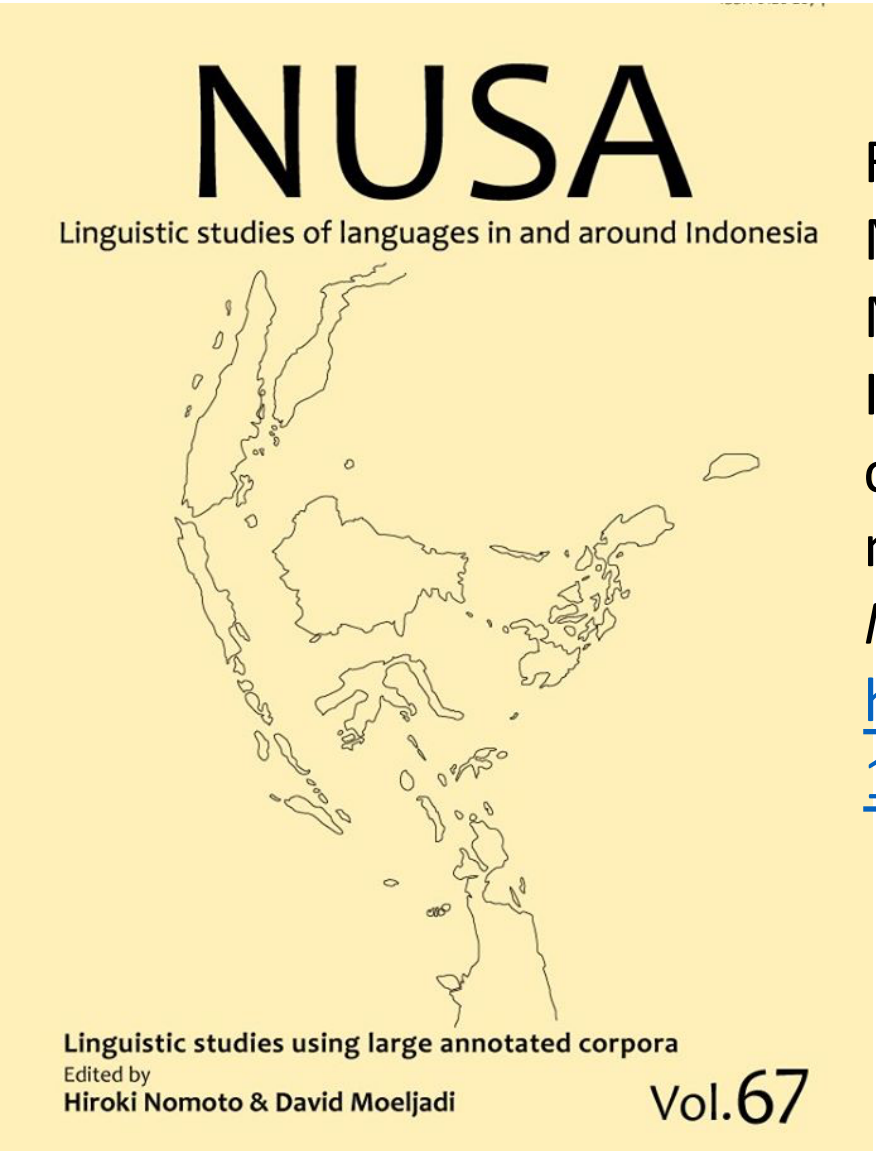
Big data, machine learning, & computational lexical semantics

A case study with Indonesian denominal verbs

Gede Primahadi Wijaya Rajeg; Karlina Denistia; Simon Musgrave

Universitas Udayana; Eberhard Karls Universität Tübingen; Monash University

Fakultas Ilmu Budaya (FIB) Research Talk (FReTalk) Session 2
Universitas Udayana, Indonesia
9 Mei 2020



Rajeg, G. P. W., Denistia, K., & 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*, *67*, 35–76. http://repository.tufs.ac.jp/handle/10108/94452

Supplementary materials also #OpenAccess

- **Dataset, incl. the Vector Space Model**
  - Rajeg, G. P. W., Denistia, K., & Musgrave, S. (2019). *Dataset for* Vector space model and the usage patterns of Indonesian denominal verbs. figshare. https://doi.org/10.6084/m9.figshare.8187155

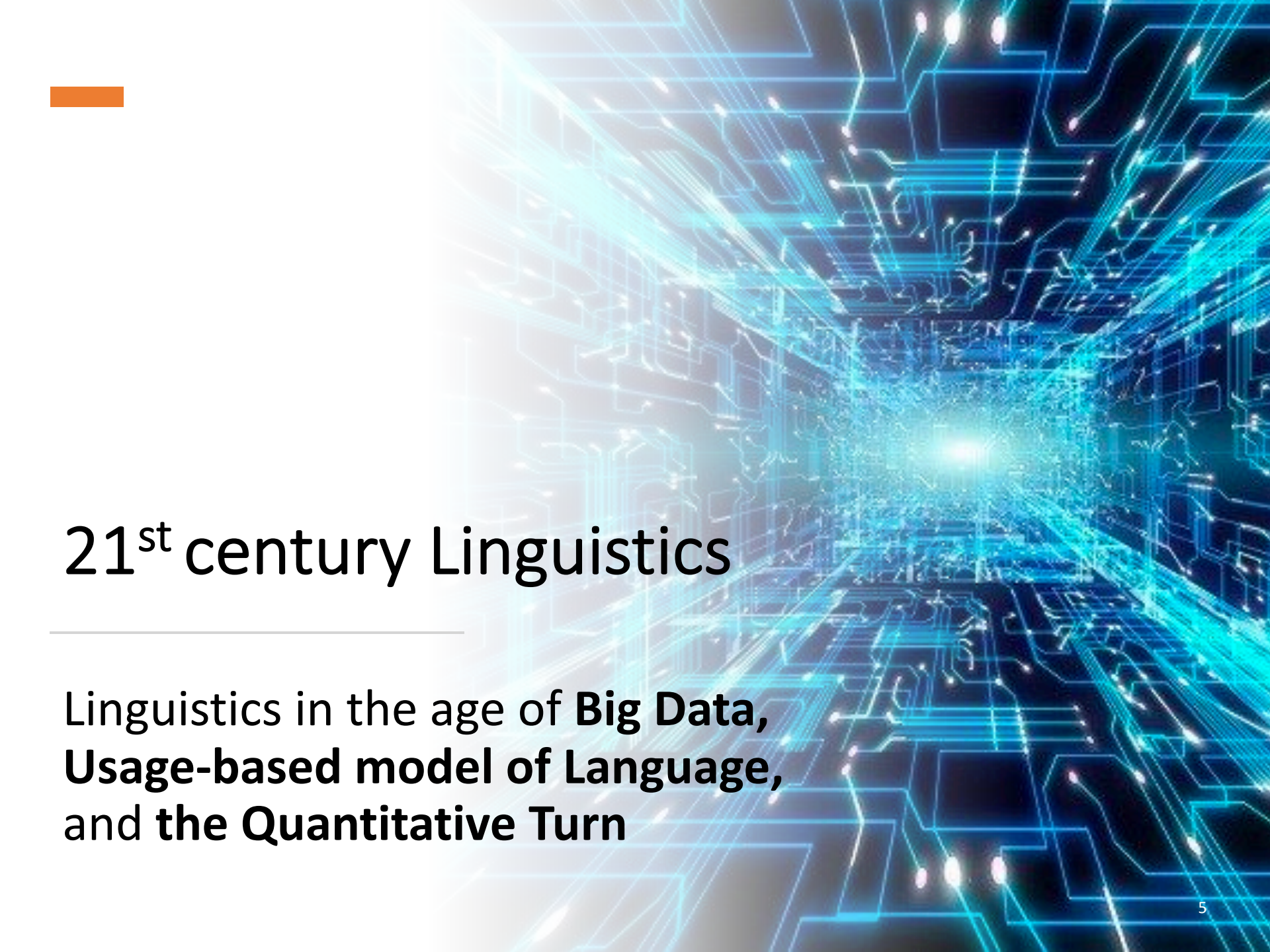
- **R Markdown Notebook (containing the R programming codes for statistical analyses)**
  - Rajeg, G. P. W., Denistia, K., & Musgrave, S. (2019). *R Markdown Notebook for* Vector space model and the usage patterns of Indonesian denominal verbs. figshare. https://doi.org/10.6084/m9.figshare.9970205

