What does formal language theory tell us about the nature of reduplication?

Yang Wang

UCLA

yangwangx@g.ucla.edu

Formal Language Theory in Morphology and Phonology

2024 LSA organized session

January 6th, 2024

Reduplication

Morpho-phonological "copying" → identity-based patterns on the surface

- **Meaning-changing operations**

Dyirbal plurals (Pama-Nyungan; North Queensland)

midi → midi~midi

Glosses: *little; small* → *lots of little ones*

Agta plurals (Austronesian; Philippines)

labáng → lab~labáng

patch → PL-*patch*

Reduplication

Morpho-phonological "copying" → identity-based patterns on the surface

- **Semantics-free**

Tagalog pseudo-reduplication (Austronesian; Philippines)

patpát *pat

"N: stick; piece of split bamboo"

Reduplication-like representation for pseudo-reduplicated words (Zuraw, 2002)
→ supported by a MEG study on visual inputs (Wray et al., 2022)

Why reduplication

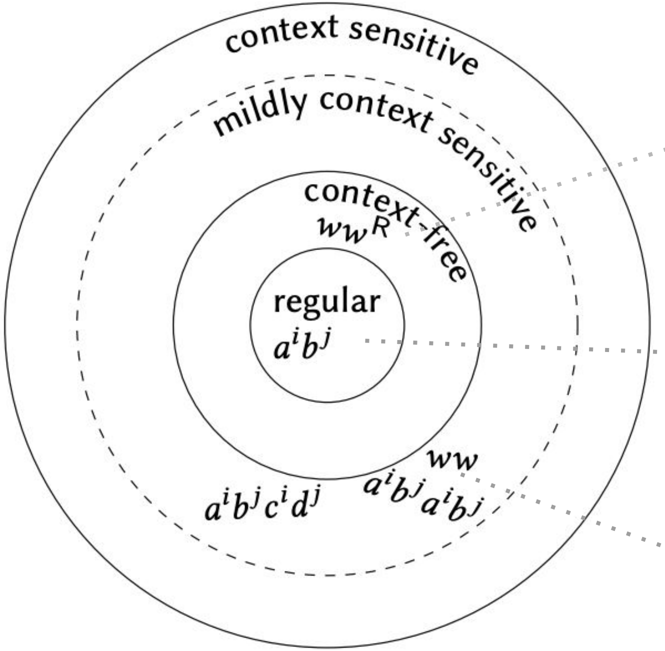
- **Well-attested** in natural languages, typologically rich with many sub-types

  Theoretical proposals for morpho-phonology (e.g., Marantz, 1982; McCarthy & Prince, 1986; Steriade, 1988; Gafos, 1998; Raimy, 2000; Inkelas & Zoll, 2005; Kiparsky, 2010; McCarthy et al., 2012; Zimmermann, 2021)

- Experimental works suggest humans are **highly sensitive** to identity-based patterns (e.g., Marcus et al., 1999; Gerken, 2006; Marcus et al., 2007; Kovács & Mehler, J., 2009a, 2009b; Gerken, 2010; Gervain et al., 2012; Wray et al., 2022; Gallagher, 2013; Berent et al., 2016, 2017; Moreton et al., 2021; Wang & Wilson, In prep) and reduplicative patterns aid speech segmentation (Ota & Skarabela, 2018) and facilitate lexical learning (Ota & Skarabela, 2016) .

- A **long-standing challenge** for formal languages theory and computational modeling (but see Frank & Tenenbaum, 2011; Berent et al, 2012; Prickett et al., 2022; Dolatian & Heinz, 2020; Beguš, 2021; Beguš & Zhou, 2022)

The puzzle of reduplication



String reversal

- rare, confined to language games (Marantz, 1982; Bagemihl, 1989)
- use explicit reasoning but not implicit linguistic knowledge (Moreton et al., 2021)

Most phonology & morphology
(e.g., Johnson, 1972; Kaplan & Kay, 1994; Heinz, 2007; Chandlee, 2014; Chandlee, 2017)

Unbounded copying & surface repetitions

- well-attested (e.g., Moravcsik, 1978; Rubino, 2013)
- use implicit linguistic knowledge (Moreton et al., 2021)

Question 1:

How can we fit in reduplication with the rest of the (morpho-)phonology while excluding some unattested context-free patterns, such as reversals?

Copying, but not reversal

As a **morphological generation** process, $w \rightarrow ww$
(Dolatian & Heinz, 2018, 2019, 2020)

- Classifying the computation of the reduplicative typology based on **2-way (D-)Finite-state transducers**
  - **1-way** FST: only right movement along the input
  - **2-way** FST: move left and right along the input
  - **2-way rotating** FST: do not output anything while moving right-to-left $\rightarrow$ no reversal

Copying, but not reversal

As a morphological generation process, $w \rightarrow ww$
(Dolatian & Heinz, 2018, 2019, 2020)

- Classifying the computational properties of reduplicative typology based on **2-way (D-)Finite-state transducers**
  - **1-way** FST: only right movement along the input
  - **2-way** FST: move left and right along the input
  - **2-way rotating** FST: do not output anything while moving right-to-left → no reversal

Morphological analysis $ww \rightarrow w$?

String-set problem

Namely, the computational properties of the surface phonological forms created by reduplication?

Copying, but not reversal

As a morphological generation process, $w \to ww$
(Dolatian & Heinz, 2018, 2019, 2020)

- Classifying the computational properties of reduplicative typology based on **2-way (D-)Finite-state transducers**
  - **1-way** FST: only right movement along the input
  - **2-way** FST: move left and right along the input
  - **2-way rotating** FST: do not output anything while moving right-to-left → no reversal

Morphological analysis $ww \to w$?

String-set problem

Namely, the computational properties of the surface phonological forms created?

My proposal: a formal characterization of regular languages (most phonology and morphophonology) and languages derived from them through a primitive copying operation.

People wanted such a proposal long ago….

We do not know whether there exists an independent characterization of ***the class of languages that includes the regular sets and languages derivable from them through reduplication,****. . . this class might be relevant to the characterization of NL [natural language] word-sets.*

(Gazdar & Pullum 1985, p 258).

Rather than grudgingly clambering up the Chomsky Hierarchy towards Context-sensitive Grammars, we should consider going back down to ***Regular Grammars*** *and striking out in a different direction. The simplest alternative proposal is a class of grammars which intuitively have the same relation to* ***queues*** *that CFGs have to stacks.*

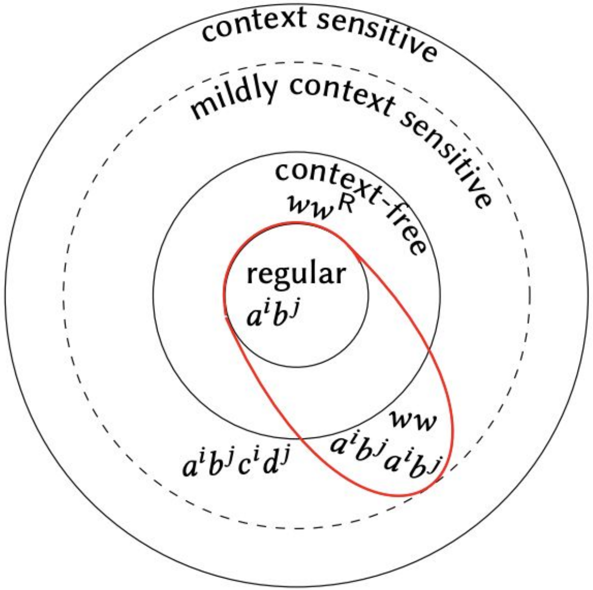
(Manaster-Ramer 1986, p 87)

Proposal: Finite-state buffered machines

A finite-state automata + a copying mechanism (Wang & Hunter, 2023)

- Unbounded memory buffer, with queue storage and restricted ways of interaction with the input
  - same alphabet
  - once one symbol is removed, everything else must also be emptied
- Two modalities
  - Normal mode: similar to a normal FSA
  - Buffering mode: storing a copy of input symbols to the buffer
- Two special sets of states (indicating when to copy, and when to end)

A proposed formal class (Wang 2021 a, b; Wang & Hunter, 2023)



Surveyed closure properties	Closed ?
union	✔
concatenation	✔
Kleene star	✔
homomorphism	✔
Intersection with regular languages	✔
Inverse homomorphism	✘
Recursive copying	✘
intersection	✘
complementation	✘

On regular copying languages

Yang Wang and Tim Hunter
University of California, Los Angeles

ABSTRACT

This paper proposes a formal model of regular languages enriched with unbounded copying. We augment finite-state machinery with the ability to recognize copied strings by adding an unbounded memory buffer with a restricted form of first-in-first-out storage. The newly introduced computational device, finite-state buffered machines (FS-BMs), characterizes the class of regular languages and languages derived from them through a primitive copying operation. We name this language class *regular copying languages* (RCLs). We prove a pumping lemma and examine the closure properties of this language class. As suggested by previous literature (Gazdar and Pullum 1985, p.278), regular copying languages should approach the correct characterization of natural language word sets.

