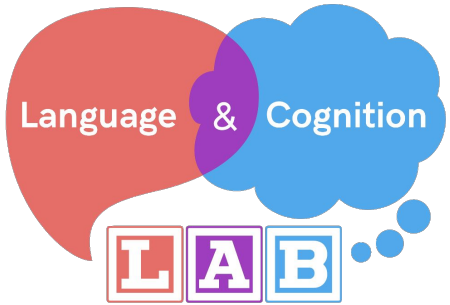
Using Wordbank for Language Development Research

Alvin Tan & Michael C. Frank



board games

meta-analysis

big team science

cross-situational word learning

open science

grounded language

language models

pragmatics

lexical acquisition

social cognition

children's books

???

???

You (probably)

bluegrass

Roadmap

- Introduction
  - CDIs
  - Wordbank
  - What can we learn?
- Practical session
  - `tidyverse` refresher
  - `wordbankr`
  - Case study: Vocabulary and grammar
- What have we learnt? (Research talk – Mike)

Pre-workshop survey

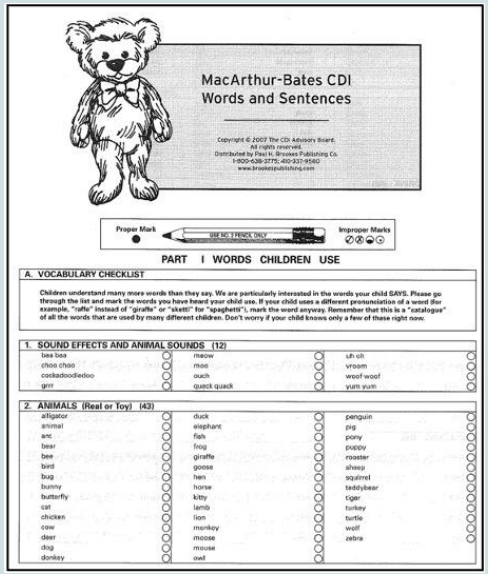
Measuring language development

Direct production measurement
(LENA Team, 2020)

Picture-based vocabulary tests
(Dunn, 2018; Brownell, 2010)

Looking while listening
(Fernald et al., 2008; Byers-Heinlein, 2018)

Measuring language development



Communicative Development Inventories (CDIs)

Caregiver-report checklists measuring early lexical comprehension and production (and gesture use, emerging morphology and syntax…)

Direct

ng
n, 2018)

(Dunn, 2018; Brownell, 2010)

CDIs

- Words & Gestures (8–18mo; receptive and expressive vocabulary)
- Words & Sentences (16–30mo; expressive vocabulary)

English (American) WS

1. *Sound Effects and Animal Sounds* (12)

baa baa	grrr	ouch	vroom
choo choo	meow	quack quack	woof woof
cockadoodledoo	moo	uh oh	yum yum

2. *Animals (Real or Toy)* (43)

alligator	cow	horse	puppy
animal	deer	kitty	rooster
ant	dog	lamb	sheep
bear	donkey	lion	squirrel
bee	duck	monkey	teddybear
bird	elephant	moose	tiger

Why CDIs?

- Comprehensive view of child's abilities
- Less susceptible to child temperament/mood
- Reliable and valid (Jarůšková et al., 2023)

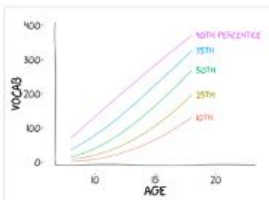
- Inexpensive and easy to administer
- Adaptations in >120 languages/dialects
- Online implementation in Web-CDI (DeMayo et al., 2021)

wordbank.stanford.edu

2022

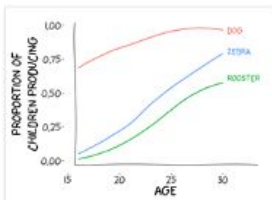
Wordbank

An open database of children's vocabulary development



Vocabulary Norms

Explore vocabulary size growth curves for various languages and demographic groups.



Item Trajectories

Explore trajectories of individual words, word categories, and grammar items.