Overview

- 21st century Linguistics

- Distributional semantics and *Vector Space Model*
  - A case study on semantic (dis)similarity and cluster between denominal verbs with *meN-, meN-/-kan,* and *meN-/-i* affixes

- Conclusion & outlook



21st century Linguistics

Linguistics in the age of **Big Data, Usage-based model of Language,** and **the Quantitative Turn**

1. Big Data

What is “Big Data”?

“new ways of taking large amounts of data and using software tools **to identify previously undiscovered patterns, trends, correlations, and associations**.”

DePalma (n.d.)

DePalma, D. (n.d.). *Where Does Language Fit in with Big Data? | The Chronicle*. Retrieved 22 March 2020, from https://www.ata-chronicle.online/featured/where-does-language-fit-in-with-big-data/

Big data in linguistics

- The advent of massive electronic texts
  - (Language) corpus (plural: corpora) – from Latin: 'body'
  - (Transcribed) natural spoken and written texts
- Amounting to millions as well as billions of words
  - Check https://www.english-corpora.org/
- Difficult to process without specialised computational tools (e.g. as in corpus linguistics)
  - Recently, with the use of programming languages (e.g. R or Python) (cf. Gries, 2009)
- Quantitative-Qualitative mix is possible

Gries, S. Th. (2009). *Quantitative Corpus Linguistics with R: A Practical Introduction*. Routledge.

Big data in linguistics

- Potential advantages are obvious:
  - Rapidly changing the way linguists do research
  - Allowing empirical study of rare phenomena impossible otherwise (e.g. when using traditional, intuitive method)
  - Addressing old, key research questions in a new light
  - Questioning previous findings and existing theories with new, empirical analyses

Hiltunen, T., McVeigh, J., & Säily, T. (2017). How to turn linguistic data into evidence? *Studies in Variation, Contacts and Change in English*, *19*. http://www.helsinki.fi/varieng/series/volumes/19/

2. Usage-based linguistics

(Diessel, 2017; Langacker, 1988; Tummers et al., 2005, among others)

- Diessel, H. (2017). Usage-based linguistics. In M. Aronoff (Ed.), *Oxford Research Encyclopedia of Linguistics*. Oxford University Press. https://doi.org/10.1093/acrefore/9780199384655.013.363
- Langacker, R. W. (1988). A usage-based model. In B. Rudzka-Ostyn (Ed.), *Topics in Cognitive Linguistics* (pp. 127–161). John Benjamins Publishing Company.
- Tummers, J., Heylen, K., & Geeraerts, D. (2005). Usage-based approaches in Cognitive Linguistics: A technical state of the art. *Corpus Linguistics and Linguistic Theory*, *1*(2), 225–261.

Big data in linguistics & Usage-based model of language

- Availability of non-elicited usage-data (e.g. a linguistic corpus) underlies the methodological state-of-the-art for usage-based (cognitive) linguists

- Strong motivation for empirical research:

  "the usage data constitute the empirical foundation from which general patterns can be abstracted" (Tummers et al., 2005, p. 228)

Tummers, J., Heylen, K., & Geeraerts, D. (2005). Usage-based approaches in Cognitive Linguistics: A technical state of the art. *Corpus Linguistics and Linguistic Theory*, *1*(2), 225–261.

Big data in linguistics & Usage-based model of language

- Dialectic relationship between language system and use:

> "the grammar does not only constitute the knowledge repository to be employed in language use, but it is also itself the product of language use."
>
> (Tummers et al., 2005, p. 228)

Tummers, J., Heylen, K., & Geeraerts, D. (2005). Usage-based approaches in Cognitive Linguistics: A technical state of the art. *Corpus Linguistics and Linguistic Theory*, *1*(2), 225–261.

"When you can measure what you are speaking about, and express it in numbers, you know something about it; but when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meagre and unsatisfactory kind"

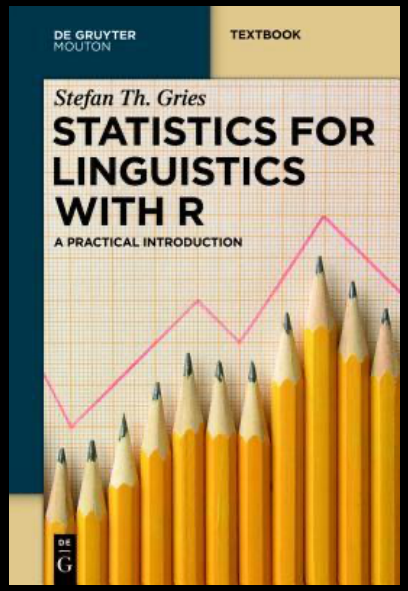
William Thomson, Lord Klevin
Lecture on "Electrical Units of Measurement"(3 May 1883)
https://en.wikiquote.org/wiki/William_Thomson

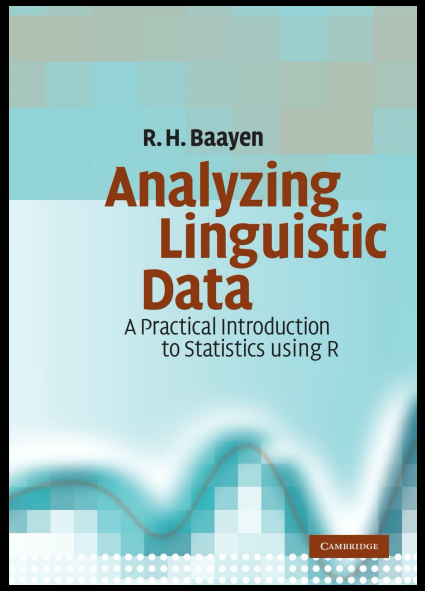
3. The Quantitative Turn

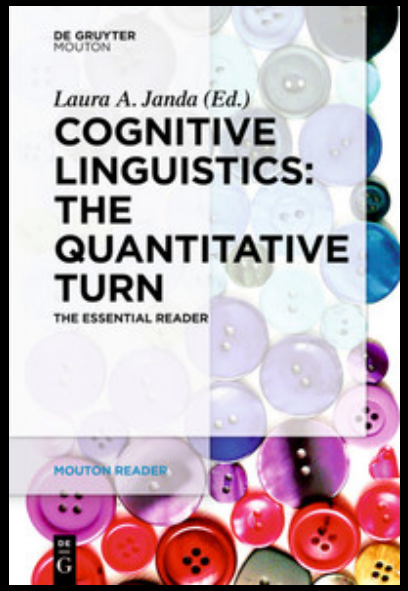
(Janda, 2013, 2016, 2017)