Keywords:
reduplication,
copying,
finite-state machinery,
queue automata

Question 2:

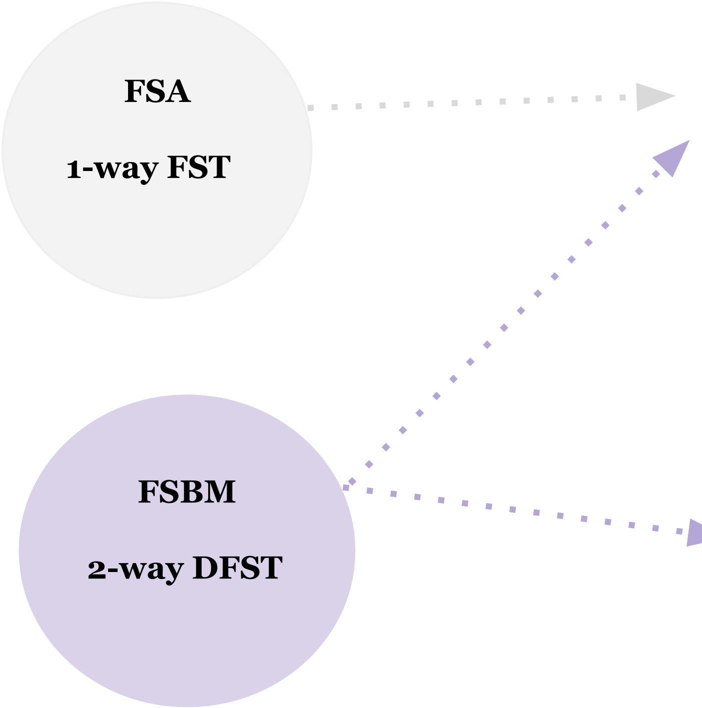
Now that reversals are excluded, does the enriched power overshoot? Should “regular” be sufficient?

Copying, or not?

Typology says **yes** for the enriched power.

Now let us see some **evidence** from a **learning** experiment.

Partial vs. total reduplication



Partial reduplication, or bounded copying

(Chandlee & Heinz, 2017)

Agta plurals (Austronesian; Philippines)

labáng → lab~labáng

patch → *PL-patch*

Total reduplication, or unbounded copying

Dyirbal plurals (Pama-Nyungan; North Queensland)

midi → midi~midi

little; small → *lots of little ones*

Hypotheses and predictions

	Total	**Partial**
FSA || 1-way FST	✘	✔
FSBM || 2-way FST	✔	✔

When people are prompted with input data that conform to both **total reduplication** and **partial reduplication**....

If we see people choose **total** over **partial** ⇒ yes for the extra copying operation

If we see people choose **partial** over **total** ⇒ inconclusive

Extrapolation paradigm (aka. *Poverty of the stimulus paradigm*; Wilson, 2006)

Training phase: impoverished inputs, compatible with many possible hypotheses.

4 singular-plural pairs, where singulars are monosyllabic CVC nominals

dug → dug~dug

Copy the full word?

…… (e.g., feature-based template based on shared features at each slot; listed allomorphy, etc)

Copy a CVC form?

Extrapolation paradigm (aka. *Poverty of the stimulus paradigm*; Wilson, 2006)

Testing phase: trials that can tease apart these different hypotheses

20 novel singulars (5 testing types; 4 for each type)

1.	CVC	Familiar	ˈnoʊg
2.	CV.CVC	Disyllabic CV	ˈpa.dis
3.	CVC.CVC	Disyllabic CVC	ˈdɛb.gɪv
4.	CV.CV.CVC	Trisyllabic	ˈteɪ.pə.gæb
5.	CV.CV.CV.CV.CVC	Five syllables	ˌgɛ.zə.ˈseɪ.kə.dɪv

Procedure

Training phase

- Participants were instructed to learn plural formation (pictures for semantic support)
- Listen to 4 singular ~ plural pairs
- Repeat the singular and the plural form

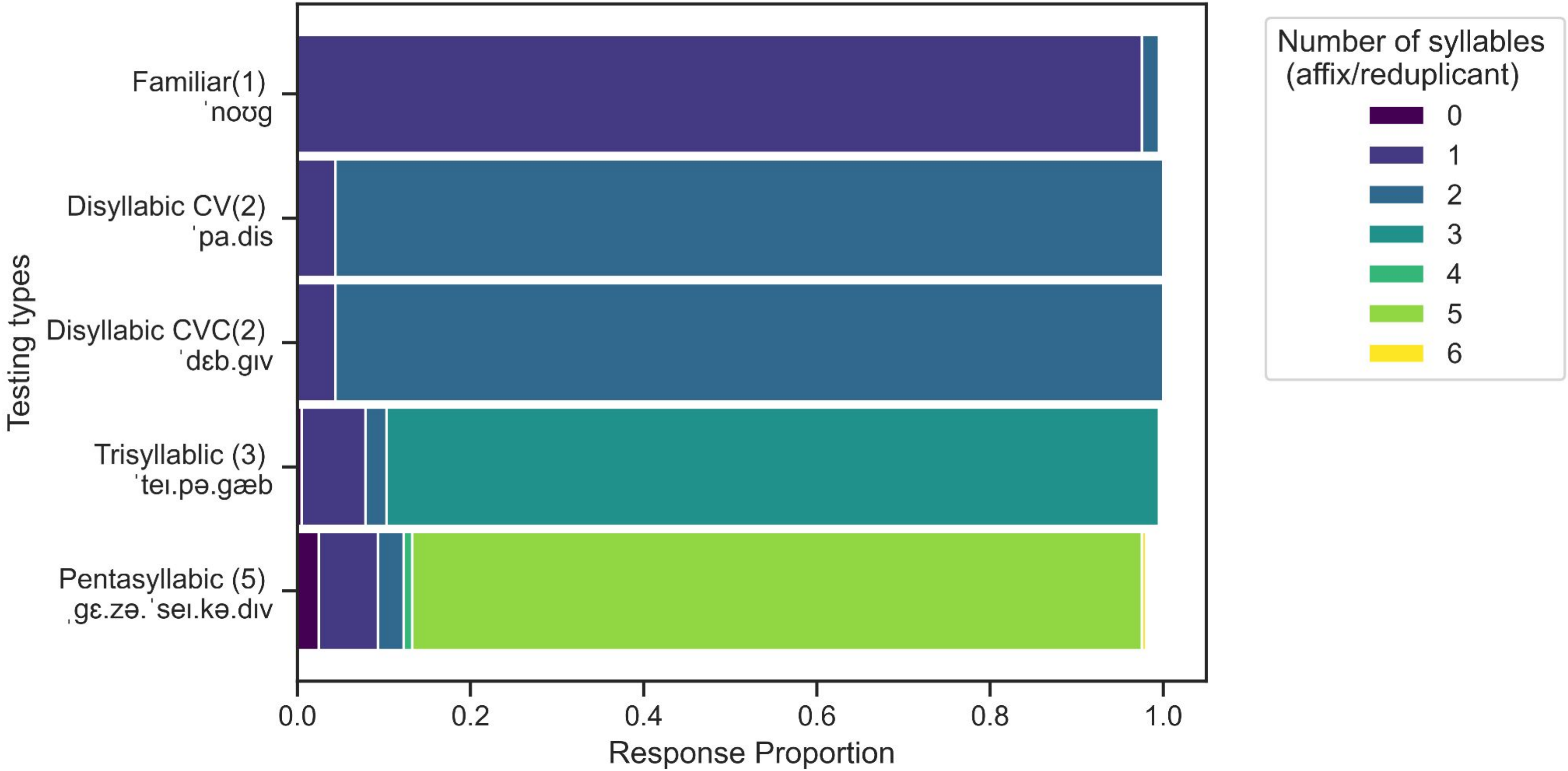
Testing phase:

- Listen to a novel singular, repeat back, and produce the plural form
- All trial types tested together, order randomized

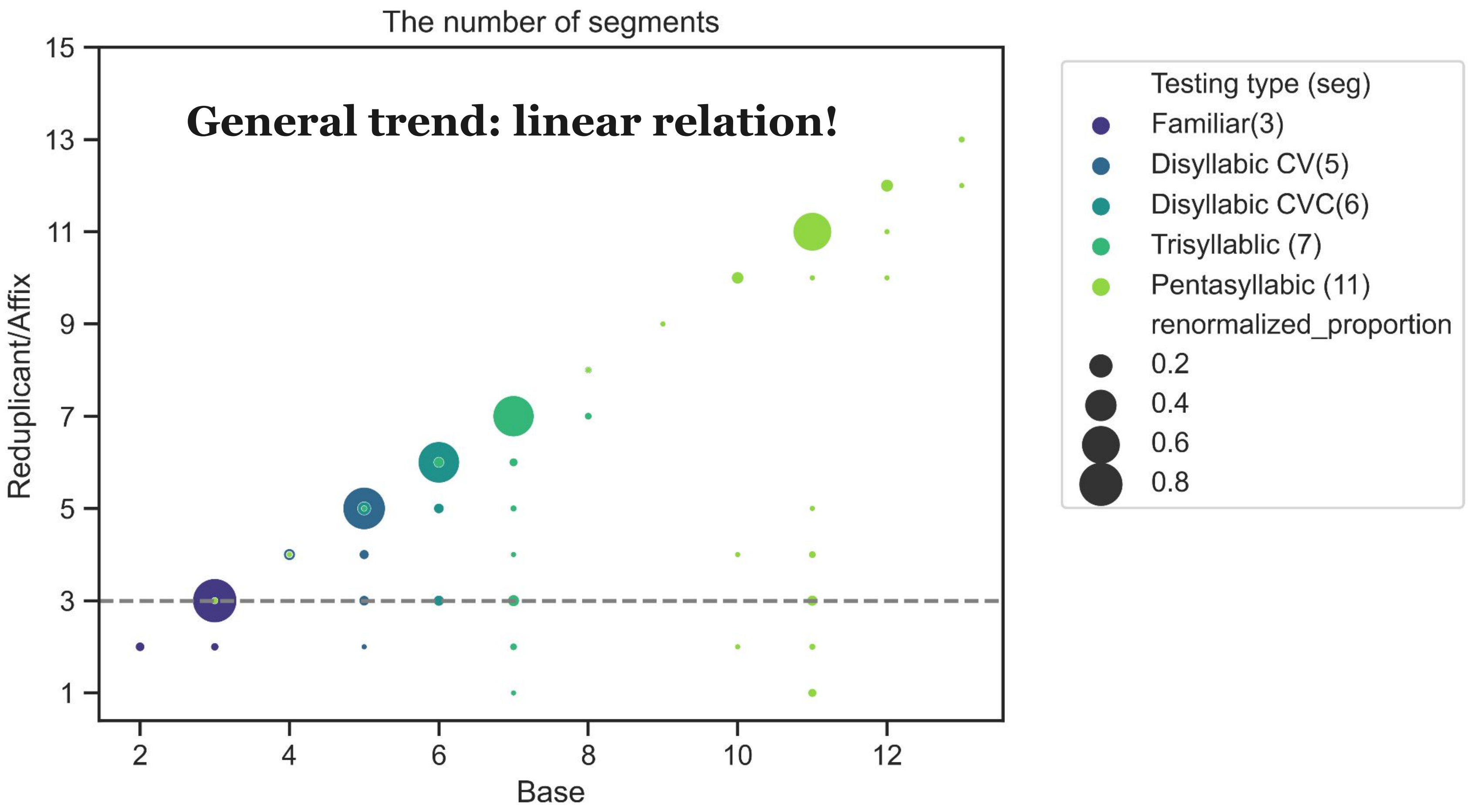
Participants

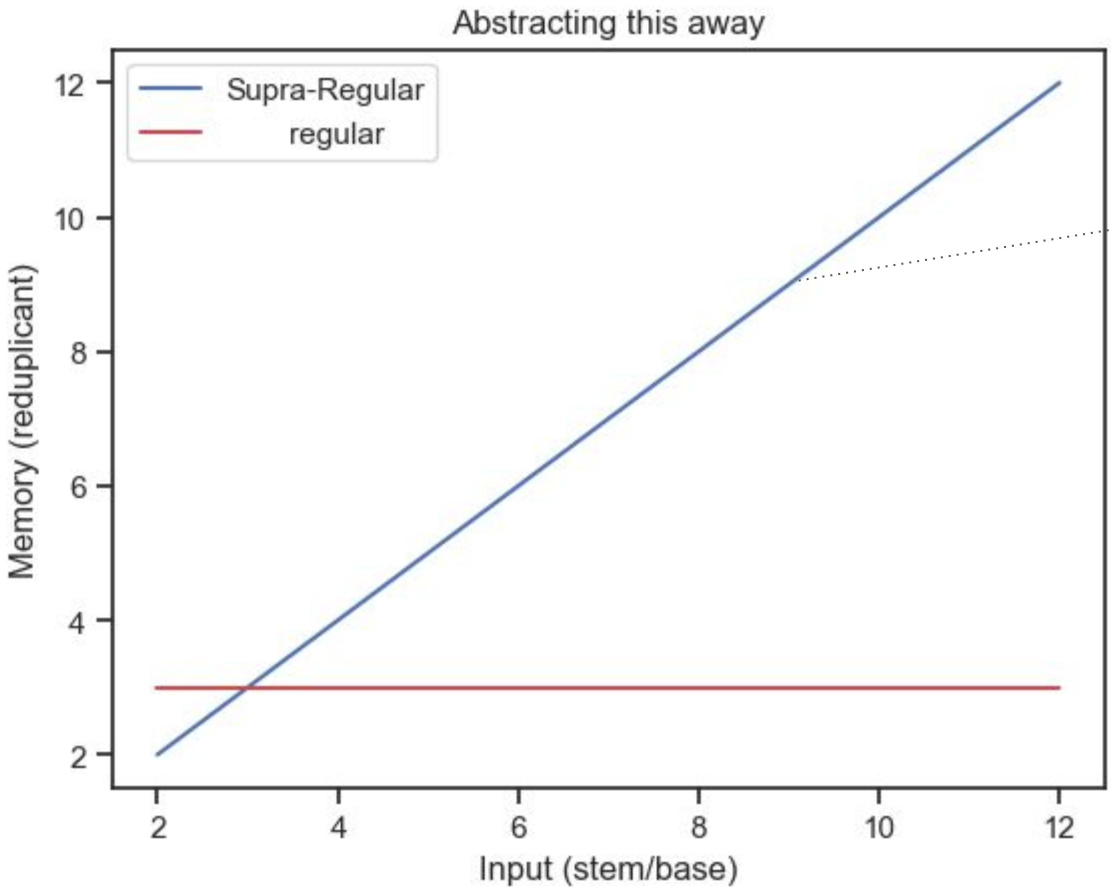
52 US-based English speakers were recruited from Prolific (51 were analyzed)

- 1 exclusion due to silent responses for all training trials and 90% of the testing trials
- 29 Female; 21 Male; 1 Other
- Age: mean = 42.48; max = 72; min = 19
- Screening on prolific: English monolingual; primary language: English; no language-related disorders.