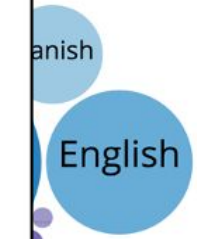
Wordbank contains data from 79,027 children and 93,258 CDI administrations, across 40 languages and 76 instruments:

2019

tains data from 75,144 children
CDI administrations, across
ages and 56 instruments:

2015

l administrations,
3 instruments:



Frank et al. (2017)

Wordbank team

Mike Frank

Virginia Marchman

Mika Braginsky

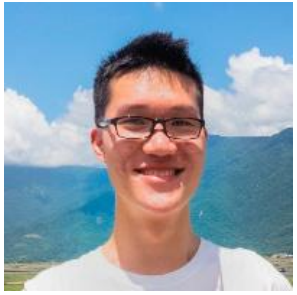
Dan Yurovsky

Danielle Kellier

Benny deMayo

George Kachergis

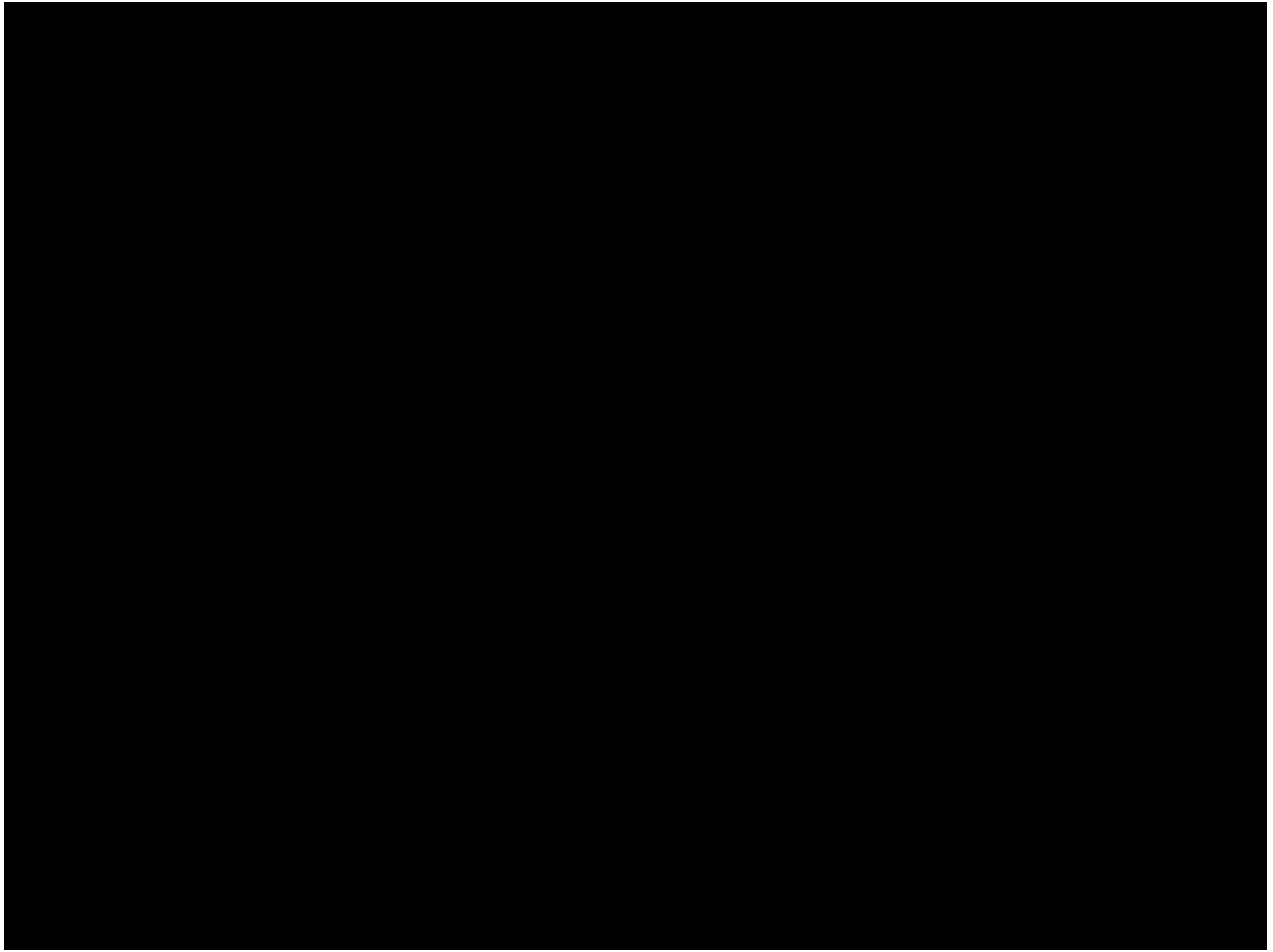
Jess Mankewitz



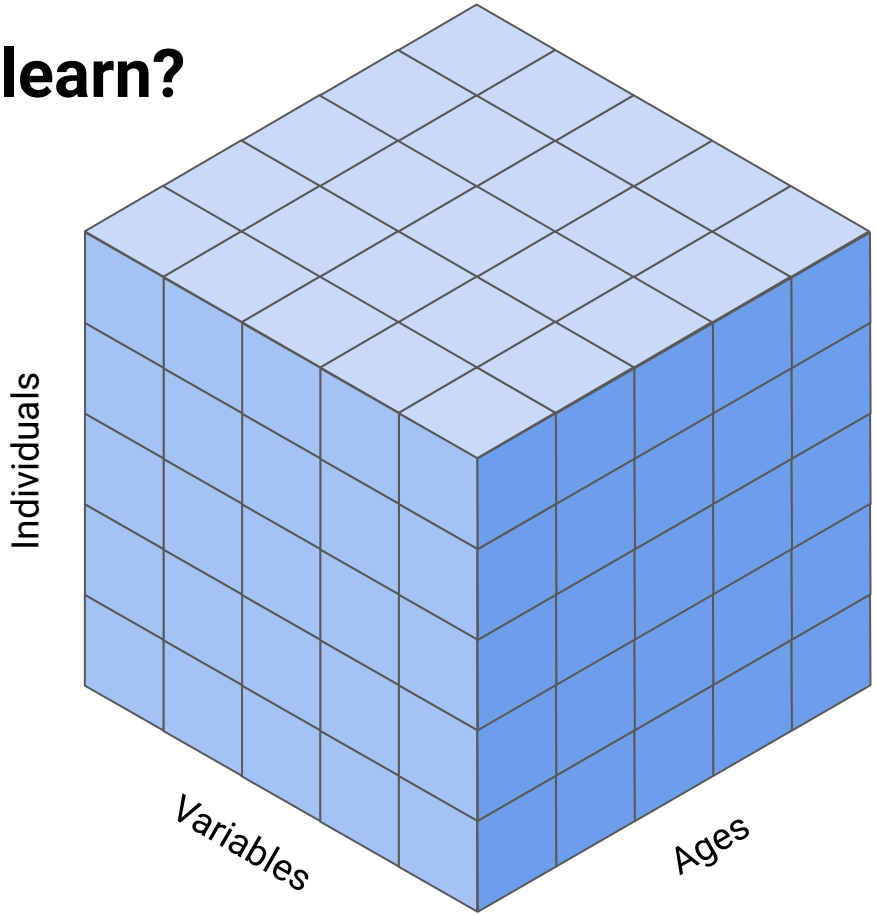
Alvin Tan

Henry Mehta

+ dozens of contributors!