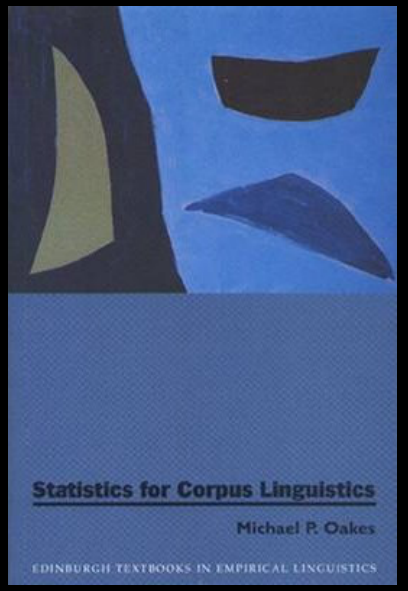
- Janda, L. A. (2013). Quantitative methods in *Cognitive Linguistics*: An introduction. In L. A. Janda (Ed.), *Cognitive Linguistics: The quantitative turn* (pp. 1–32). Mouton de Gruyter.
- Janda, L. A. (2016). Linguistic profiles: A quantitative approach to theoretical questions. *Język i Metoda*, 127–145.
- Janda, L. A. (2017). The quantitative turn. In B. Dancygier (Ed.), *The Cambridge handbook of Cognitive Linguistics* (pp. 498–514). Cambridge University Press. https://doi.org/10.1017/9781316339732.032

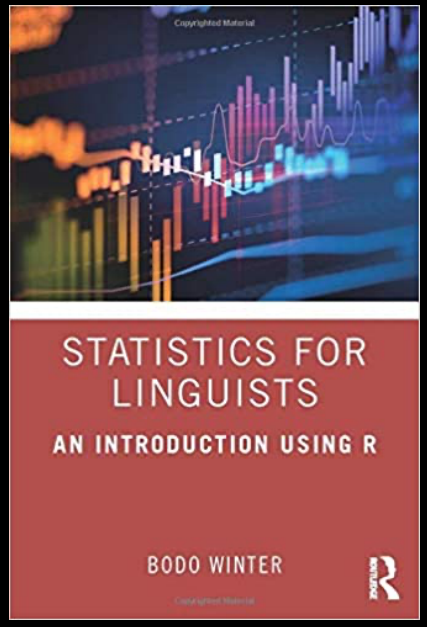
"When you can measure what you are speaking about, and express it in numbers, you know something about it" William Thomson, Lord Klevin











The Quantitative Turn

- Statistics is the only discipline that provides means to analyse, summarise, and detect pattern from large numerical dataset derived from extensive corpus linguistic and/or experimental linguistic studies.

The Quantitative Turn

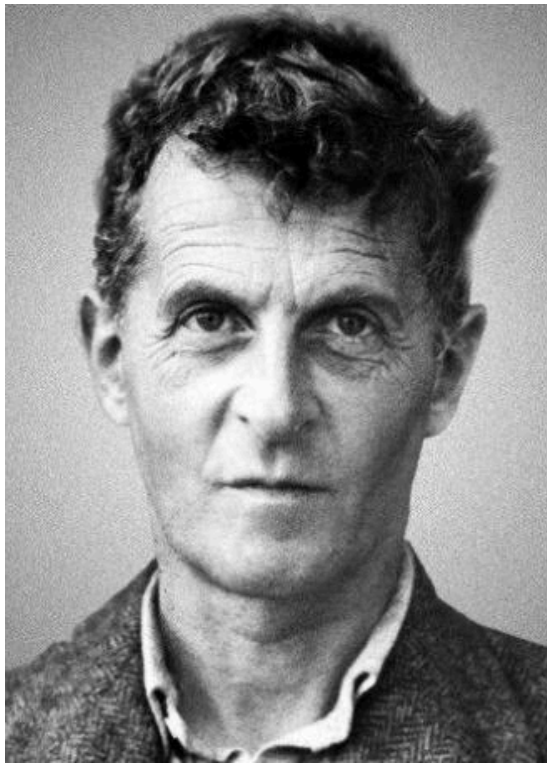
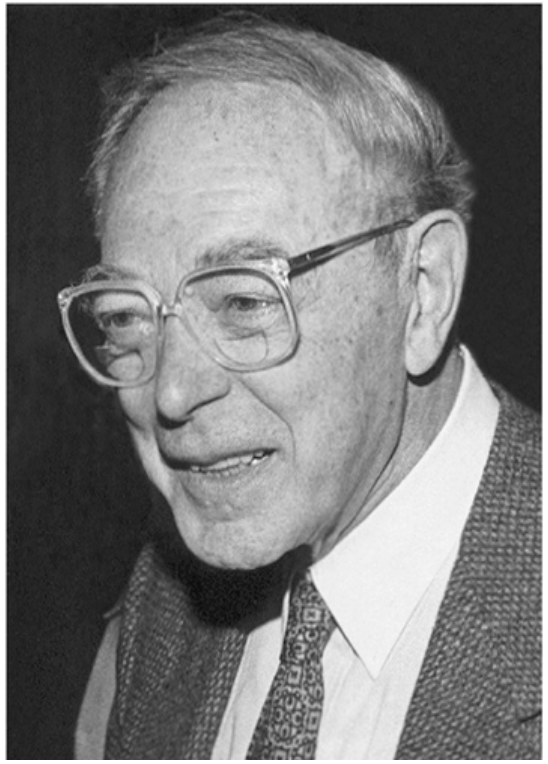
- Computational Linguistics has developed statistically-based techniques to learn patterns from massive amount of data (e.g. in Natural Language Processing [NLP])
  - Thus **machine learning**, letting the computer to learn pattern from data based on statistical algorithm
  - statistical machine translation, document similarity

Brownlee, J. (2017, September 21). What Is Natural Language Processing? Retrieved 8 May 2020, from Machine Learning Mastery website: https://machinelearningmastery.com/natural-language-processing/