Familiar(1)
ˈnoʊg
Disyllabic CV(2)
ˈpa.dis
Disyllabic CVC(2)
ˈdɛb.gɪv
Trisyllablic (3)
ˈteɪ.pə.gæb
Pentasyllablic (5)
ˌgɛ.zə.ˈseɪ.kə.dɪv
Testing types
0.0
0.2
0.4
0.6
0.8
1.0
Response Proportion
Number of syllables
(affix/reduplicant)
0
1
2
3
4
5
6


The number of segments
General trend: linear relation!
Reduplicant/Affix
Base
15
13
11
9
7
5
3
1
2
4
6
8
10
12
Testing type (seg)
Familiar(3)
Disyllabic CV(5)
Disyllabic CVC(6)
Trisyllablic (7)
Pentasyllabic (11)
renormalized_proportion
0.2
0.4
0.6
0.8



This is what has been learned!

Hypotheses and predictions

	Total	**Partial**
FSA || 1-way FST	✘	✔
FSBM || 2-way FST	✔	✔

When people are prompted with input data that conform to both **total reduplication** and **partial reduplication**....

If we see people choose **total** over **partial** ⇒ yes for the extra copying operation

If we see people choose **partial** over **total** ⇒ inconclusive

Question 3:

How is reduplication learned?

From grammatical knowledge to learning biases

The presented experiment addresses two levels of questions.

1. Whether the extra power to copy should be there in the grammatical knowledge?

   Answer: yes

2. Is there any biases that guide the learner through the learning process?

   Answer: An inductive bias that prefers total reduplication over a segment-based length restricting hypothesis (at least for pluralization)

What about other attested patterns?

Extrapolation paradigm (aka. *Poverty of the stimulus paradigm*; Wilson, 2006)

Training phase: “impoverished” inputs, compatible with many possible hypotheses.

4 singular-plural pairs, where singular are monosyllabic CVC nominals

dug → du~dug

Delete the last coda?

……(e.g., copy a CV of a final syllable/stressed syllable)

Copy an initial CV?

Participants

- 36 US-based English speakers were recruited from Prolific (more replication in progress)

  ▸ 1 participant were excluded due to failure to follow the experimental instruction by giving English words of the pictures

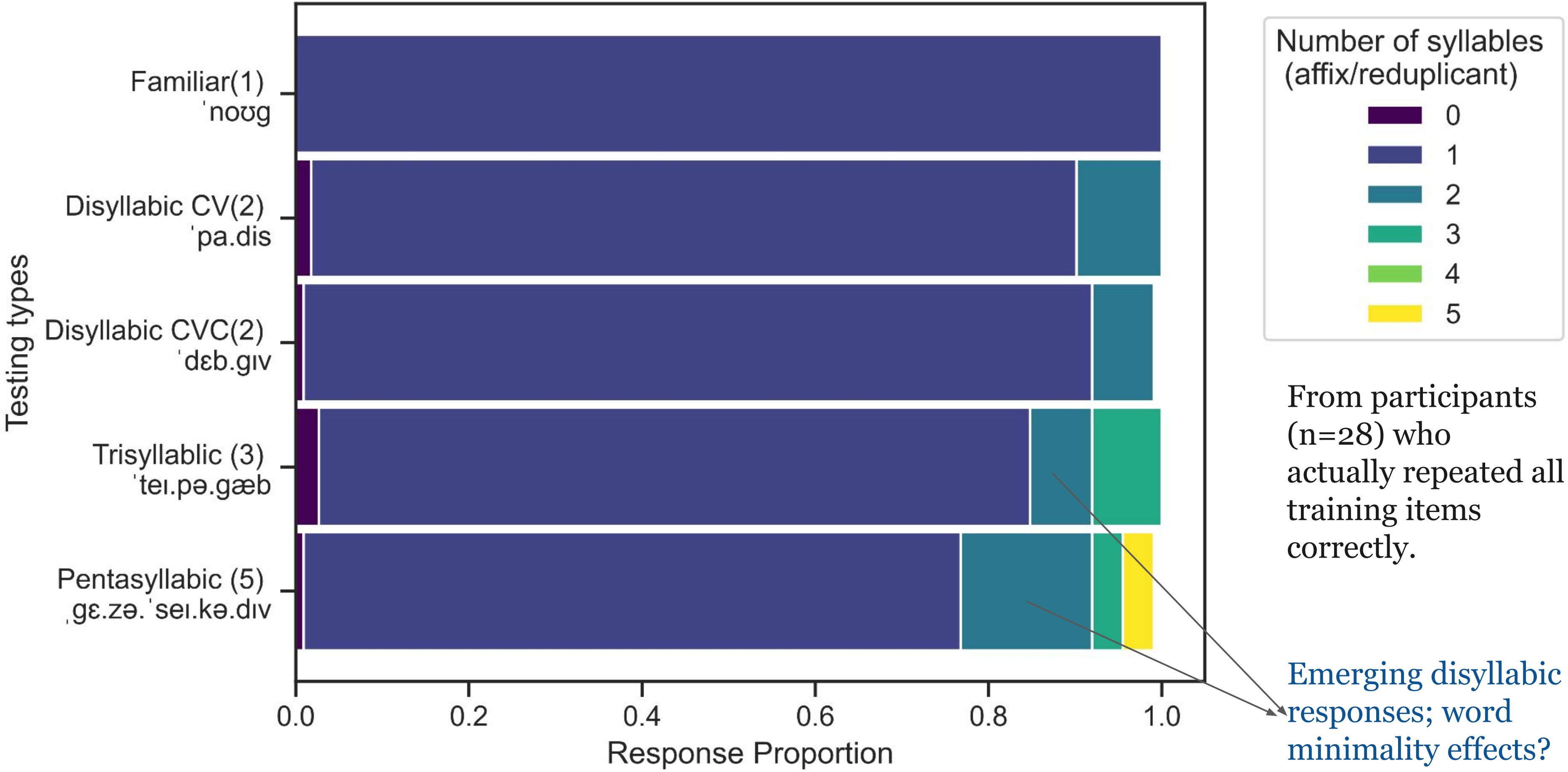
  ▸ 2 participants were excluded because of exposure to languages with attested grammatical reduplication (Hebrew and Japanese)

