *Australian Corpus of English*; frequency-list display via *AntConc* software

Word frequency list

AntConc 3.5.8 (Macintosh OS X) 2019

Corpus Files

- ACE_A.TXT
- ACE_B.TXT
- ACE_C.TXT
- ACE_D.TXT
- ACE_E.TXT
- ACE_F.TXT
- ACE_G.TXT
- ACE_H.TXT
- ACE_J.TXT
- ACE_K.TXT
- ACE_L.TXT
- ACE_M.TXT
- ACE_N.TXT
- ACE_P.TXT
- ACE_R.TXT
- ACE_S.TXT
- ACE_W.TXT

Concordance Concordance Plot File View Clusters/N-Grams Collocates Word List Keyword List

Word Types: 39795 Word Tokens: 786783 Search Hits: 0

Rank	Freq	Word
1	50758	the
2	25693	of
3	21877	and
4	19595	to
5	18305	a
6	16035	in
7	7893	is
8	7404	for
9	7356	that
10	6611	it
11	6518	was
12	5397	with
13	5390	on
14	5308	as
15	5302	s
16	5281	i
17	4643	be
18	4602	he
19	4502	by
20	3865	at

Word Types: 39,795

Size of the *ACE* corpus (measured in word-types)

Number of different word-types

Search Term Words Case Regex Hit Location Search Only 0

Advanced

Start Stop Sort

Sort by Invert Order Sort by Freq

Lemma List Loaded Word List Loaded Clone Results

Total No. 17 Files Processed

Data from the *Australian Corpus of English*; frequency-list display via *AntConc* software

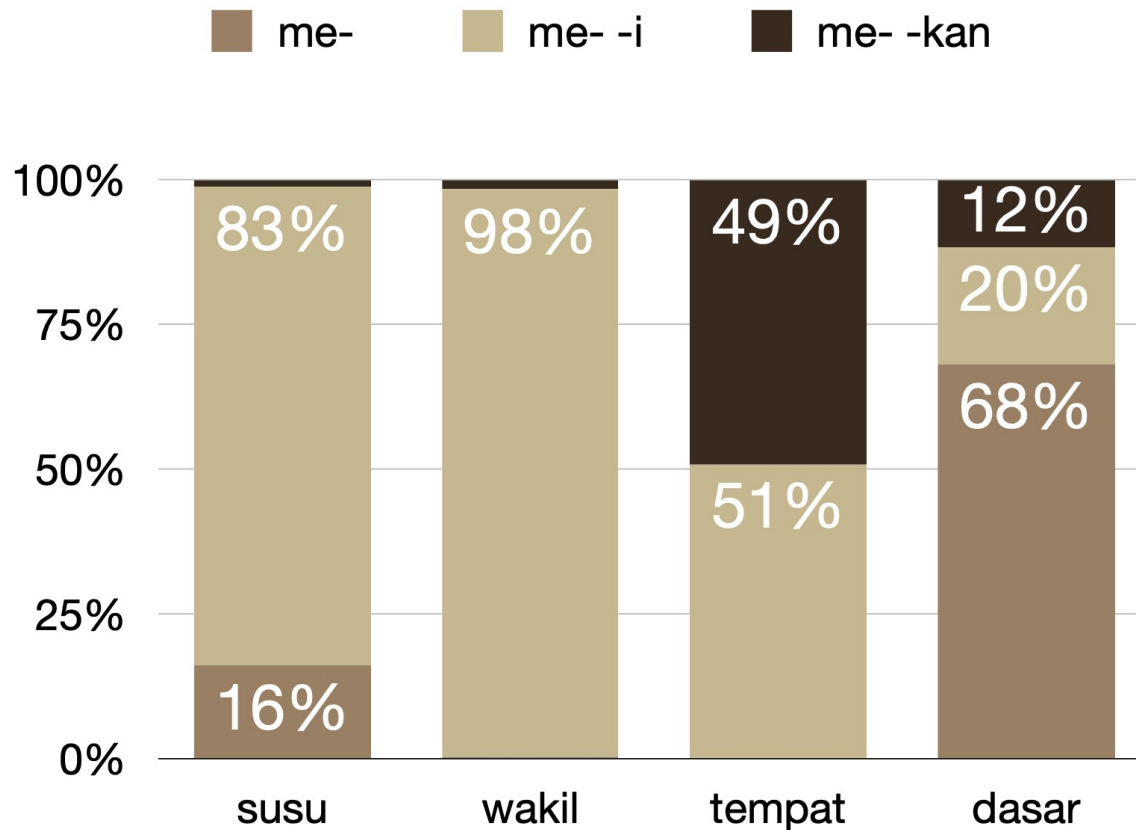
Word frequency list

- The basis for calculating:
 - collocational strength of words with a node/target word
 - keywords in a given target corpus (in comparison to the reference corpus)
 - all these use some forms of statistical significance tests (cf. Gries 2009; 2010)

- Gries, Stefan Th. 2009. *Statistics for linguistics with R: A practical introduction*. Berlin: Mouton de Gruyter.
- Gries, Stefan Th. 2010. Useful statistics for corpus linguistics. In Aquilino Sánchez & Moisés Almela (eds.), *A mosaic of corpus linguistics: selected approaches*, 269–291. Frankfurt am Main: Peter Lang.

Word frequency list

- Morphological profiles of a base word



Word-sequence frequency list

Word-sequence frequency-list

AntConc 3.5.8 (Macintosh OS X) 2019

Concordance Concordance Plot File View **Clusters/N-Grams** Collocates Word List Keyword List

Total No. of N-Gram Types 18911 Total No. of N-Gram Tokens 100245

Rank	Freq	Range	N-gram
1	293	17	one of the
2	151	10	a number of
3	149	16	some of the
4	145	14	as well as
5	139	17	the end of
6	137	16	part of the
7	128	16	i don t
8	126	17	out of the
9	120	17	it was a
10	115	14	there is a
11	110	16	to be a
12	107	16	a lot of
13	104	10	new south wales
14	104	17	there was a
15	103	11	the use of
16	102	13	it is a
17	97	13	there is no
18	90	15	it would be
19	89	8	the number of

Search Term Words Case Regex N-Grams

N-Gram Size Min. 3 Max. 3

Min. Freq. 3 Min. Range 1

Sort by Invert Order On Left On Right

Start Stop Sort

Clone Results

Total No. 17 Files Processed

All three-word-sequences in the *ACE* corpus

Word-sequence frequency-list

AntConc 3.5.8 (Macintosh OS X) 2019

Concordance Concordance Plot File View Clusters/N-Grams Collocates Word List Keyword List

Total No. of Cluster Types 51 Total No. of Cluster Tokens 208

Rank	Freq	Range	Cluster
1	33	9	social and
2	25	7	social security
3	11	5	social justice
4	8	3	social welfare
5	7	1	social relations
6	6	2	social democracy
7	6	4	social issues
8	5	3	social democratic
9	4	3	social change
10	4	1	social contract
11	4	4	social life
12	4	2	social order
13	4	1	social partnership
14	4	4	social services
15	3	1	social alternatives
16	3	2	social history
17	3	3	social policy
18	3	2	social practices
19	3	2	social realism

Search Term Words Case Regex N-Grams

social Advanced Cluster Size Min. 2 Max. 2

Start Stop Sort

Sort by Invert Order Search Term Position Min. Freq. 2 Min. Range 1

Sort by Freq On Left On Right

Clone Results

Total No. 17 Files Processed

Two-word-
sequences
based on the
node word *social*

In *AntConc*, it is
called *clusters*

Main analytical tools in corpus linguistics

Concordances/Key
word in context
(KWIC)

Collocates Tables

Frequency Tables

Word Frequency-list

Word-sequence
Frequency-list

Collocation & Collocate

- Collocation:

“actual words in **habitual company**” (Firth 1957: 14)

Collocation & Collocate

- Collocation:

“actual words in **habitual company**” (Firth 1957: 14)

“the **phenomenon** surrounding the fact that **certain words are more likely to occur in combination with other words** in certain contexts.” (Baker et al. 2006: 36)

Collocation & Collocate

- Collocation:

“actual words in **habitual company**” (Firth 1957: 14)

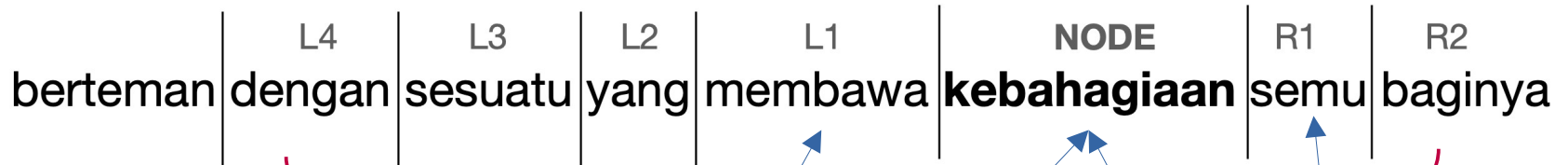
“the **phenomenon** surrounding the fact that **certain words are more likely to occur in combination with other words** in certain contexts.” (Baker et al. 2006: 36)