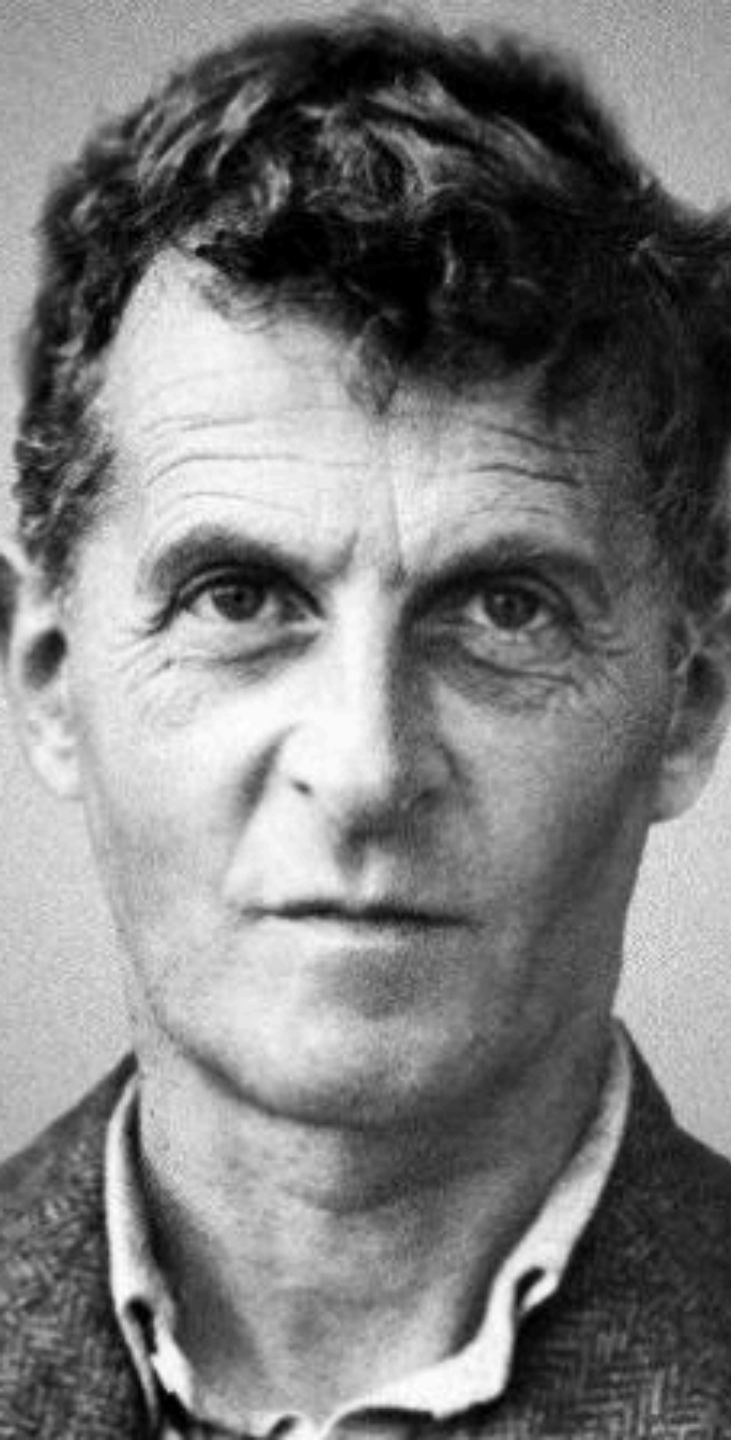
The Quantitative Turn

- We apply one of the machine learning tools by the name of **Vector Space Model (VSM)**
  - rooted in distributional semantics (next section)
  - implemented in **word2vec** algorithm developed by Google team (Mikolov et al., 2013; cf. Schmidt, 2015; Heuser, 2016)

- Heuser, R. (2016, June 1). Word Vectors in the Eighteenth Century, Episode 2: Methods – Adventures of the Virtual. Retrieved 6 August 2017, from http://ryanheuser.org/word-vectors-2/
- Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *ArXiv:1301.3781 [Cs]*. Retrieved from http://arxiv.org/abs/1301.3781
- Schmidt, B. (2015, October 25). Vector Space Models for the Digital Humanities. Retrieved 24 January 2016, from http://bookworm.benschmidt.org/posts/2015-10-25-Word-Embeddings.html

From Distributional Semantics to Vector Space Models



Distributional semantics

"The meaning of a word is its use in the language"

(Wittgenstein, 1953, Section 43)

Wittgenstein, L. (1953). *Philosophical investigations*. New York: Macmillan.

Distributional semantics

"You shall know a word by the company it keeps"

(Firth, 1957, p. 11)

Firth, J. R. (1957). *A synopsis of linguistic theory 1930-55. 1952–59*, 1–32.

Distributional semantics

"If we consider words or morphemes *A* and *B* to be more different in meaning than *A* and *C*, then we will often find that the distribution of *A* and *B* are more different than the distributions of *A* and *C*. In other words, **difference of meaning correlates with difference in distribution**."

(Harris, 1954, p. 156; boldface is ours)

Harris, Z. S. (1954). Distributional Structure. *WORD*, *10*(2–3), 146–162. https://doi.org/10.1080/00437956.1954.11659520

How do those **theoretical** ideas make their ways into *vector space model*?

Vector Space Model (VSM): from Words 2 Vectors

- A word in the corpus is represented by a vector:
  - co-occurrence frequencies of the target word with other words in the corpus (i.e. the context words)
  - within a certain window-span around the target word (e.g., up to 3 or 5 words to the left and right of the target)
  - organised in a words-contexts co-occurrence table/matrix
  - trying to make precise the ideas in distributional semantics

- Turney, P. D., & Pantel, P. (2010). From frequency to meaning: Vector Space Models of semantics. *Journal of Artificial Intelligence Research*, *37*(1), 141–188. https://doi.org/10.1613/jair.2934
- Erk, K. (2012). Vector space models of word meaning and phrase meaning: A survey. *Language & Linguistics Compass*, *6*(10), 635–653. https://doi.org/10.1002/lnco.362

Vector Space Model (VSM): from Words 2 Vectors

context words > analogised to multidimensional **semantic space** of the target words

	bergabung	bukan	dalam	demikian	depan	di	fu	ia	ibunya	kaki	kakinya	ke
melangkah	0	0	0	0	1	0	0	1	0	1	0	3
melangkahi	1	1	0	0	0	0	1	0	0	0	0	0
melangkahkan	0	0	1	1	0	1	0	1	1	1	2	2

target words

vectors

The meaning of “vector” here is derived from its programming sense: ‘an array or a list of numbers’ (Heuser, 2016)

Heuser, R. (2016, June 1). Word Vectors in the Eighteenth Century, Episode 2: Methods – Adventures of the Virtual. Retrieved 6 August 2017, from http://ryanheuser.org/word-vectors-2/

create word-context co-occurrence table/matrix

A case study with Indonesian denominal verbs

with *meN-, meN-/-kan,* & *meN-/-i* affixes

Methods in a nutshell

- Schmidt, B., & Li, J. (2017). wordVectors: Tools for creating and analyzing vector-space models of texts (Version 2.0) [R package]. Retrieved from http://github.com/bmschmidt/wordVectors

Snippet of vector space model for 184,666 word-types

```
> vsm <- read.binary.vectors("data/leipzig_w2v_vector_full.bin")
Reading a word2vec binary file of 184666 rows and 100 columns
  |======================================================================| 100%
> vsm
A VectorSpaceModel object of  184666  words and  100  vectors
              [,1]         [,2]         [,3]         [,4]         [,5]         [,6]
</s>    0.004002686  0.004419403 -0.003830261 -0.003278046  0.001366577  0.003021088
yang   -0.054897364  0.020181134  0.022839403 -0.020136114 -0.060841653  0.135852650
dan     0.054462079  0.009463400 -0.016201755  0.137052819 -0.102377541 -0.107336603
di      0.026733818  0.063937731  0.080151923  0.126316801  0.160071045  0.120005257
dengan  0.123633221 -0.119772665 -0.054115091  0.032494470 -0.134128243  0.015288961
ini     0.103471845 -0.089435019 -0.046082057  0.054030232  0.114544198 -0.007455951
untuk  -0.038178634 -0.126596004  0.005099069 -0.017128143 -0.061365727 -0.082834117
dari   -0.114230752 -0.124413706  0.011286141  0.041064896  0.053682573 -0.118735529
itu     0.056879677  0.020124484 -0.060551416 -0.041253708  0.036803395 -0.006934925
dalam  -0.125667900 -0.145851433  0.028326087 -0.071725696 -0.058170870 -0.066587009
```

Methods in a nutshell

Indonesian Leipzig Corpora
180+ million word-tokens

Pre-processing
175+ million word-tokens
184.666 word-types/vocabs

word2vec
wordVectors R package

Retrieve denominal verbs with *MorphInd*

- *me-* words tagged as _VSA with <n> tag for the root.
- with *me-* , *me-/-kan*, and *me-/-i*
- N >= 20 tokens overall

51 verbs

VSM-based Cluster Analysis & Nearest neighbours

- Larasati, S. D., Kuboň, V., & Zeman, D. (2011). Indonesian Morphology Tool (MorphInd): Towards an Indonesian Corpus. *Systems and Frameworks for Computational Morphology*, 119–129. doi: 10.1007/978-3-642-23138-4_8
- Schmidt, B., & Li, J. (2017). wordVectors: Tools for creating and analyzing vector-space models of texts (Version 2.0) [R package]. Retrieved from http://github.com/bmschmidt/wordVectors