- Data from 33 participants were analyzed

  ▸ 20 Female; 12 Male; 1 Other

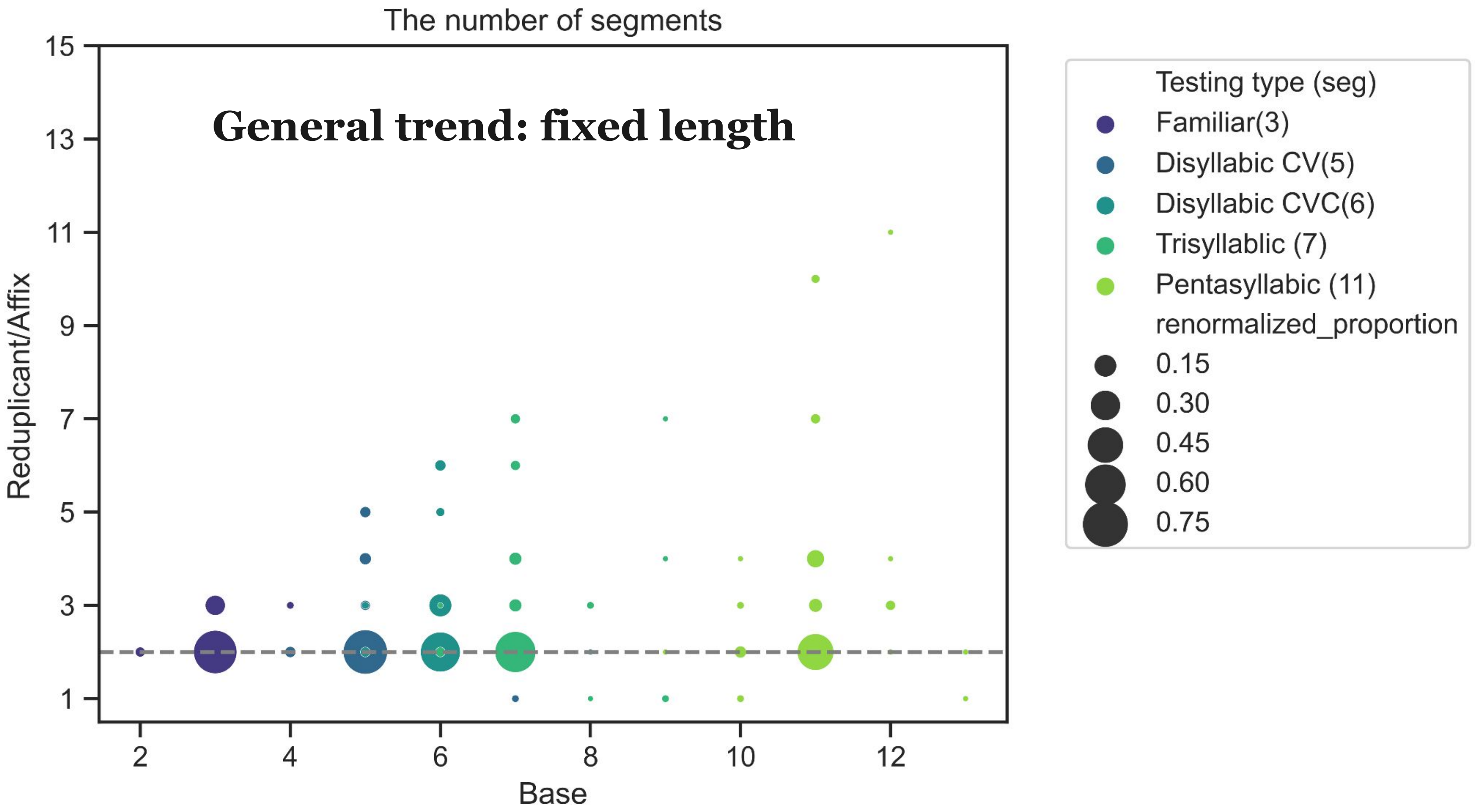
  ▸ Age: mean = 39.30; max = 68; min = 18

  ▸ Screening on prolific: English monolingual; primary language: English; no language-related disorders



From participants (n=28) who actually repeated all training items correctly.

Emerging disyllabic responses; word minimality effects?


The number of segments
General trend: fixed length
Reduplicant/Affix
Base
15
13
11
9
7
5
3
1
2
4
6
8
10
12
Testing type (seg)
Familiar(3)
Disyllabic CV(5)
Disyllabic CVC(6)
Trisyllablic (7)
Pentasyllablic (11)
renormalized_proportion
0.15
0.30
0.45
0.60
0.75

Take-home message

- A computational device for natural language morphology and phonology requires a kind of the **primitive copying** operation.

- When trained with complete monosyllabic copying, people generalize total reduplication to longer forms, but not heavy syllable copying.

- Incomplete copying leads to more responses of invariant shapes (the one that gets trained on), with some emerging typological variations.

    - (ask me about this if you want to hear more!)

Concluding remarks

Formal language theory

Typology

Empirical testing

Thank you!

Contact: *yangwangx@g.ucla.edu*

Thanks to Scott, Jon and Jeff for organizing this session. Also to Dylan Bumford, Hossep Dolatian, Bruce Hayes, Tim Hunter, Claire Moore-Cantwell, Iza Sola-Llonch, Megha Sundara, Colin Wilson, Lily Xu, Sam Zukoff, Kie Zuraw, and the attendees of UCLA Phonology seminar and {Psycho, Comp}+Ling seminar for their help, comments, and insights for various bits covered in this presentation. Also to Mariana Cui, Jacob Hanna, Jenessa Lathrop, Shawdi Sani, Alexandria Zarko and Boyi Zheng, for data transcription and annotation. This project is supported by UCLA Linguistics Research Funds and a Dissertation Year Fellowship from UCLA Graduate Division.

Selected Reference I

Berent, I., Bat-El, O., & Vaknin-Nusbaum, V. (2017). The double identity of doubling: Evidence for the phonology–morphology split. *Cognition*, 161, 117-128.

Dawson, C., & Gerken L. (2009). From domain-generality to domain-sensitivity: 4-month-olds learn an abstract repetition rule in music that 7-month-olds do not. *Cognition*, 111(3), 378-82.

Dolatian, H., & Heinz, J. (2020). Computing and classifying reduplication with 2-way finite-state transducers. *Journal of Language Modelling*, 8(1),179–250.

Endress, A.D., Dehaene-Lambertz, G., & Mehler., J. (2007). Perceptual constraints and the learnability of simple grammars. *Cognition*, 105(3), 577–614.

Endress, A.D., Nespor, M. and Mehler, J. (2009). Perceptual and memory constraints on language acquisition. *Trends in cognitive sciences*, *13*(8), 348-353.

Ferguson, B., & Lew-Williams, C. (2016). Communicative signals support abstract rule learning by 7-month-old infants. *Scientific Reports*, 6, 25434.

Finley, S. & Christiansen, M. (2011). Multimodal transfer of repetition patterns in artificial grammar learning. In *Proceedings of the annual meeting of the Cognitive Science Society*, 33 (33).

Frank, M.C., Slemmer, J.A., Marcus, G.F., & Johnson, S.P. (2009). Information from multiple modalities helps 5-month-olds learn abstract rules. *Developmental Science*, 12(4), 504–509.

Frank, M. C., & Tenenbaum, J. B. (2011). Three ideal observer models for rule learning in simple languages. *Cognition, 120*(3), 360–371

Gallagher, G. (2013). Learning the identity effect as an artificial language: bias and generalisation. *Phonology,* 30(2), 253–295.

Selected Reference II

Gerken, L. (2006). Decisions, decisions: Infant language learning when multiple generalizations are possible. *Cognition, 98*(3), B67-B74.

Gervain, J., Berent, I., & Werker, J. (2012). Binding at birth: Newborns detect identity relations and sequential position in speech. *Journal of Cognitive Neuroscience*, 24 (3), 564 –574.

Haley, C., & Wilson, C. (2021). Deep neural networks easily learn unnatural infixation and reduplication patterns. *Proceedings of the Society for Computation in Linguistics*, Vol 4.

Inkelas, S. & Zoll, C. (2005). *Reduplication: Doubling in morphology* (106). Cambridge University Press.

Kiparsky, P. (2010). Reduplication in Stratal OT. In Uyechi, L., and L. H. Wee (Eds.), *Reality exploration and discovery: pattern interaction in language and life*, 125–42. Stanford: CSLI Publications.

Kovács, Á.M. and Mehler, J. (2009a). Cognitive gains in 7-month-old bilingual infants. *Proceedings of the National Academy of Sciences*, *106*(16), 6556-6560.

Kovács, Á.M. and Mehler, J. (2009b). Flexible learning of multiple speech structures in bilingual infants. *Science*, *325*(5940), 611-612.

Levin, J. (1985). *A Metrical Theory of Syllabicity*. Doctoral dissertation, MIT, Cambridge, MA

Marantz, A. (1982). Re reduplication. *Linguistic Inquiry*, 13(3), 435–482.

Marcus, G.F., Vijayan. S., Rao. S.B. & Vishton, P.M. (1999). Rule learning by seven-month-old infants. *Science 283(5398)*, 77-80.

Marcus, G. F., Fernandes, K. J., & Johnson, S. P. (2007). Infant rule learning facilitated by speech. *Psychological Science, 18*(5), 387–391.

Selected Reference III

McCarthy, J., & Prince., A. (1986). Prosodic morphology. Ms., University of Massachusetts, Amherst, and Brandeis University, Waltham, Mass.

McCarthy, J., Kimper, W., & Mullin., K. (2012). Reduplication in harmonic serialism. *Morphology*, 22(2), 173–232.

Moravcsik, E. (1978). Reduplicative constructions. In J, Greenberg (Ed), *Universals of Human Language*, 3, 297–334. Stanford University Press, Stanford.

Moreton, E., Prickett, B., Pertsova, K., Fennell. J., Pater, J., & Sanders, L. (2021). Learning repetition, but not syllable reversal. In: R., Bennett, R., Bibbs, M.L., Brinkerhoff, M. J. Kaplan, S., Rich, A.. Rysling, N.. Van Handel, & M.W., Cavallaro (Eds.), *Supplemental Proceedings of the 2020 Annual Meeting on Phonology*.