- Collocate:

“a **word** which occurs within the neighbourhood of another word” (Baker et al. 2006: 36-37)

Syntactic collocation:
verb-subject

linear collocation



Linguistik korpus kuantitatif dan kajian semantik leksikal sinonim emosi bahasa Indonesia. *Linguistik Indonesia* 38(2). 123–150. <https://doi.org/10.26499/li.v38i2.155>.

verb-direct.object
syntactic collocation:

noun-adjective
syntactic collocation:

membesarkan vs. *memperbesar*

Focusing on the one word to the right of the verbs
(R1 collocates – linear collocation)

c29+ mill. tokens of Indonesian Leipzig Corpora
(5 files of the newspapers corpus)

Corpus Files

ind_news_2008_300K-se
 ind_news_2009_300K-se
 ind_news_2010_300K-se
 ind_news_2011_300K-se
 ind_news_2012_300K-se

Concordance Concordance Plot File View Clusters/N-Grams Collocates Word List Keyword List

Total No. of Collocate Types: 21 Total No. of Collocate Tokens: 229

Rank	Freq	Freq(L)	Freq(R)	Stat	Collocate
1	72	0	72	541.01547	partai
2	33	0	33	311.83144	hati
3	29	0	29	172.55814	anak
4	13	0	13	110.42321	pkb
5	10	0	10	88.94921	namanya
6	7	0	7	57.35639	nu
7	6	0	6	46.61582	pan
8	8	0	8	40.83980	nama
9	3	0	3	32.43699	ppi
10	3	0	3	26.88485	dede
11	3	0	3	24.56146	partainya
12	4	0	4	23.83965	pks
13	4	0	4	20.15517	organisasi
14	4	0	4	15.98213	ketiqa
15	3	0	3	10.77529	industri
16	3	0	3	10.50653	enam
17	3	0	3	4.41405	kedua
18	4	0	4	0	pd
19	3	0	3	0	mereka
20	5	0	5	0	dirinya
21	9	0	9	0	dan

Strong collocates for
membesarkan

IMPORTANT IS BIG
conceptual metaphor
(Lakoff & Johnson 1999: 50)

nama besar; hari raya;
partai besar; rakyat kecil;
wong cilik

Lakoff, George & Mark Johnson. 1999. *Philosophy in the flesh: The embodied mind and its challenge to Western thought*. New York: Basic Books.

Search Term Words Case Regex

\b(?:)membesarkan\b

Advanced

Window Span Same

From... 0 To... 1R

Start

Stop

Sort

Min. Collocate Frequency

3

Sort by Invert Order

Sort by Stat

Clone Results

Total No.

5

Files Processed

Corpus Files

Ind_news_2008_300K-se
 Ind_news_2009_300K-se
 Ind_news_2010_300K-se
 Ind_news_2011_300K-se
 Ind_news_2012_300K-se

Concordance Concordance Plot File View Clusters/N-Grams Collocates Word List Keyword List

Total No. of Collocate Types: 24 Total No. of Collocate Tokens: 281

Rank	Freq	Freq(L)	Freq(R)	Stat	Collocate
1	107	0	107	1450.97094	keunggulan
2	46	0	46	402.30159	kemenangan
3	23	0	23	150.68988	gol
4	18	0	18	134.70741	peluang
5	4	0	4	51.91349	keunggulannya
6	5	0	5	48.68492	porsi
7	6	0	6	36.87455	harapan
8	9	0	9	34.27479	jumlah
9	4	0	4	32.44518	selisih
10	5	0	5	31.47010	jarak
11	4	0	4	27.09690	risiko
12	8	0	8	25.88534	pasar
13	4	0	4	24.74586	kekhawatiran
14	4	0	4	24.37497	rekor
15	4	0	4	23.95496	akses
16	4	0	4	23.68716	kapasitas
17	5	0	5	20.39368	kemungkinan
18	3	0	3	19.00962	volume
19	3	0	3	17.24832	skor
20	3	0	3	14.91043	pendapatan
21	3	0	3	13.20241	jaringan
22	3	0	3	10.43497	anggaran

Strong collocates for
memperbesar

Relative/gradable entity (?)

QUANTITY IS SIZE

MORE (OF QUANTITY) IS BIG
(Lakoff & Núñez 2000: 55-56)

*makan besar; diskon gede-
gedean/besar-besaran*

Lakoff, George & Rafael Núñez. 2000. *Where mathematics comes from: How the embodied mind brings mathematics into being*. New York: Basic Books.

Search Term Words Case Regex

\b(?:)memperbesar\b

Start

Stop

Sort

Advanced

Window Span Same

From... 0

To... 1R

Min. Collocate Frequency

3

Sort by Invert Order

Sort by Stat

Total No.

5

Files Processed

Clone Results

Rajeg, Gede Primahadi Wijaya & I Made Rajeg. 2019. Analisis Koleksem Khas dan potensinya untuk kajian kemiripan makna konstruksional dalam Bahasa Indonesia. In I Nengah Sudipa (ed.), *ETIKA BAHASA* Buku persembahan menapaki usia pensiun: I Ketut Tika, vol. 1, 65–83. Denpasar, Bali, Indonesia: Swasta Nulus. <https://doi.org/10.26180/5bf4e49ea1582>. <https://osf.io/preprints/inarxiv/uwzts/> (30 January, 2019).

Outlines

- ~~Defining “corpus” and other key concepts~~
- ~~Main analytical tools in corpus linguistics~~
 - ~~concordance/keyword in context (KWIC)~~
 - ~~frequency list~~
 - ~~collocation~~
- Applications
 - Form-meaning relationship in Indonesian voice-morphological constructions – case study with *kena-i* & *kena-kan*

Form-meaning relationship in Indonesian voice-morphological constructions

- **Pemahaman kuantitatif dasar dan penerapannya dalam mengkaji keterkaitan antara bentuk dan makna** (G. P. W. Rajeg & I M. Rajeg 2019) - *Linguistik Indonesia* (OA paper, data, & R codes)
 - metaphorical vs. literal meanings of morphologically related words based on the root *panas*
- **Corpus-based approach meets LFG: The puzzling case of voice alternations of *kena*-verbs in Indonesian** (G.P.W. Rajeg, I M. Rajeg & I W. Arka 2020) - *LFG'20 Proceedings* (OA paper, data, & R codes)
 - association of senses of *kenai* vs. *kenakan* in Active & Passive constructions (cxns)
- **Corpus linguistic and experimental studies on meaning-preserving hypothesis in Indonesian voice alternation** (I M. Rajeg, G. P. W. Rajeg & I W. Arka to appear) - *Linguistics Vanguard* (OA paper, data, & R codes)
 - association of senses of *majukan*, *ajukan*, *mundurkan*, *undur(kan)* in Active & Passive cxns in the corpus and in the mind

Form-meaning relationship in Indonesian voice-morphological constructions

- Voice alternation (active-passive)
- “Meaning-preserving alternation” (Kroeger 2005: 271)

“meaning is essentially the same” (in active and passive clauses based on the same verb) – “they describe the same kind of event, and it would be impossible for one to be true while the other is false.”

Meaning-preserving alternation

1) Murid Go Bie-Pay yang **meng-(k)ena-kan** baju warna hitam (...)

Student NAME REL **AV**-be.hit-CAUS shirt colour black

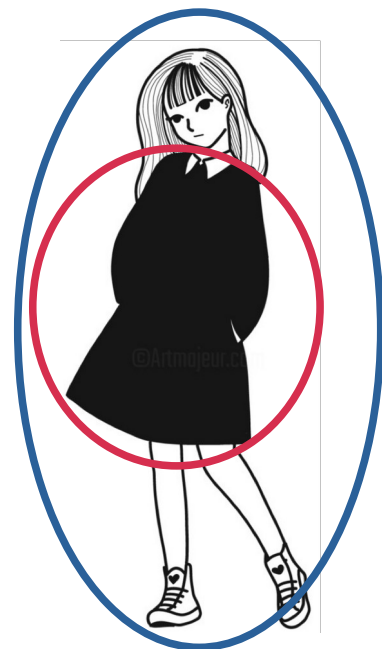
'Go Bie-Pay's student who *wears/puts on* a black shirt (...)