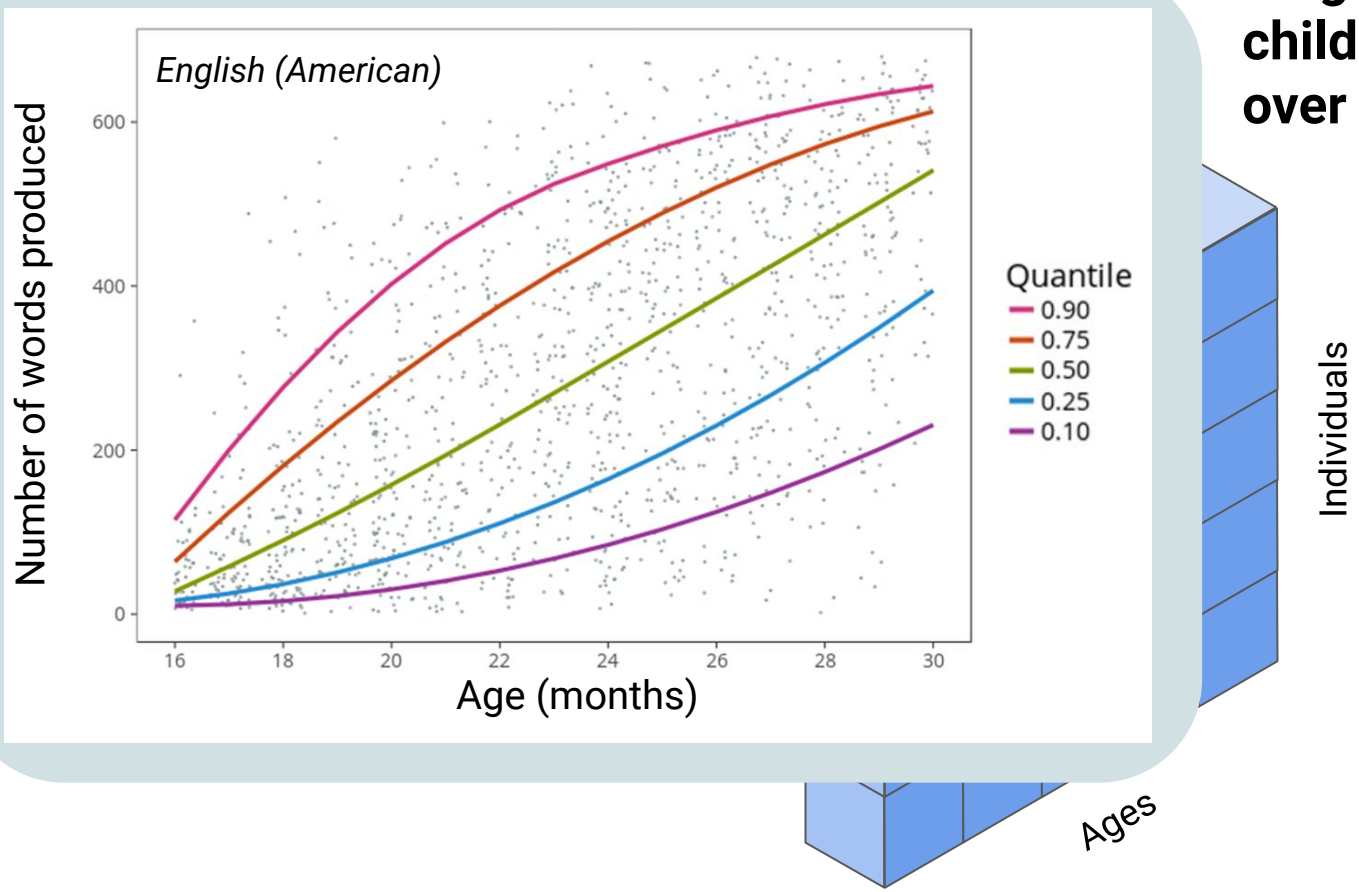


What can we learn?



Cattell (1966)