Nelson, M., Dolatian, H., Rawski, J., & Prickett, B. (2020). Probing RNN Encoder-Decoder Generalization of Subregular Functions using Reduplication. In *Proceedings of the Society for Computation in Linguistics*. Vol. 3.

Prickett, B., Traylor, A., & Pater., J. (2022). Learning reduplication with a neural network that lacks explicit variables. *Journal of Language Modelling, 10(1)*, 1–38.

Rubino, C. (2013). Reduplication. In: M., Dryer & M., Haspelmath (Eds.) WALS Online (v2020.3) [Data set]. Zenodo. (Available online at http://wals.info/chapter/27).

Selected Reference IV

Saffran, J.R., Pollak, S.D., Seibel, R.L., & Shkolnik, A. (2007). Dog is a dog is a dog: Infant rule learning is not specific to language. *Cognition*, 105(3), 669–680.

Steriade, D. (1988). Reduplication and syllable transfer in Sanskrit and elsewhere. *Phonology*, 5(1), 73–155.

Wang, Y. & Hunter, T. (2023). On Regular Copying Languages. *Journal of Language Modelling* 11(1), 1–66.

Wang, Y. & Wilson, C. (In prep). Inductive bias in learning partial reduplication: Evidence from artificial grammar learning.

Wilson, C. (2006). Learning phonology with substantive bias: An experimental and computational study of velar palatalization. *Cognitive science*, *30*(5), 945-982.

Wray, S., Stockall, L., & Marantz, A. (2022). Early Form-Based Morphological Decomposition in Tagalog: MEG Evidence from Reduplication, Infixation, and Circumfixation. *Neurobiology of Language,* 3(2): 235–255.

Zimmermann, E. (2021). Faded copies: Reduplication as distribution of activity. *Glossa*: *a journal of general linguistics*, 6(1), 58.

Appendix: A worked example

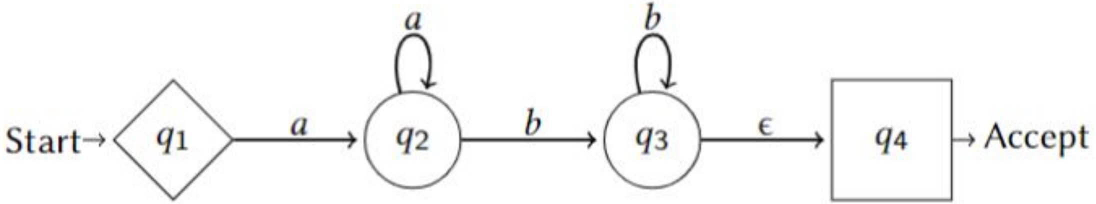


Figure: *An FSBM* M_2 *with* $G = \{q_1\}$ *and* $H = \{q_4\}$. $L(\mathsf{M}_2) = \{a^i b^j a^i b^j \mid i, j \geq 1\}$

Appendix: A worked example

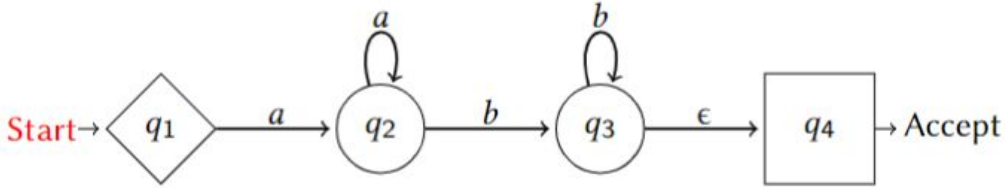


Figure: *An FSBM* M_2 *with* $G = \{q_1\}$ *and* $H = \{q_4\}$. $L(\mathsf{M}_2) = \{a^i b^j a^i b^j \mid i, j \geq 1\}$

input	state	buffer	mode
abbabb	q_1	ϵ	N

Appendix: A worked example

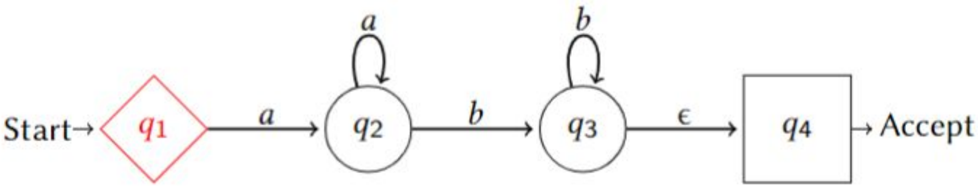


Figure: *An FSBM* M_2 *with* $G = \{q_1\}$ *and* $H = \{q_4\}$. $L(\mathsf{M}_2) = \{a^i b^j a^i b^j \mid i, j \geq 1\}$

input	state	buffer	mode
abbabb	q_1	ϵ	N
abbabb	q_1	ϵ	B

Appendix: A worked example

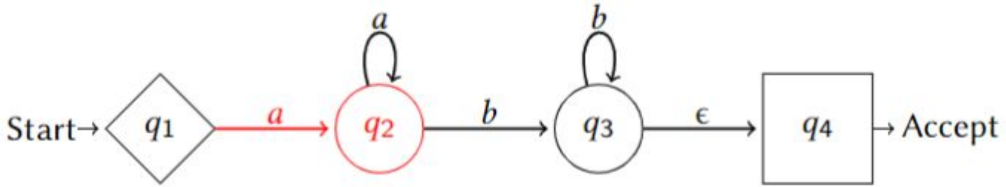


Figure: *An FSBM* M_2 *with* $G = \{q_1\}$ *and* $H = \{q_4\}$. $L(\mathsf{M}_2) = \{a^i b^j a^i b^j \mid i, j \geq 1\}$

input	state	buffer	mode
abbabb	q_1	ϵ	N
abbabb	q_1	ϵ	B
bbabb	q_2	a	B

Appendix: A worked example

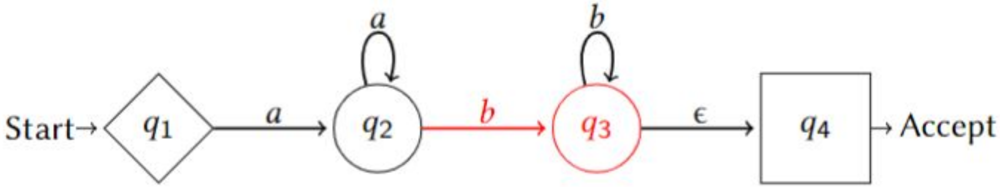


Figure: *An FSBM* M_2 *with* $G = \{q_1\}$ *and* $H = \{q_4\}$. $L(\mathsf{M}_2) = \{a^i b^j a^i b^j \mid i, j \geq 1\}$

input	state	buffer	mode
abbabb	q_1	ϵ	N
abbabb	q_1	ϵ	B
bbabb	q_2	a	B
babb	q_3	ab	B

Appendix: A worked example

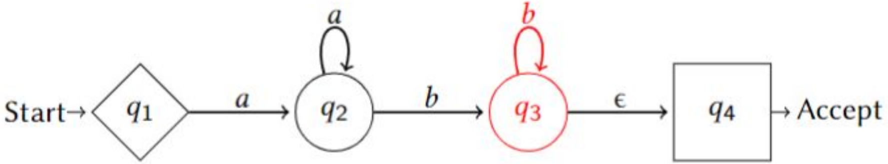


Figure: *An FSBM* M_2 *with* $G = \{q_1\}$ *and* $H = \{q_4\}$. $L(\mathsf{M}_2) = \{a^i b^j a^i b^j \mid i, j \geq 1\}$

input	state	buffer	mode
abbabb	q_1	ϵ	N
abbabb	q_1	ϵ	b
bbabb	q_2	a	B
babb	q_3	ab	B
abb	q_3	abb	B