2) Gaun yang **di-kena-kan** ber-warna hitam

dress REL **PASS**-be.hit-CAUS have-colour black

'The dress that is *worn/put on* is black'

The 'wearing' sense of *kenakan* is preserved in AV and PASS forms **IN THESE EXAMPLES**



Form-meaning relationship in Indonesian voice-morphological constructions

- Voice alternation (active-passive)
- “Meaning-preserving alternation” (Kroeger 2005: 271)
 - Implicitly assumed to be applicable to verbal polysemy
 - Any sense expressed in active is also predicted to be preserved in passive
 - No prediction about asymmetric likelihood for the expression of a given sense in a given voice (cf. McDonnell 2016: 243)
 - No prediction for the conventional association of certain sense with certain (voice) type (cf. Bernolet & Coleman 2016)

- Kroeger, Paul. 2005. *Analyzing grammar: An introduction*. Cambridge: Cambridge University Press.
- McDonnell, Bradley. 2016. *Symmetrical voice constructions in Besemah: A usage-based approach*. Santa Barbara, USA: University of California, Santa Barbara PhD dissertation.
- Bernolet, Sarah & Timothy Coleman. 2016. Sense-based and lexeme-based alternation biases in the Dutch dative alternation. In Jiyoung Yoon & Stefan Th. Gries (eds.), *Corpus-based approaches to Construction Grammar*, 165–198. Amsterdam: John Benjamins Publishing Company.

Puzzling behaviour of *kenakan* and *kenai* in AV/PASS

(Rajeg, Rajeg & Arka 2020: 311)

3) Pengusaha **meng-(k)ena-kan/*meng-(k)ena-i** pajak

entrepreneur AV-be.hit-CAUS/AV-be.hit-APPL tax

‘Entrepreneurs *imposes/charges* tax (to their consumers)...’

4) motor kedua akan **di-kena-kan/di-kena-i** pajak sebesar 2 persen

motor second FUT **PASS**-be.hit-CAUS/-APPL tax as.large 2 percent

‘...the second motorbike will be *imposed/subject to/charged with* 2% tax.’

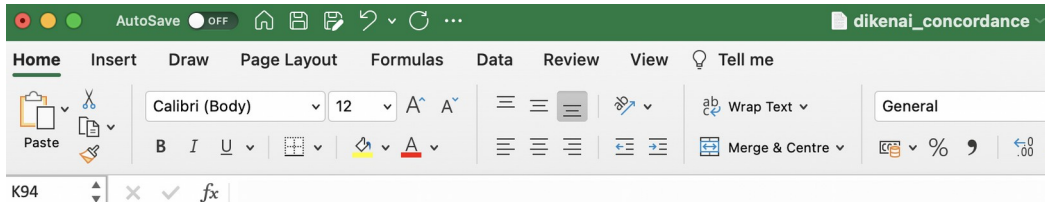
kenakan and *kenai* in PASS *di-* are interchangeable to convey the same sense of ‘imposing’.
Why does the AV form *mengenai* is infelicitious to convey the ‘impose’ sense, unlike its PASS form *dikenai*?

Methodological aspects: **overview**

- Indonesian Leipzig Corpora:
 - One file: ind_mixed_2012_1M-sentences.txt (c15mill. tokens)
- Concordance/KWIC of all tokens of *kenai/kan* in AV and PASS
 - *mengenai* (N=284 tokens) & *dikenai* (N=139)
 - *mengenakan* (N=1,101) & *dikenakan* (N=446)
- Qualitative data analyses of the concordance for each verb in each voice-morphological forms – in MS Excel
 - Manual coding of the senses for each verb
- Quantitative (statistical) analyses on the results of qualitative data analyses – in R

How do we do, and organise, the qualitative analyses (i.e., annotating the senses) of verbs' concordances in MS Excel?

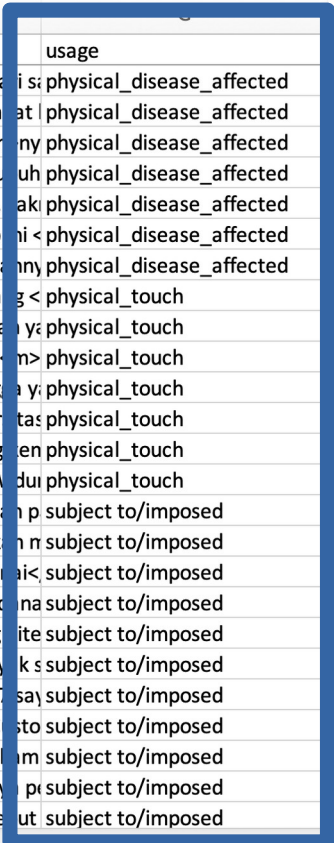
Annotation for **qualitative** analysis of corpus data (i.e., concordance data) in MS Excel/spreadsheet software



	A	B	C	D	E	F	G
1	corpus	sent_id	left	node	right	node_sentences	usage
2	ind_mixed_2012_1M	888090	Manifestasinya tergantung dari saraf yang	dikenai	.	Manifestasinya tergantung dari s	physical_disease_affected
3	ind_mixed_2012_1M	188653	angan adalah tempat lain yang juga lazim	dikenai	.	Punggung tangan adalah tempat	physical_disease_affected
4	ind_mixed_2012_1M	924074	Individu yang	dikenai	nya dapat mencapai usia dewasa m	Individu yang <m>dikenai</m>	physical_disease_affected
5	ind_mixed_2012_1M	74483	, maka yang nampak itu masih akan dapat	dikenai	nya, menyusup di antara lindungan	Asal saja masih ada bagian tu	physical_disease_affected
6	ind_mixed_2012_1M	429566	ari udara, air, tumbuhan dan bakteri serta	dikenai	oleh proses mekanika seperti perut	Weathering zone: zona lapik,	aki
7	ind_mixed_2012_1M	225710	Kalau sisa bakteri yang hidup ini	dikenai	penisilin dari dosis yang sama, mak	Kalau sisa bakteri yang hidup	ni <
8						ngah dari keturunan	physical_disease_affected
9						eberapa orang yang <	physical_touch
10						alah ,Áupenuangan ya	physical_touch
11						ajurit yang telah <m>	physical_touch
12						keadaan Ki Ranga y	physical_touch
13						hews dari Univer	physical_touch
14						g berbasis tabung	physical_touch
15						ka ia melihat Ki W	physical_touch
16						saja, dalam sebulan	subject to/imposed
17						ersebut merupakan	subject to/imposed
18						rtwright <m>dikenai<	subject to/imposed
19						embelian buku dic	subject to/imposed
20						opoli garam yang	ite subject to/imposed
21						n diperlukan bany	k s subject to/imposed
22						da 7 Agustus 2007	sa subject to/imposed
23						aya melapor ke C	sto subject to/imposed
24						biaya yang harus	lm subject to/imposed
25						>dikenai</m> biay	pe subject to/imposed
26	ind_mixed_2012_1M	55466	cepat ketimbang Google Maps dan tanpa	dikenai	biaya roaming untuk data.	Menariknya, akses peta terse	subject to/imposed

The annotated concordance files for all verbs are then imported into R for the statistical, **quantitative** analyses

(cf. my [YouTube tutorial on analysing concordance data in MS Excel](#))



Methodological aspects:

Qualitative, data annotation

- Semantic reference and class of the collocates (e.g., direct object) as guidance for categorising senses (cf. Stefanowitsch 2007)
 - See our paper for examples
- Consult with the online KBBI
- Qualitative, semantic interpretation involved

Methodological aspects:

Quantitative/Statistical analyses

- Bivariate design of quantitative analyses:
 - **FORM** variable (different voice-morphological form of a base verb)
 - **SENSE/MEANING** variable (different senses/meanings evoked by each verb in each voice-morphological form)
- How many times are **sense A, B, C**, etc. expressed by verb **V** in **Active** vs. **Passive forms**?
- Chi-square (or Fisher Exact) significance test
 - Visualisation with barplot and association plot

Methodological aspects: Quantitative/Statistical analyses

FORM variable

	<i>AV:mengenai</i>	<i>PASS:dikenai</i>	
MEANING variable	Sense 1	Freq of sense 1 with <i>AV</i>	Freq of sense 2 with <i>PASS</i>
	Sense 2	Freq of sense 2 with <i>AV</i>	Freq of sense 2 with <i>PASS</i>
	Sense 3	Freq of sense 3 with <i>AV</i>	Freq of sense 3 with <i>PASS</i>
	...		