Snippet of vector space model for 51 denominal verbs

Console Terminal

~/Google Drive/MeNasal/

```
> vsm_tgt
A VectorSpaceModel object of  51  words and  100  vectors
                    [,1]        [,2]        [,3]       [,4]         [,5]         [,6]
mengatakan    -0.15838896  0.05143370  0.02442524  0.24287735 -0.037270460 -0.128047243
menggunakan   -0.10439277  0.07838520  0.18227956  0.38019755 -0.012107351 -0.019670846
mewakili      -0.25855595 -0.09868021  0.05797758  0.19394790 -0.219191238 -0.049940147
menempatkan    0.01809703 -0.01772323  0.24649994  0.12339904  0.090888657 -0.063201532
menempati     -0.17115426 -0.10566988  0.16301627  0.17471313 -0.105571389 -0.090712421
mengakhiri     0.24585645  0.12317820 -0.12496730  0.18348111 -0.032905310 -0.200849608
melangkah      0.14984114  0.02910020  0.04853236 -0.34504056 -0.339647442 -0.008438474
mencontohkan  -0.24159168  0.01203183  0.07734830  0.13243185  0.002526026  0.068748802
mendasar      -0.45044890 -0.07471947  0.37076223 -0.02369694  0.357669443  0.234728336
menandai       0.01873370  0.19389461  0.17055722  0.15514162  0.026616529 -0.258213133
```

Methods in a nutshell

Indonesian Leipzig Corpora
180+ million word-tokens

Pre-processing
175+ million word-tokens
184.666 word-types/vocabs

word2vec
wordVectors R package

Retrieve denominal verbs with *MorphInd*

- *me-* words tagged as _VSA with <n> tag for the root.
- with *me-* , *me-/-kan*, and *me-/-i*
- N >= 20 tokens overall

51 verbs

VSM-based Cluster Analysis & Nearest neighbours

- Larasati, S. D., Kuboň, V., & Zeman, D. (2011). Indonesian Morphology Tool (MorphInd): Towards an Indonesian Corpus. *Systems and Frameworks for Computational Morphology*, 119–129. doi: 10.1007/978-3-642-23138-4_8
- Schmidt, B., & Li, J. (2017). wordVectors: Tools for creating and analyzing vector-space models of texts (Version 2.0) [R package]. Retrieved from http://github.com/bmschmidt/wordVectors