Appendix: A worked example

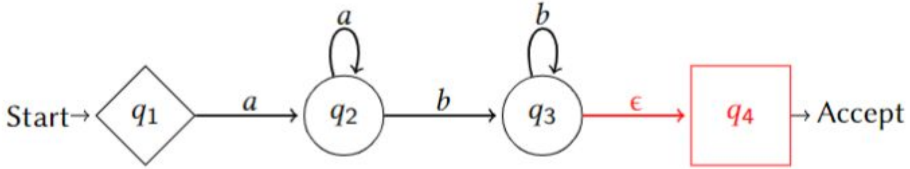


Figure: *An FSBM* M_2 *with* $G = \{q_1\}$ *and* $H = \{q_4\}$. $L(\mathsf{M}_2) = \{a^i b^j a^i b^j \mid i, j \geq 1\}$

input	state	buffer	mode
abbabb	q_1	ϵ	N
abbabb	q_1	ϵ	B
bbabb	q_2	a	B
babb	q_3	ab	B
abb	q_3	abb	B
abb	q_4	abb	B

Appendix: A worked example

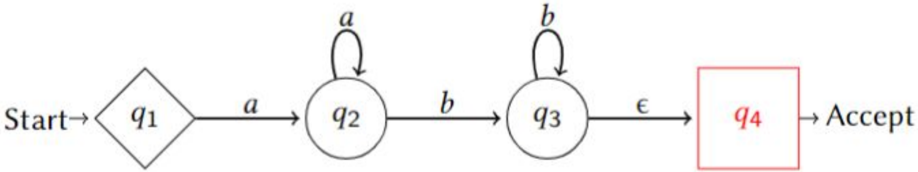


Figure: *An FSBM* M_2 *with* $G = \{q_1\}$ *and* $H = \{q_4\}$. $L(\mathsf{M}_2) = \{a^i b^j a^i b^j \mid i, j \geq 1\}$

input	state	buffer	mode
abbabb	q_1	ϵ	N
abbabb	q_1	ϵ	B
bbabb	q_2	a	B
babb	q_3	ab	B
abb	q_3	abb	B
abb	q_4	abb	B

input	state	buffer	mode
ϵ	q_4	ϵ	N

Appendix: A worked example

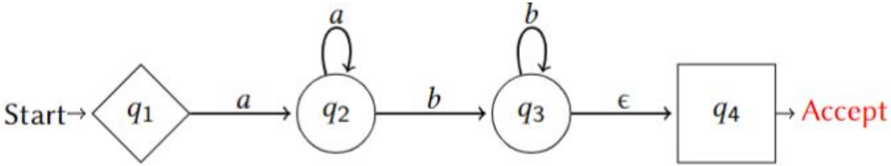


Figure: *An FSBM* M_2 *with* $G = \{q_1\}$ *and* $H = \{q_4\}$. $L(\mathsf{M}_2) = \{a^i b^j a^i b^j \mid i, j \geq 1\}$

input	state	buffer	mode
abbabb	q_1	ϵ	N
abbabb	q_1	ϵ	B
bbabb	q_2	a	B
babb	q_3	ab	B
abb	q_3	abb	B
abb	q_4	abb	B

input	state	buffer	mode
ϵ	q_4	ϵ	N
ACCEPT			

Appendix: Emerging variation

- There are a lot of variations in terms of individual behaviors, which I did not have the time to show.

  ▸ Quite a few instances of no-copying but listed allomorphy: ga-/da- prefixation; or deleting the last coda in the base.

  ▸ Frequently, we observe “the emergence of the unmarked”: vowel reduction in the reduplicant

  ▸ People vary in terms of which part to copy and where to place the reduplicant/affix (see slide 50)

Appendix: Emerging variation

- One par variable total reduplication up to the disyllabic forms; for trisyllabic forms and five syllable word, copy the final syllable

ˈdɛb.gɪv-ˈdɛb.gɪv

ˈteɪ.pə.gæb-gæb

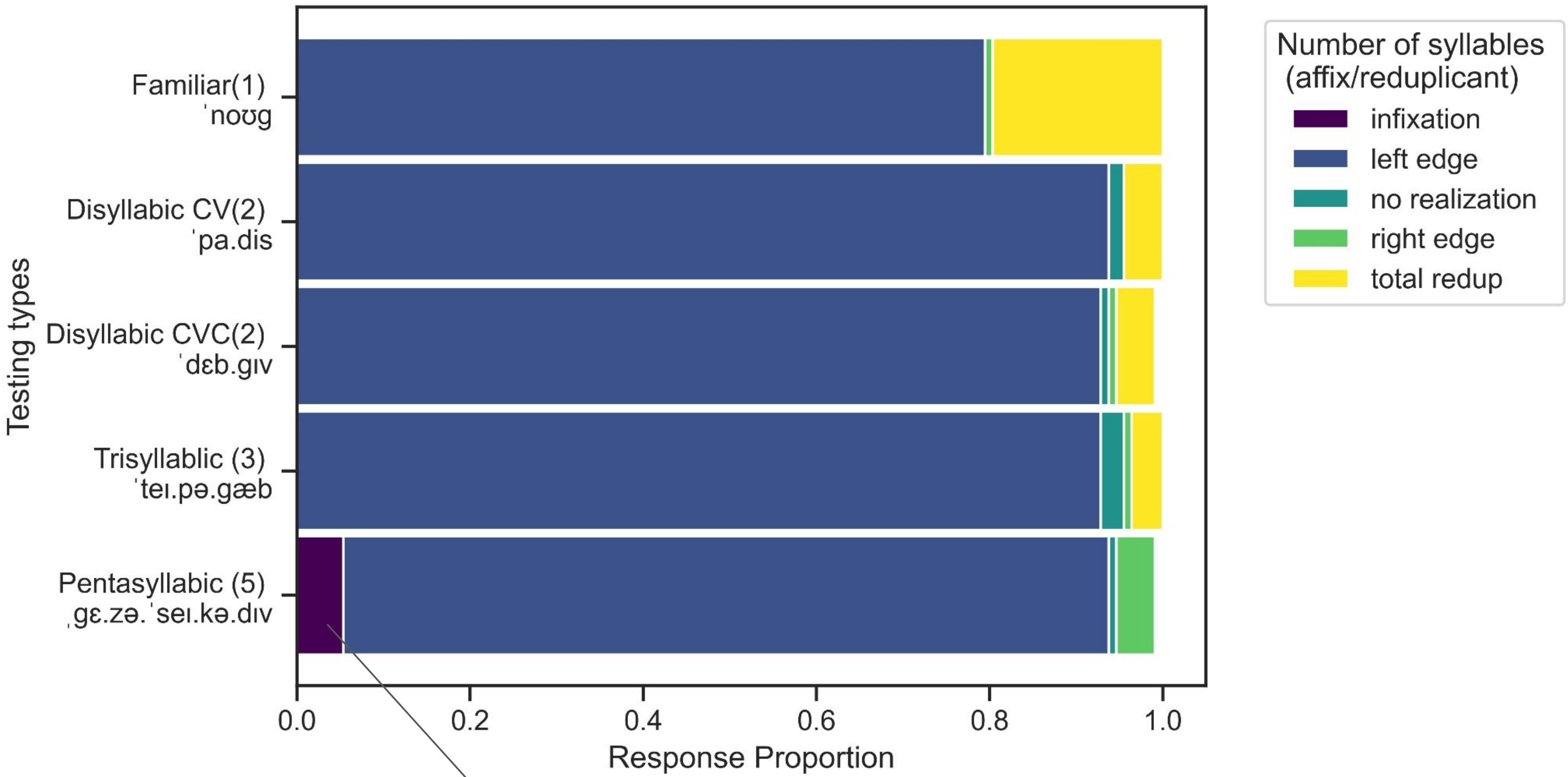
ˌgɛ.zə.ˈseɪ.kə.dɪv-dɪv

- One participant shows 3 instances of truncating both copies when the stem is long (templatic back-copying?)