Bivariate design for the statistical analyses

Methodological aspects: Quantitative/Statistical analyses

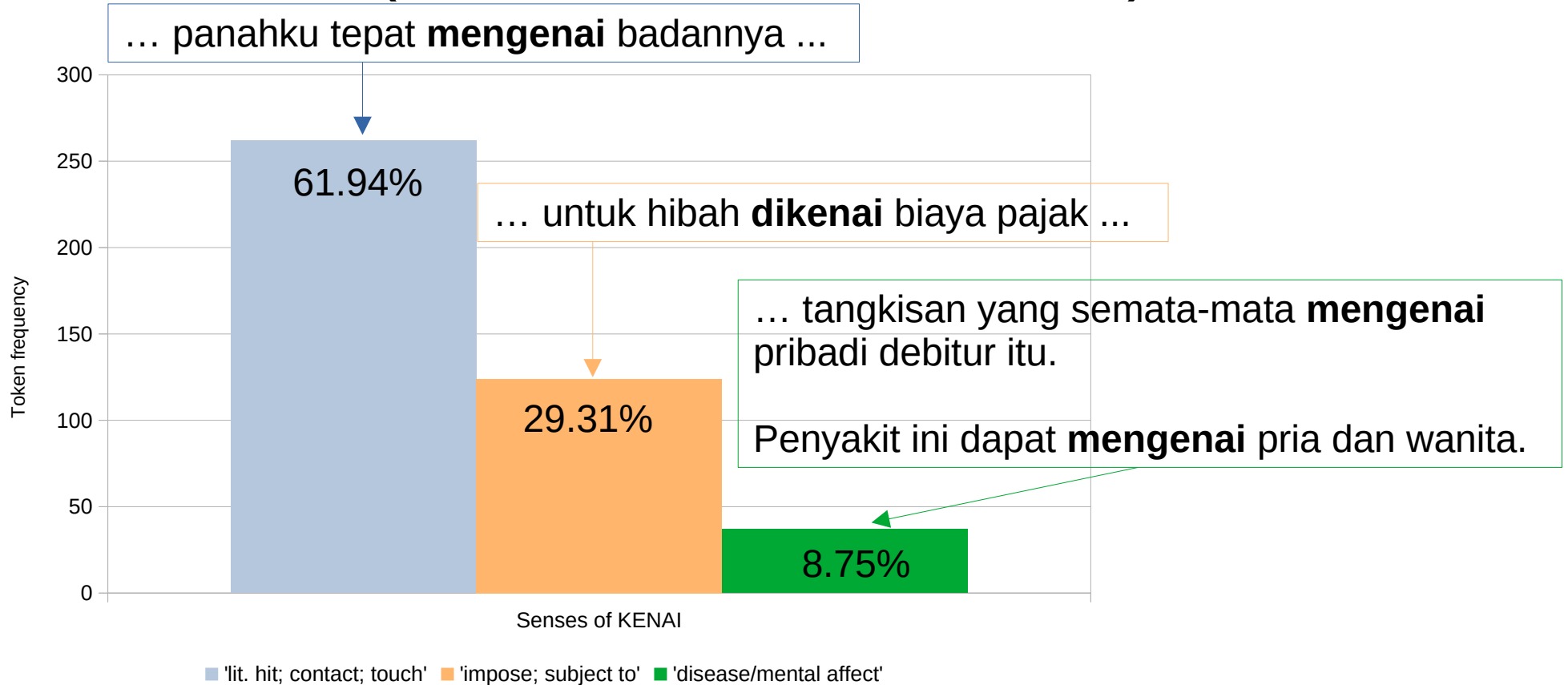
FORM variable

	<i>AV:mengenai</i>	<i>PASS:dikenai</i>
MEANING variable		
come into touch; contact; hit	255	7
subject to; impose	0	124
affect (mental; disease)	29	8

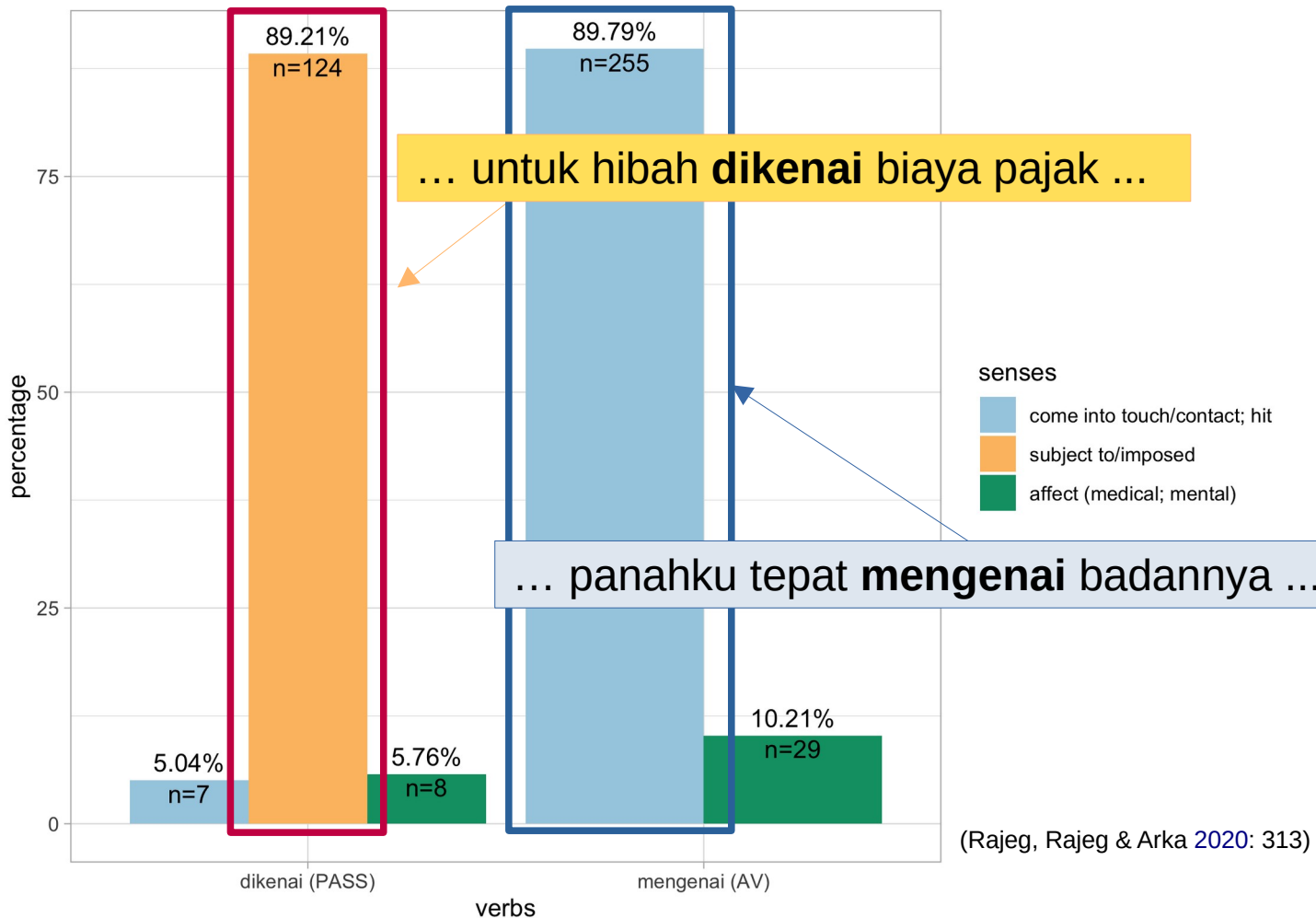
Bivariate design for the statistical analyses

Results for the senses of *kenai* in AV and PASS

Senses of *kenai* (combined in AV & PASS)