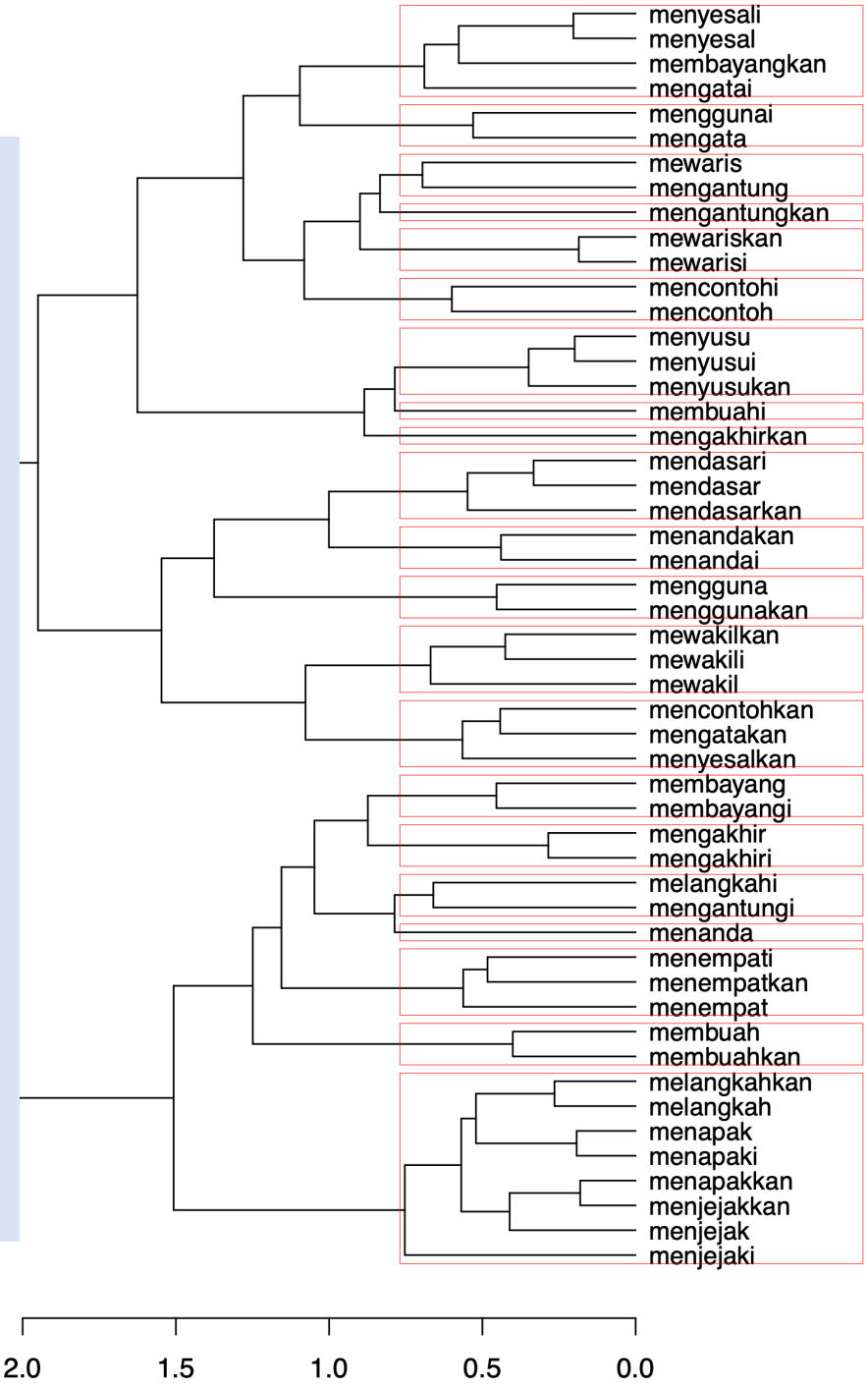
Cluster analysis on the denominal verbs

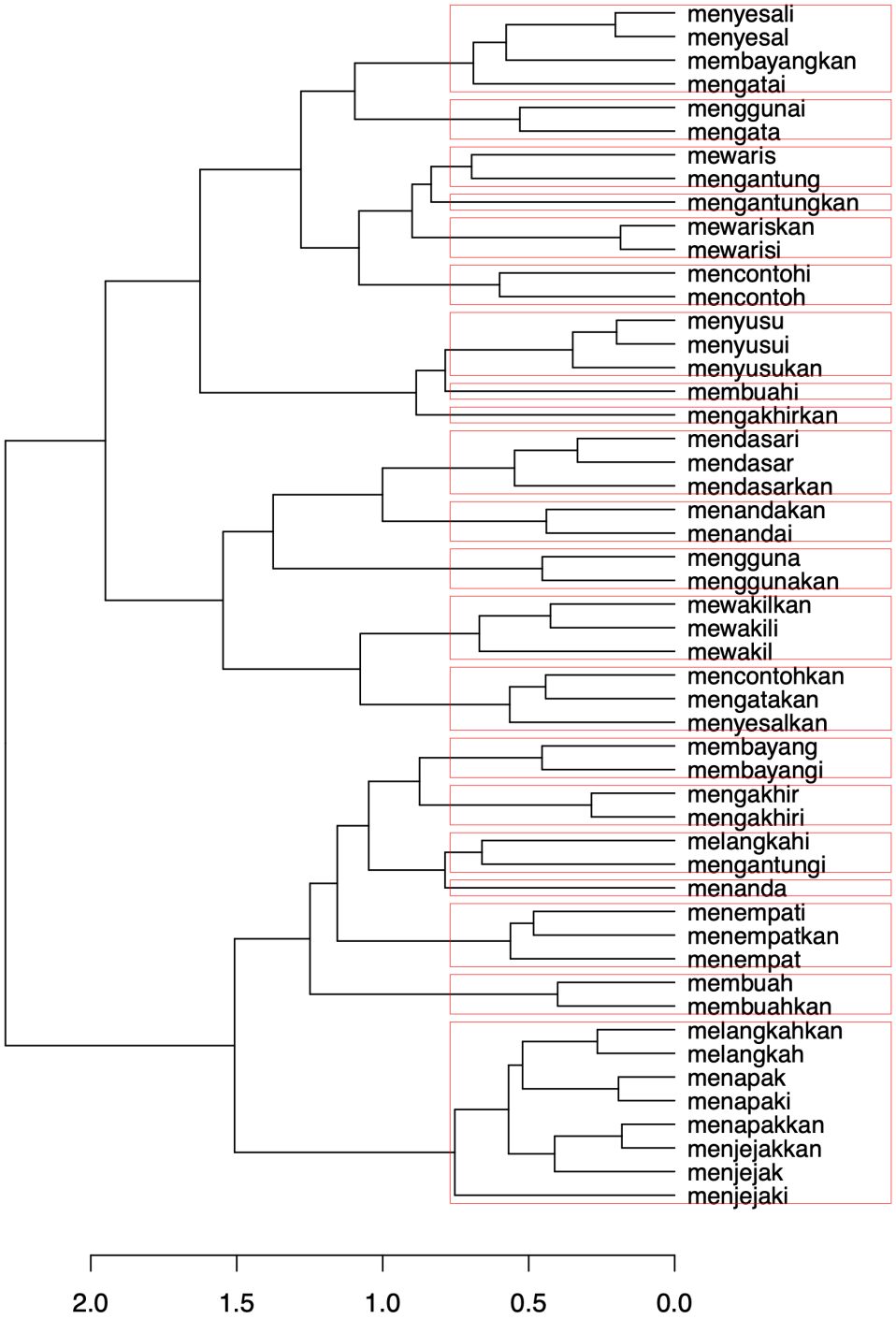
Dendrogram visualising results of *Hierarchical Agglomerative Cluster* analysis

The number of clustering solution are determined automatically using *Average Silhouette Width* (ASW) measure.

The highest ASW score points to 21-cluster solution

Splitting the cluster into sub-clusters (red boxes) are automatically performed in R