ˈdi.zə.gɛd → diz-diz

ˌkɪpəˈzudətɛf→ kɪp-kɪp

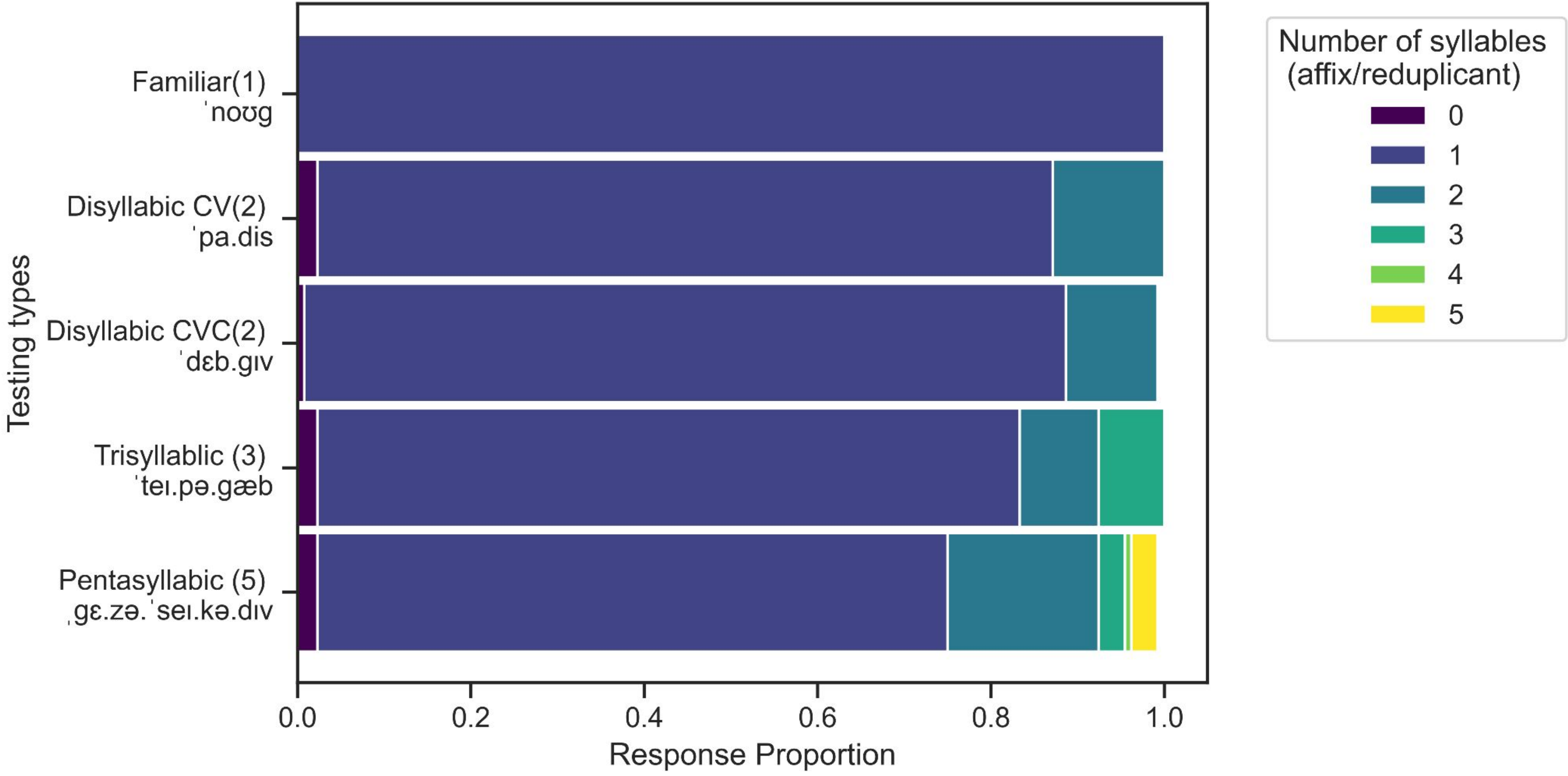
ˌpi.sə.ˈgoʊ.bɛ.kət→ pis-pis

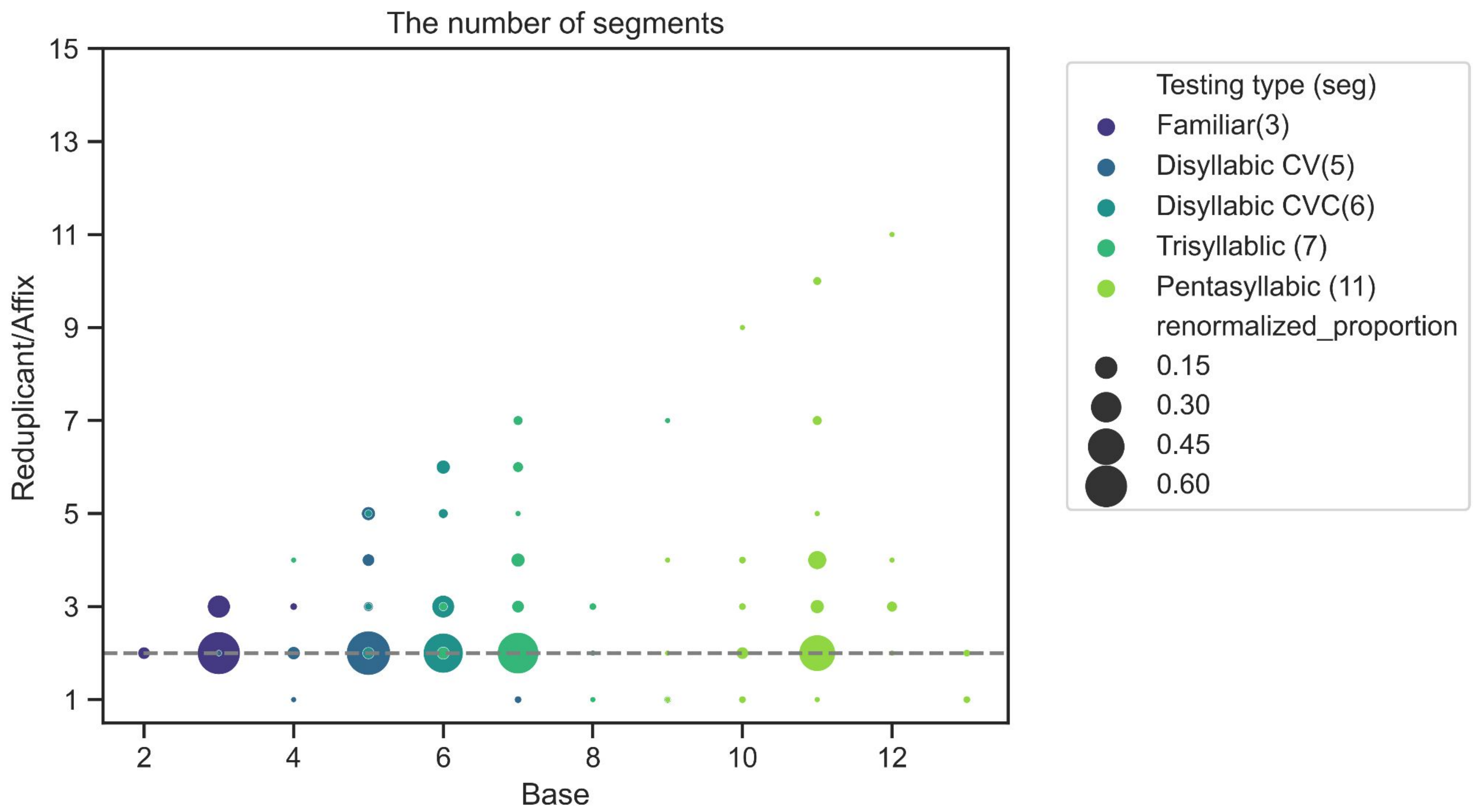


ˌveɪ.kə.-ˈfɑ.zu-ˈfɑ.zu.bɛd

ˌkɪ .pə.-ˈzu.də -ˈzu.də.tɛf

Appendix: all participants (Exp 2)




The number of segments
15
13
11
9
7
5
3
1
Reduplicant/Affix
2
4
6
8
10
12
Base
Testing type (seg)
Familiar(3)
Disyllabic CV(5)
Disyllabic CVC(6)
Trisyllablic (7)
Pentasyllabic (11)
renormalized_proportion
0.15
0.30
0.45
0.60