Distribution of senses for *kenai* in PASS and AV

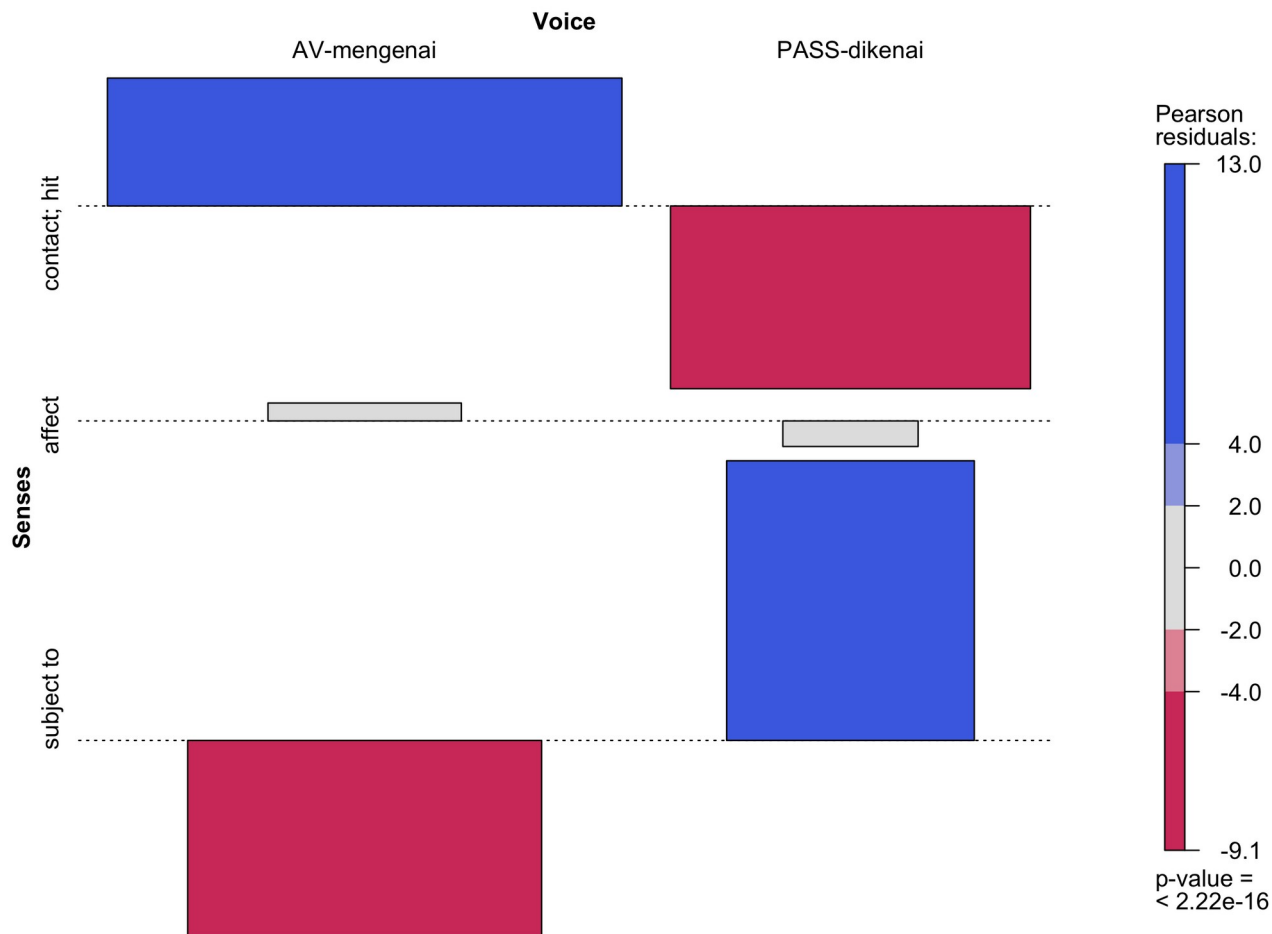


(Rajeg, Rajeg & Arka 2020: 313)

- ‘impose; subject to’ sense of *kenai* is directly constructed at, and here strongly associated with, the passive morphological cxn (cf. Booij 2010)
- This sense is NOT DERIVED from (an imaginary) active form *mengenai*
- AV:*mengenai* ‘impose’ is a significantly absent (negative evidence) form-meaning pairing
- AV:*mengenai* is strongly associated with literal, physical sense of ‘contact; hit’

$X^2 = 363.699$, $df = 2$, $p_{\text{two-tailed}} < 0.001$, Cramer's $V = 0.927$

Association plot between senses of *kenai* and voice



Bluish shading indicates positive residuals while redish shows negative residuals.
 Significant positive association (bluish): strong preference of 'hit' for AV and 'subject to' for PASS.

Rajeg, Gede Primahadi Wijaya, I Made Rajeg & I Wayan Arka. 2020. Supplementary materials for "Corpus-based approach meets LFG: Puzzling voice alternation in Indonesian." Open Science Framework. <https://doi.org/10.17605/OSF.IO/YMD2V>.

mengenai has been **grammaticalised** into prepositional meaning ‘regarding to; concerning; about’

1) Ia tidak ingin teman-temannya tahu **mengenai** siapa kakaknya itu
3SG NEG want friend.PL know concerning who older.sibling DEM

‘(S)he does not want h(is/er) friends know *about/regarding* who h(is/er) older sibling is (...)

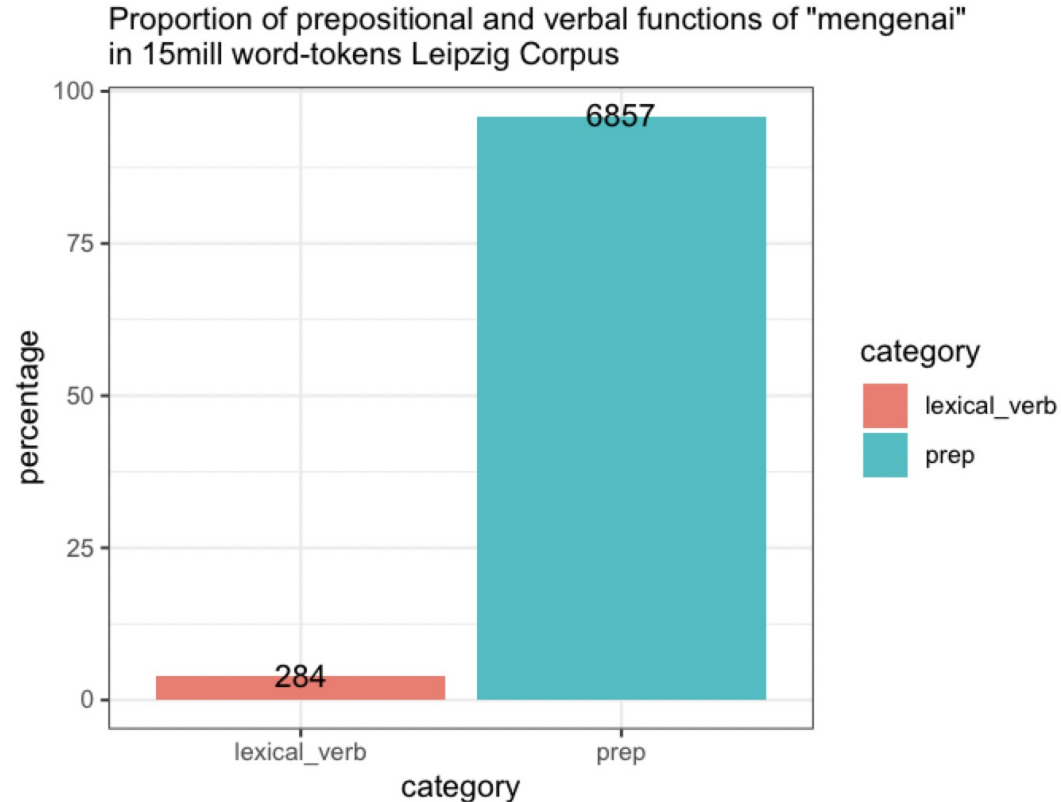
2) Bukti yang paling nyata **mengenai** hal ini adalah ...
evidence REL most real concerning matter DEM BE

‘The most concrete evidence *regarding* this matter is ...’

3) **Mengenai** apa yang disampaikan itu menjadi hal berikutnya.
Concerning what REL PASS.deliver=3SG DEM become matter subsequent

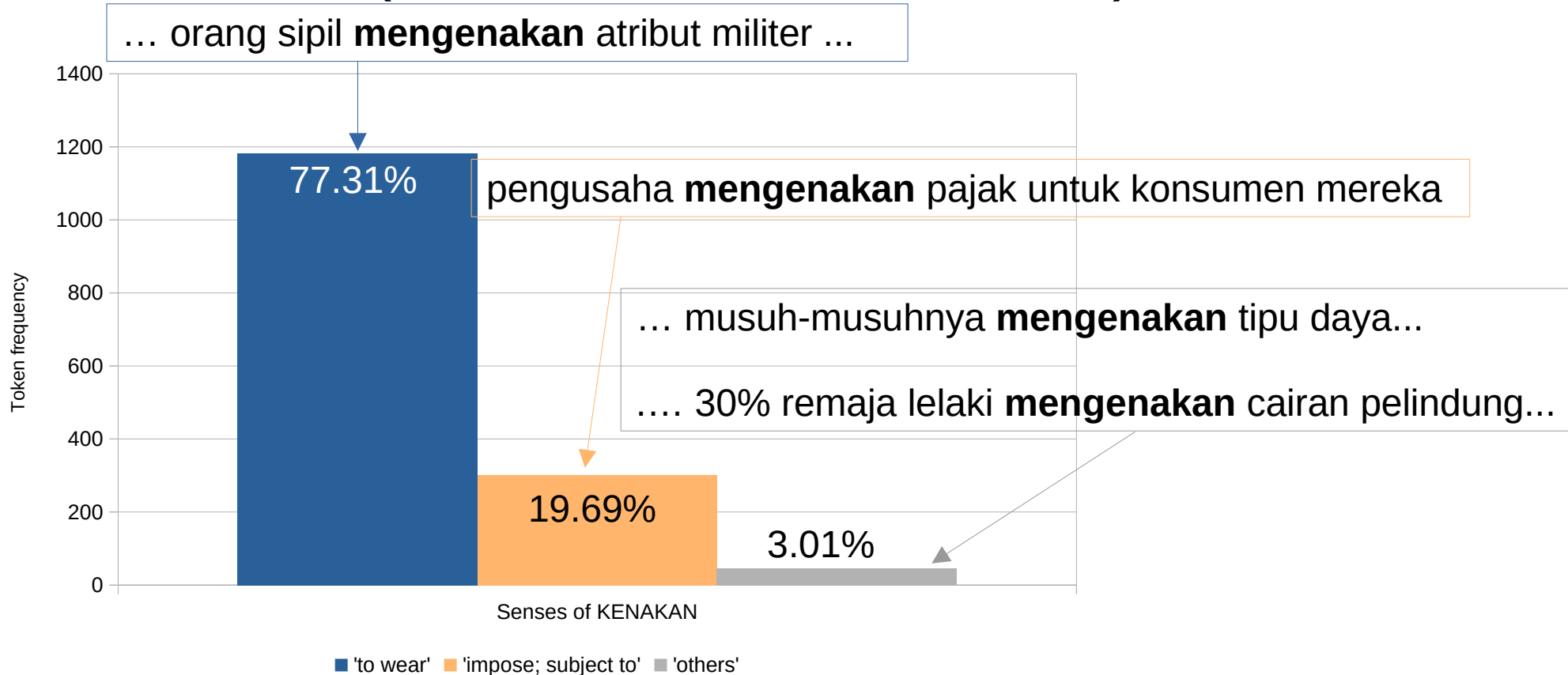
‘*Regarding* what (s)he delivered becomes the subsequent/next matter’

mengenai has been **grammaticalised** into prepositional meaning 'regarding to; concerning; about'