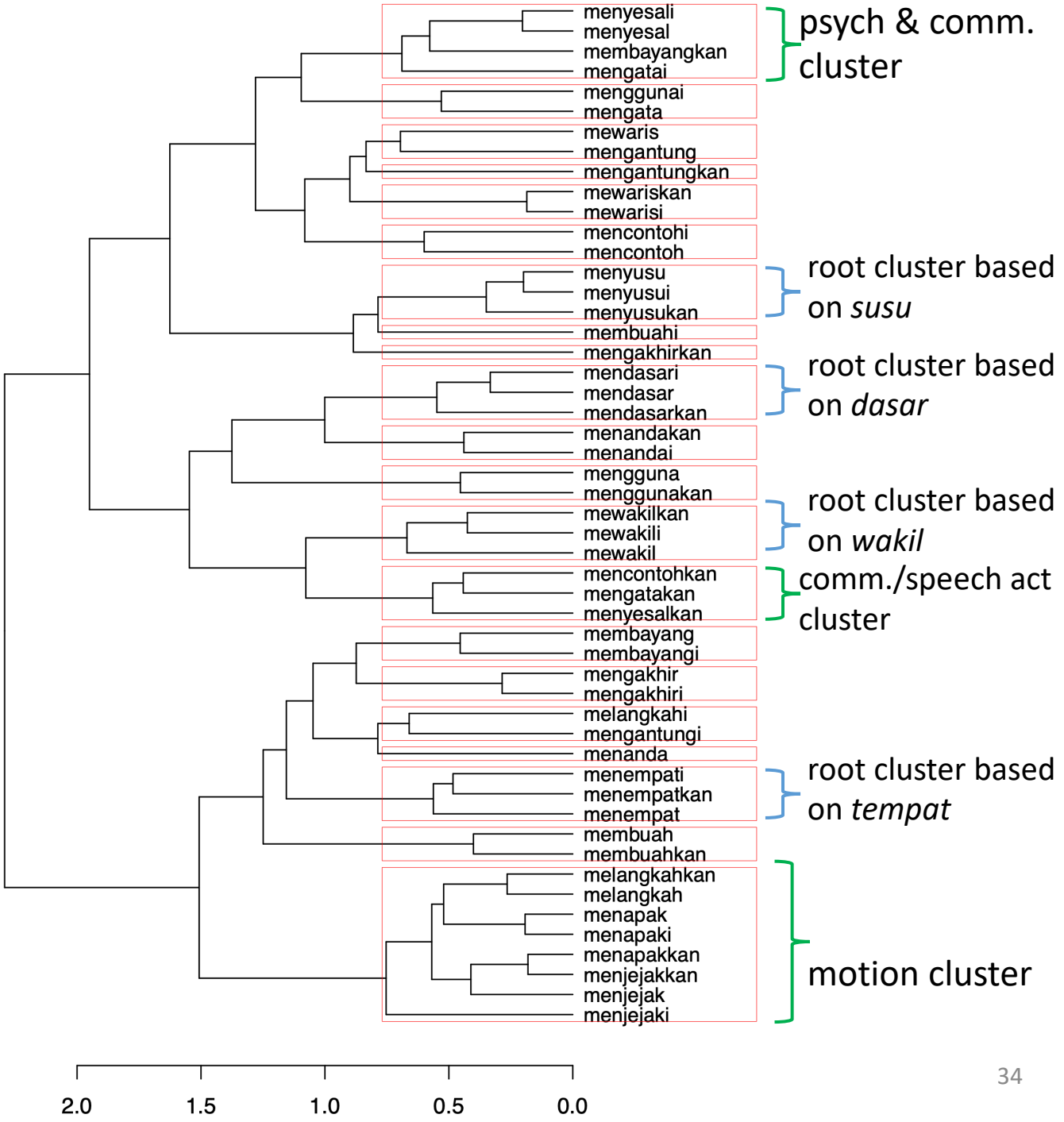




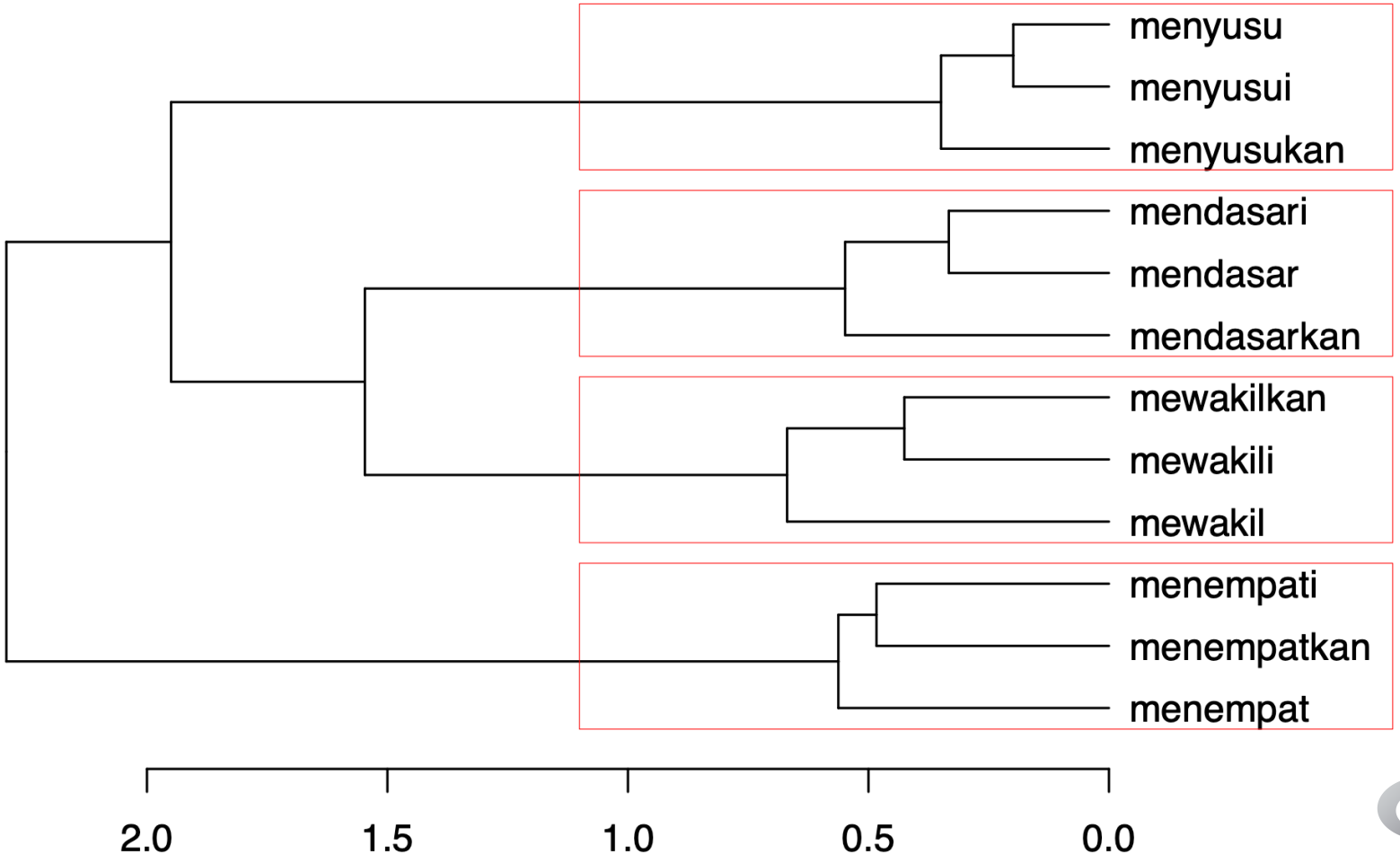
psych & comm. cluster

comm./speech act cluster

motion cluster


menyesali
menyesal
membayangkan
mengatai
menggunai
mengata
mewaris
mengantung
mengantungkan
mewariskan
mewarisi
mencontohi
mencontoh
menyusu
menyusui
menyusukan
membuahi
mengakhirkan
mendasari
mendasar
mendasarkan
menandakan
menandai
mengguna
menggunakan
mewakilkan
mewakili
mewakil
mencontohkan
mengatakan
menyesalkan
membayang
membayangi
mengakhir
mengakhiri
melangkahi
mengantungi
menanda
menempati
menempatkan
menempat
membuah
membuahkan
melangkahkan
melangkah
menapak
menapaki
menapakkan
menjejakkan
menjejak
menjejaki
psych & comm. cluster
root cluster based on *susu*
root cluster based on *dasar*
root cluster based on *wakil*
comm./speech act cluster
root cluster based on *tempat*
motion cluster
2.0
1.5
1.0
0.5
0.0

Root clustering: same root different morphologies



Two hypotheses for the contrast between *–i/-kan* verb-pairs

"[w]ith a number of words the distinction between -**kan** and -**i** is blurred in common usage. In some cases, both -**i** and -**kan** occur with the same meaning. With some there is a recipient or locative object, while with others the object is the patient."

(Sneddon et al., 2010, p. 101)

<< We'll call this the **similarity hypothesis** >>

Sneddon, J. N., Adelaar, A., Djenar, D. N., & Ewing, M. C. (2010). *Indonesian reference grammar* (2nd ed.). Allen & Unwin.

Two hypotheses for the contrast between *–i/-kan* verb-pairs

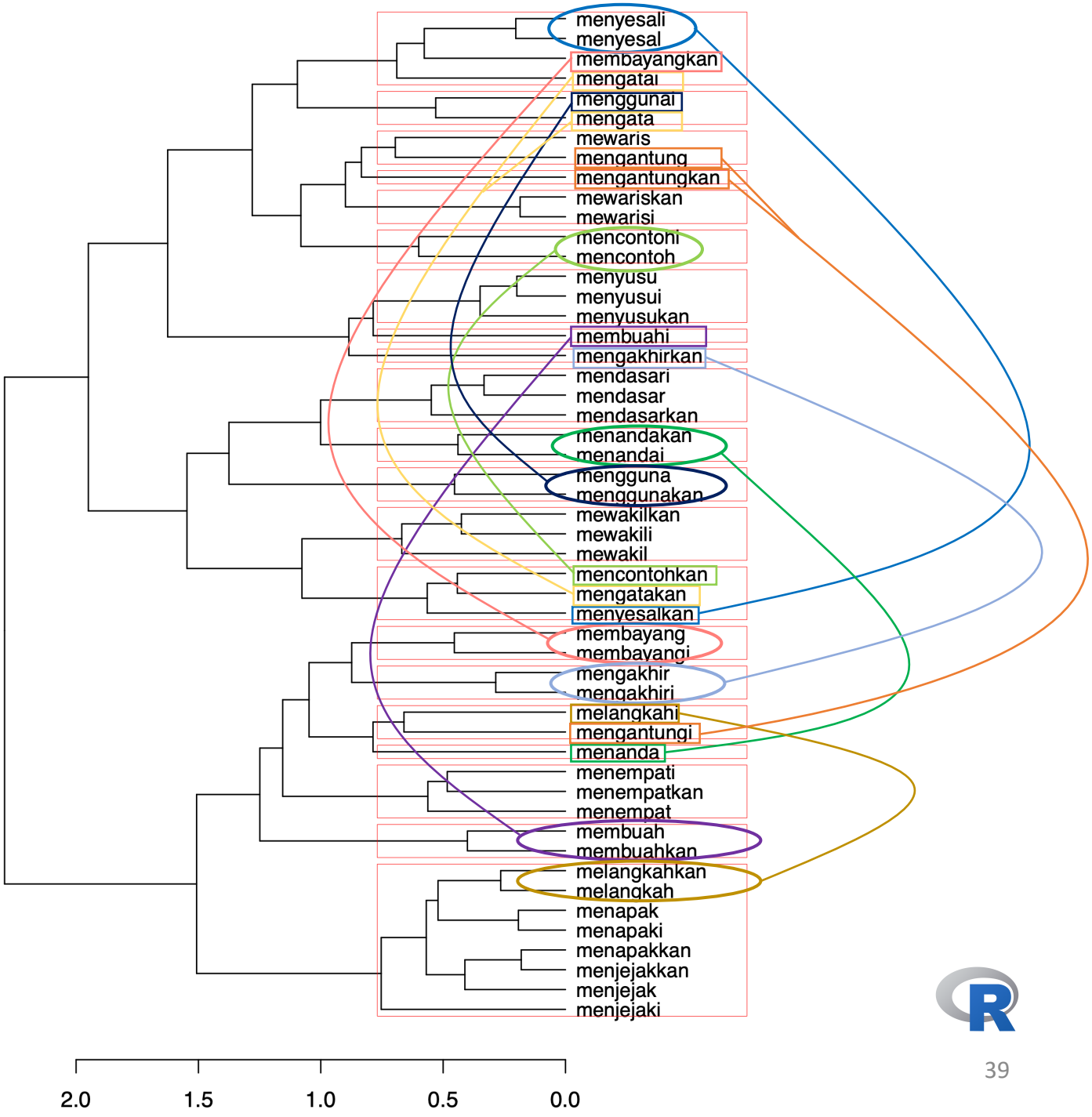
"[m]any verbs can take both -**kan** and -**i**, usually with a clear distinction in meaning, -**kan** marking the object as patient and -**i** marking it as location or recipient"