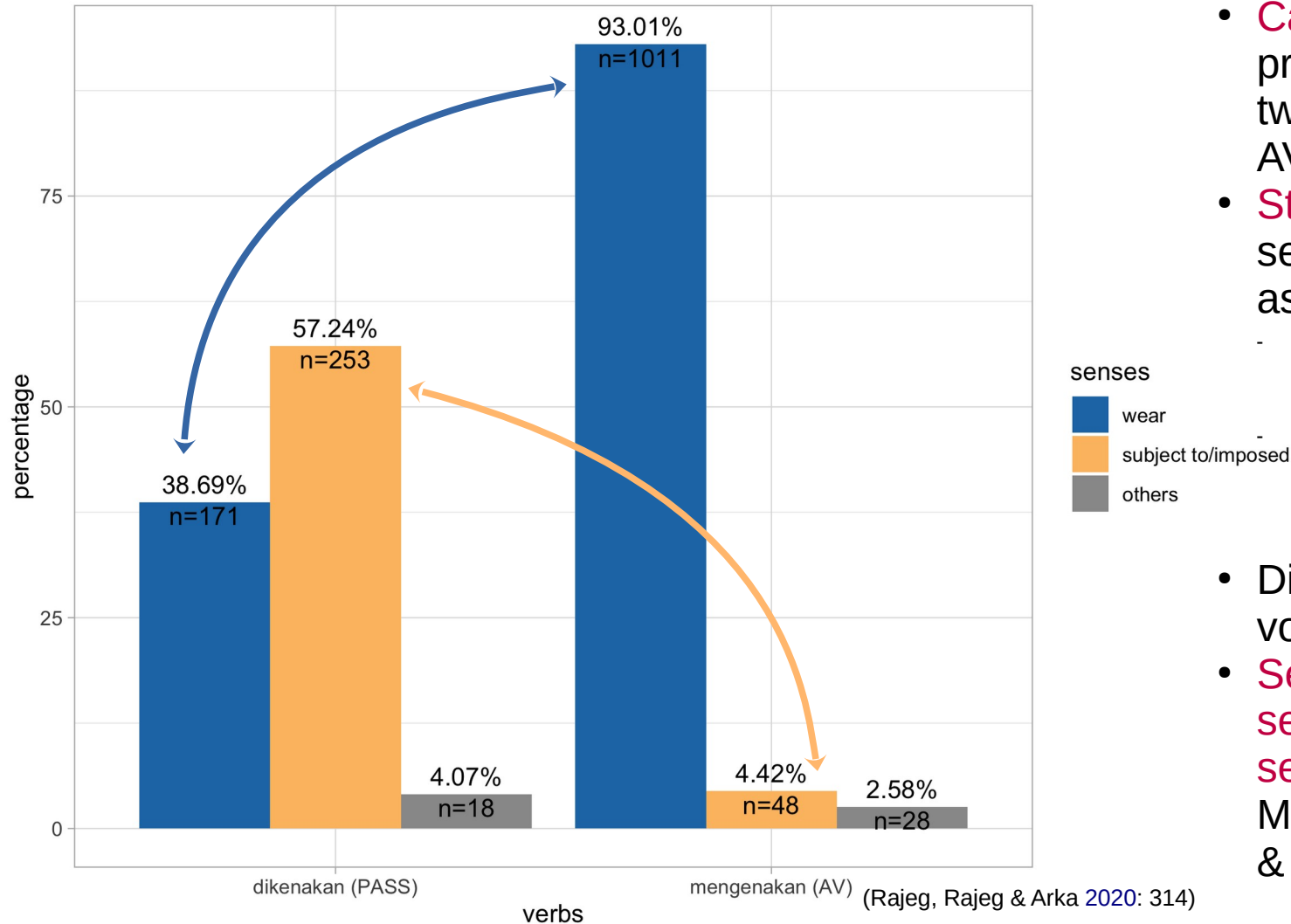


Results for the senses of *kenakan* in AV and
PASS

Senses of *kenakan* (combined in AV & PASS)

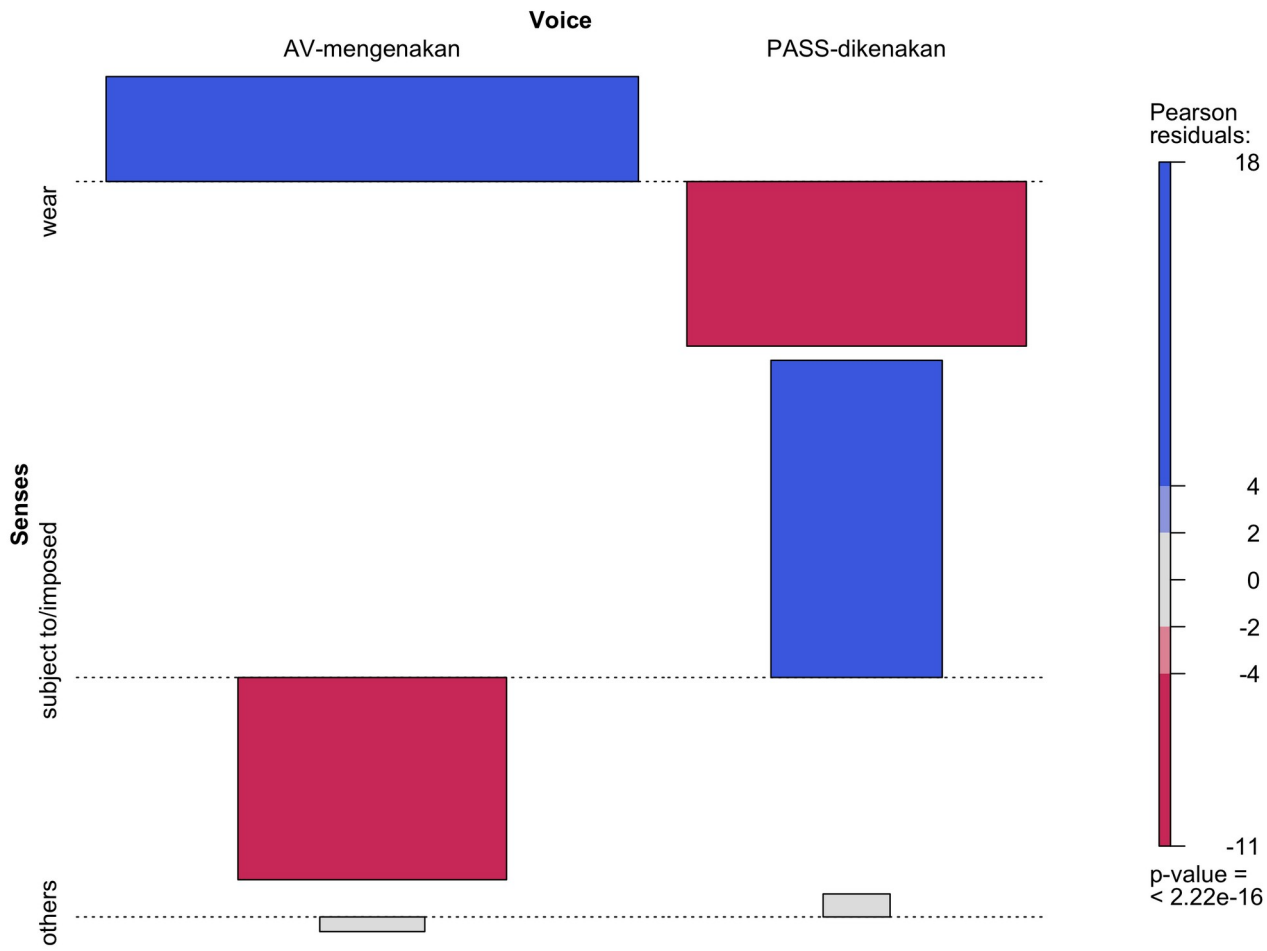


Distribution of senses for *kenakan* in PASS and AV



- **Categorically** meaning-preserving. That is, the two senses are attested in AV and PASS.
- **Statistically**, the two senses exhibit significant asymmetric distribution:
 - 'wear' strongly prefers AV
 - 'impose' strongly prefers PASS
- Distributional nuance of voice alternation
- **Semantic factor and sensitivity in voice selection/alternation** (cf. McDonnell 2016; Bernolet & Coleman 2016)

Association plot between senses of *kenakan* and voice



Bluish shading indicates positive residuals while redish shows negative residuals.
Significant positive association (bluish): strong preference of 'wearing' for AV and 'subject to' for PASS.

Rajeg, Gede Primahadi Wijaya, I Made Rajeg & I Wayan Arka. 2020. Supplementary materials for "Corpus-based approach meets LFG: Puzzling voice alternation in Indonesian." Open Science Framework. <https://doi.org/10.17605/OSF.IO/YMD2V>.

- Why is *mengenai* infelicitous, and not interchangeable with *mengenakan*, to express ‘impose’?
 - *mengenai* ‘impose’ is a significantly absent form-meaning pairing
 - *mengenai* is predominantly used in its grammaticalised sense
 - *mengenai* is strongly associated with literal, physical hitting/contact sense
 - *mengenakan* IS ATTESTED to express ‘impose’ BUT much less typical than expressing ‘wear; put on’
- Why can PASS *dikenai* and *dikenakan* be interchangeable to convey ‘impose’?
 - These PASS forms are both positively and strongly associated with ‘impose’

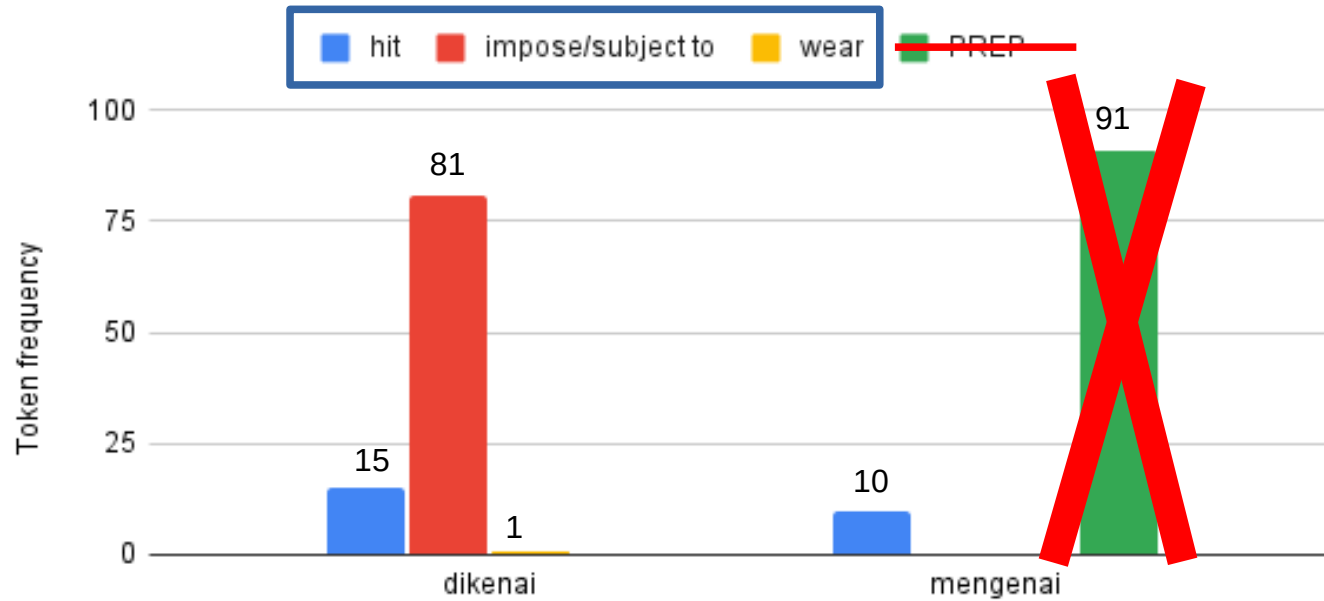
Combining corpus-based data and experimental, sentence-production data

- Assessing how strong the statistical tendency revealed via corpus data is represented in the speakers' linguistic knowledge of the verbs in questions.
 - Do speakers store such statistical association between a given verb (in a given voice morphology) and the predominant sense that the verb expresses?

Sentence-production experimental data for *kenai*

Speakers learn and store the specifics of semantic preference for *dikenai* & *mengenai*.
(cf. Goldberg 2006: 49, 56; Dąbrowska 2009)

Sentence-production experiment data



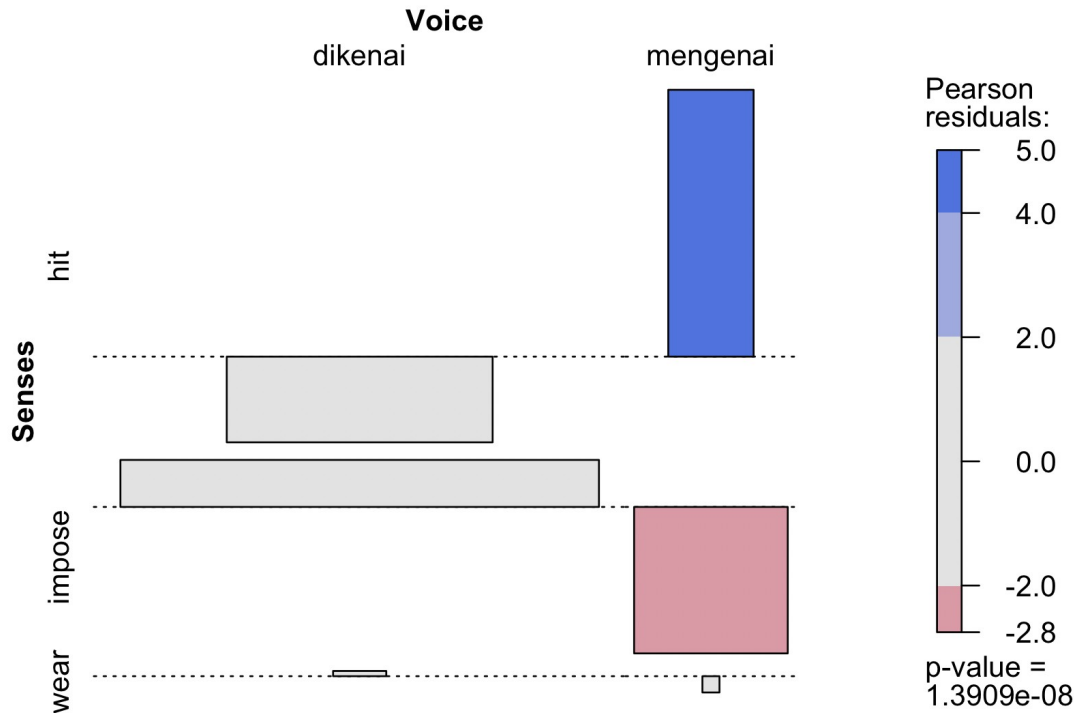
$p_{\text{fisher-exact}} < 0.001$; Cramer's $V = 0.582$

usage

- Dąbrowska, Ewa. 2009. Words as constructions. In Vyvyan Evans & Stephanie Pourcel (eds.), *New directions in cognitive linguistics*, 214–237. Amsterdam: Philadelphia: John Benjamins Publishing Company.
- Goldberg, Adele E. 2006. *Constructions at work: The nature of generalization in language*. Oxford: New York: Oxford University Press.