(Sneddon et al., 2010, p. 100)

<< We'll call this the **distinctive hypothesis** >>

Sneddon, J. N., Adelaar, A., Djenar, D. N., & Ewing, M. C. (2010). *Indonesian reference grammar* (2nd ed.). Allen & Unwin.

Clustering split and *nearest neighbours* technique


menyesali
menyesal
membayangkan
mengatai
menggunai
mengata
mewaris
mengantung
mengantungkan
mewariskan
mewarisi
mencontohi
mencontoh
menyusu
menyusui
menyusukan
membuahi
mengakhirkan
mendasari
mendasar
mendasarkan
menandakan
menandai
mengguna
menggunakan
mewakilkan
mewakili
mewakil
mencontohkan
mengatakan
menyesalkan
membayang
membayangi
mengakhir
mengakhiri
melangkahi
mengantungi
menanda
menempati
menempatkan
menempat
membuah
membuahkan
melangkahkan
melangkah
menapak
menapaki
menapakkan
menjejakkan
menjejak
menjejaki
2.0
1.5
1.0
0.5
0.0

Split between –*kan* and –*i*

- *membuah* & *membuahkan* cluster together, excluding *membuahi* (-*kan* & -*i* split)
- *melangkah* & *melangkahkan* cluster together, excluding *melangkahi* (-*kan* & -*i* split)
- *membayang* & *membayangi* cluster together, excluding *membayangkan* (-*kan* & -*i* split)
- *mengakhir* & *mengakhiri* cluster together, excluding *mengakhirkan* (-*kan* & -*i* split)
- *menyesal* & *menyesali* cluster together, excluding *menyesalkan* (-*kan* & -*i* split)
- *mencontoh* & *mencontohi* cluster together, excluding *mencontohkan* (-*kan* & -*i* split)
- *mewarisi* & *mewariskan* cluster together, excluding *mewaris* (-*kan/-i* derivatives are split from the base *meN*-)
- *menandai* & *menandakan* cluster together, excluding *menanda* (-*kan/-i* derivatives are split from the base *meN*-; orthographical issue for *menanda*)
- *mengatai, mengatakan,* & *mengata* are all split (orthographical issue for *mengata*)

Nearest neighbours

- Retrieve a set of words showing highest similarity in their vector distribution (i.e. similar contextual environment) to the target verbs
  - Reflected in *cosine similarity* scores
  - The higher their cosine similarities, the more similar those words, and hence the "closer" they are in the "semantic space"
- To contrast *–kan* & *-i* split
- To resolve orthographic anomalies

- Rajeg, G. P. W., Denistia, K., & 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*, *67*, 35–76.
- Schmidt, B. (2015, October 25). Vector Space Models for the Digital Humanities. Retrieved 24 January 2016, from http://bookworm.benschmidt.org/posts/2015-10-25-Word-Embeddings.html

Nearest neighbours: a brief demo in R

Nearest words to *membuahi*

	word	similarity to "membuahi"
1	membuahi	1.0000000
2	dibuahi	0.8372309
3	ovum	0.8330801
4	sperma	0.7965994
5	gamet	0.7326975
6	pembuahan	0.7300101
7	terbuahi	0.7016097
8	spermatozoid	0.6901719
9	spermatozoa	0.6794663
10	parthenogenesis	0.6775866
11	zigot	0.6712285

Nearest words to *membuahkan*

	word	similarity to "membuahkan"
1	membuahkan	1.0000000
2	berbuah	0.6716316
3	mem-buahkan	0.6296921
4	tercipta	0.6214626
5	membuah	0.5991543
6	menuai	0.5729809
7	tendangannya	0.5533927
8	ditepis	0.5530693
9	kerasnya	0.5528045
10	pinalti	0.5482891
11	dimentahkan	0.5478375

Nearest words to *mengatai*

	word	similarity to "mengatai"
1	mengatai	1.0000000
2	memaki	0.7476359
3	marah-marah	0.6818458
4	cerewet	0.6806308
5	mengejek	0.6795740
6	memaki-maki	0.6713904
7	jengkel	0.6675079
8	diejek	0.6645705
9	diolok-olok	0.6641099
10	meledek	0.6628723
11	berbohong	0.6597248

Nearest words to *mengatakan*

	word	similarity to "mengatakan"
1	mengatakan	1.0000000
2	menegaskan	0.8376837
3	menyatakan	0.8318030
4	mengungkapkan	0.8079668
5	mengemukakan	0.7967164
6	menuturkan	0.7925858
7	menjelaskan	0.7808896
8	menyebutkan	0.7640667
9	menerangkan	0.7568715
10	mengakui	0.7495804
11	mengatkan	0.7467988

Conclusion

- The changing faces of 21st century linguistics due to technological advances
  - data explosion
  - programming packages for data science
  - more and more interdisciplinary
- Leverage such advances for data-driven study on the lexical semantics-morphology interface in Indonesian
  - empirical evidence for claim on semantic (dis)similarity between some *–kan*/*-i* verb pairs (via cluster analysis)
  - enriched semantic, syntactic and morphological insights via nearest neighbours

21st century linguistics for the next generation of Indonesian linguists

- Data is here; don't shy away
  - Very important to know how to use the data for our research questions—this requires some creative thinking
- Quantitative thinking & coding skills are indispensable
  - Not necessarily for **doing** the research
  - But also, most importantly, for **understanding** other's research in our field; we're in the age of Quantitative Turn & we'll never go back!
- Quantitative-qualitative dualism is two faces of the same coin
  - Qualitative ideas (e.g. distributional semantics) can be operationalised in quantitative terms!
- 21st-century linguistics curriculum in Indonesia needs to incorporate all these advances!

THANK YOU FOR YOUR ATTENTION!

slides: https://bit.ly/fib-fretalk

Acknowledgement:

- Monash International Postgraduate Research Scholarship (MIPRS)
- Monash Graduate Scholarship (MGS)
- Monash Graduate Education Travel Grant
- Philip Chan – MonARCH High Performance Computer Cluster at Monash University
- Indonesian Endowment Fund for Education (LPDP) – awarded to Karlina Denistia
- Participants of ISMIL 22 (2018), at Univ. of California, Los Angeles
- Two anonymous reviewers in NUSA
- Hiroki Nomoto and David Moeljadi (eds. for NUSA special issue: *Linguistic Studies using Large Annotated Corpora*)