Sentence-production experimental data for *kenai*

Association plot between senses of *kenai* and voice



- ‘hit; physical contact’ is positively and **strongly** associated with AV *mengenai*
- ‘impose’ is positively but **weakly** associated with PASS *dikenai*
- ‘impose’ is **strongly dissociated** with AV *mengenai* (pink bar)

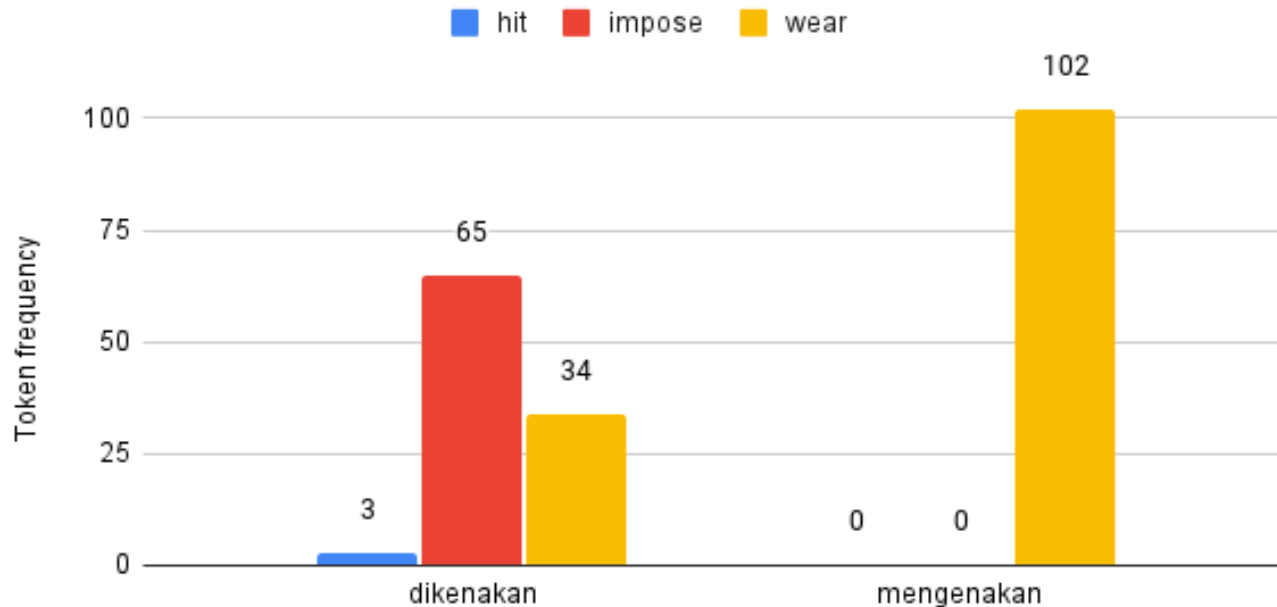
Bluish shading indicates positive residuals while redish shows negative residuals.

Significant positive association (bluish): strong preference of 'hit' for AV.

Sentence-production experimental data for *kenakan*

Speakers learn and store the specifics of semantic preference for *dikenakan* & *mengenakan*.
(cf. Goldberg 2006: 49, 56; Dąbrowska 2009)

Sentence-production experiment data

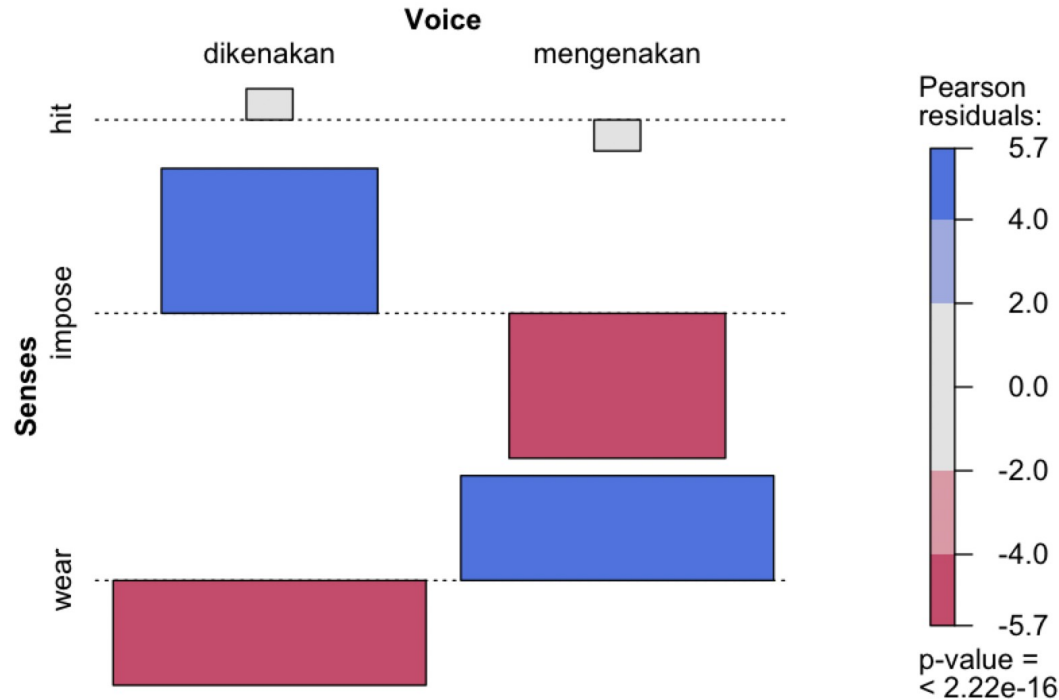


$p_{\text{fisher-exact}} < 0.001$; Cramer's $V = 0.707$ usage

- Dąbrowska, Ewa. 2009. Words as constructions. In Vyvyan Evans & Stephanie Pourcel (eds.), *New directions in cognitive linguistics*, 214–237. Amsterdam & Philadelphia: John Benjamins Publishing Company.
- Goldberg, Adele E. 2006. *Constructions at work: The nature of generalization in language*. Oxford & New York: Oxford University Press.

Sentence-production experimental data for *kenakan*

Association plot between senses of *kenakan* and voice



- 'wear; put on' is positively and **strongly** associated with AV *mengenakan*
- 'impose' is positively and **strongly** associated with PASS *dikenakan*

Bluish shading indicates positive residuals while redish shows negative residuals.

Significant positive association (bluish): strong preference of 'impose' for PASS & 'wear' for AV.

Study with causative transitive motion-verbs: *majukan, aju(kan), mundurkan, undur(kan)*

Usage-based, quantitative perspective to the meaning-preserving hypothesis

Broader contexts (II)

- Growing interests in usage-based, Construction Grammar (CxG) on the interaction of (non-)metaphoric meanings and grammatical constructions (cf. Deignan 2006; Sullivan 2013; Sokolova 2013)
- Sokolova (2013)
 - Metaphoric extension involves structural change (not only distinct collocates)
 - ▶ Russian Locative Alternation

* Deignan, Alice, 2006. The grammar of linguistic metaphors. In Anatol Stefanowitsch & Stefan Th. Gries (eds.), *Corpus-based approaches to metaphor and metonymy*, 106-122. Berlin: Mouton de Gruyter.

* Sokolova, Svetlana, 2013. Verbal predication and metaphor: How does metaphor interact with constructions? *Journal of Slavic Linguistics* 21(1), 171-204.

* Sullivan, Karen, 2013. Figures and constructions in metaphorical language. *Constructional Approaches to Language 10*. Amsterdam: John Benjamins Publishing Company.

13 https://twitter.com/Terry_McDonough/status/1221372869541867520

UCREL CRS : Usage-based perspective on the meaning-preserving hypothesis in voice alternation (Gede)

<https://youtu.be/U3Ti897MHik>

Usage-based perspective on the meaning-preserving hypothesis in voice alternation

Form-meaning relationship in Indonesian voice-morphological constructions: **SUMMARY**

- Voice alternation of the same verb does **NOT ALWAYS** preserve meaning/sense of the verb
 - Passive form is **NOT ALWAYS** derived from Active form (esp. for certain sense)
 - **Passive has distinct semantic constraints than Active** (cf. Hilpert 2014:41)
- **Certain sense of a verb tends to be (statistically speaking) strongly associated with certain voice-morphological form**
 - Semantic factor in voice selection of the verb (cf. McDonnell 2016)
- Sentence-production experiment provides some converging evidence that speakers also store the preferred sense associated with a given form in their linguistic knowledge of the verb:
 - **Frequency effect** – frequent exposure by speakers for the detailed semantic preference of the verbs in certain voice-morphological constructions (cf. Dąbrowska 2009)
 - **Item-specific representation of linguistic knowledge** in usage-based, Construction Grammar (Goldberg 2006; Diessel 2015; Dąbrowska 2009)

Concluding remarks

- About corpus linguistics:
 - Using computer software to analyse large collection of machine-readable texts
 - Access to quantitative data from large qualitative (i.e., textual) data
 - Some basic analytical tools:
 - Concordance – keyword-in-context (KWIC) display
 - Collocation – (statistical) co-occurrence of words
 - Word-sequence/cluster
 - Word frequency-list
 - These resources are of little use UNLESS coupled with research questions and aims at what to do with the large-scale textual and quantitative data (e.g., in the context of theoretically motivated questions/hypothesis to be tested/investigated)
 - Knowledge about statistics is essential to analyse the quantitative corpus-based data so that it can shed light on, and answer, the research questions
 - Primarily viewed as methodology: **a means to an end**, not necessarily the end in itself.

An overview of corpus linguistics and its application to form-meaning relationship in Indonesian voice-morphological constructions

Gede Primahadi Wijaya Rajeg

Bachelor of English, Faculty of Humanities, Universitas Udayana, Indonesia

Keynote presentation at the Linguistics Master's Program of Universitas Sebelas Maret, Indonesia

Wednesday, 25 August 2021

 <https://orcid.org/0000-0002-2047-8621>

 @PrimahadiWijaya