Large variability in children's vocabularies over development



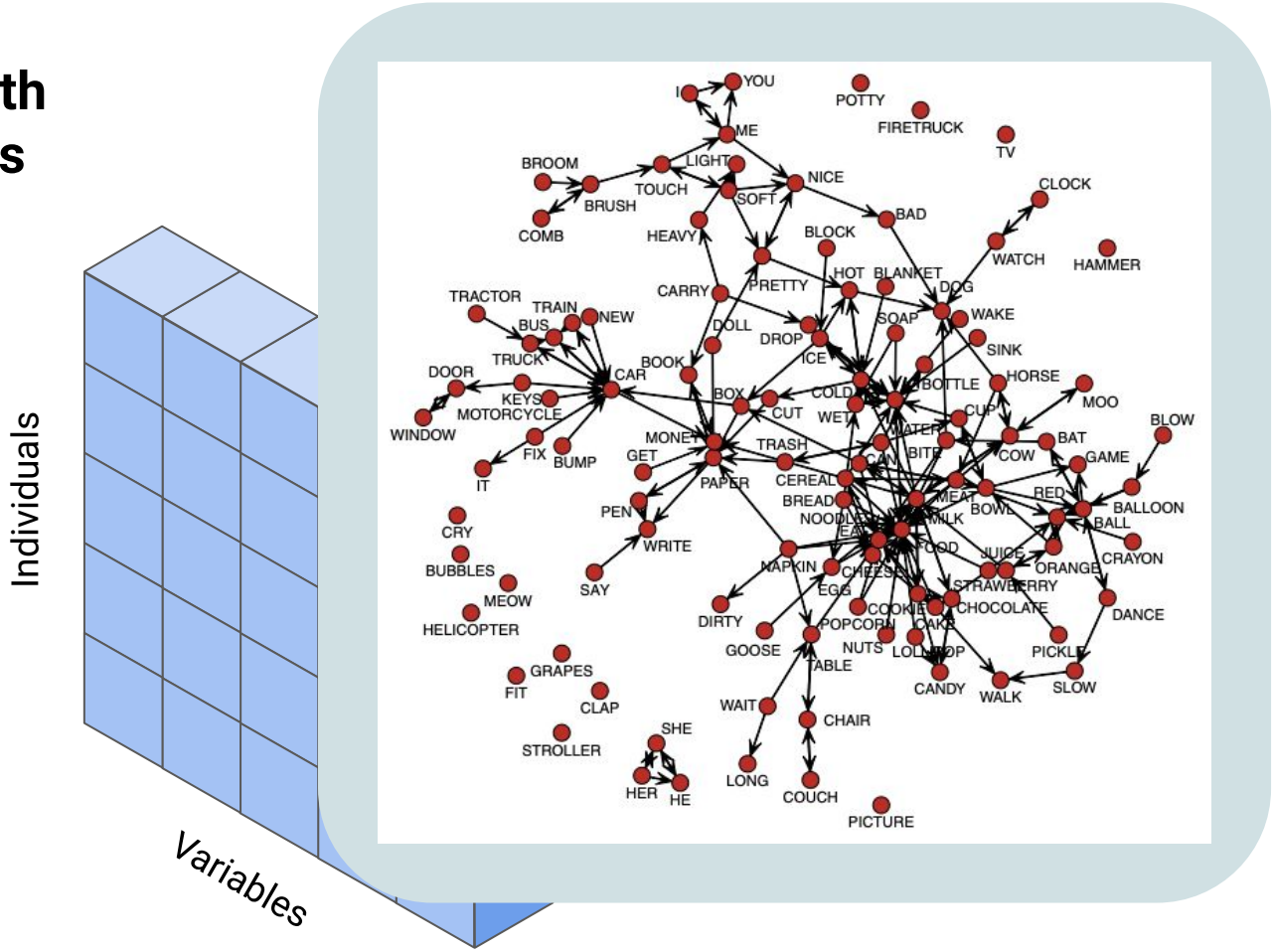
Frank et al. (2021)

	Croatian	Danish	English (American)	Italian	Korean	Norwegian	Russian	Spanish (Mexican)	Swedish	Turkish
1	**grandma**	**peekaboo**	sister	**peekaboo**	**child's name**	**daddy**	**grandma**	milk	**child's name**	**mommy**
2	**mommy**	pattycake	brother	**daddy**	again	**peekaboo**	**daddy**	**bye**	**mommy**	**daddy**
3	**daddy**	**child's name**	**daddy**	**mommy**	money	**mommy**	grandpa	**yum yum**	**daddy**	**yum yum**
4	**bye**	**daddy**	**mommy**	baby food	aunt	**no**	**yum yum**	**daddy**	**peekaboo**	water
5	**child's name**	**mommy**	**child's name**	**child's name**	cracker	**child's name**	need	**mommy**	**bye**	outing
6	grandpa	**no**	bottle	**hi**	play	**hi**	can	water	bath	**child's name**
7	woof woof	**yum yum**	**peekaboo**	**grandma**	**peekaboo**	**bye**	weee	**no**	**hi**	baby food
8	cat	**hi**	**no**	water	**grandma**	pattycake	thump	ouch	**no**	up
9	vroom	**bye**	**bye**	**no**	with	**yum yum**	hide and seek	ball	lamp	clap
10	**yum yum**	ouch	bath	woof woof	food	good night	meow	tickle	look	ball

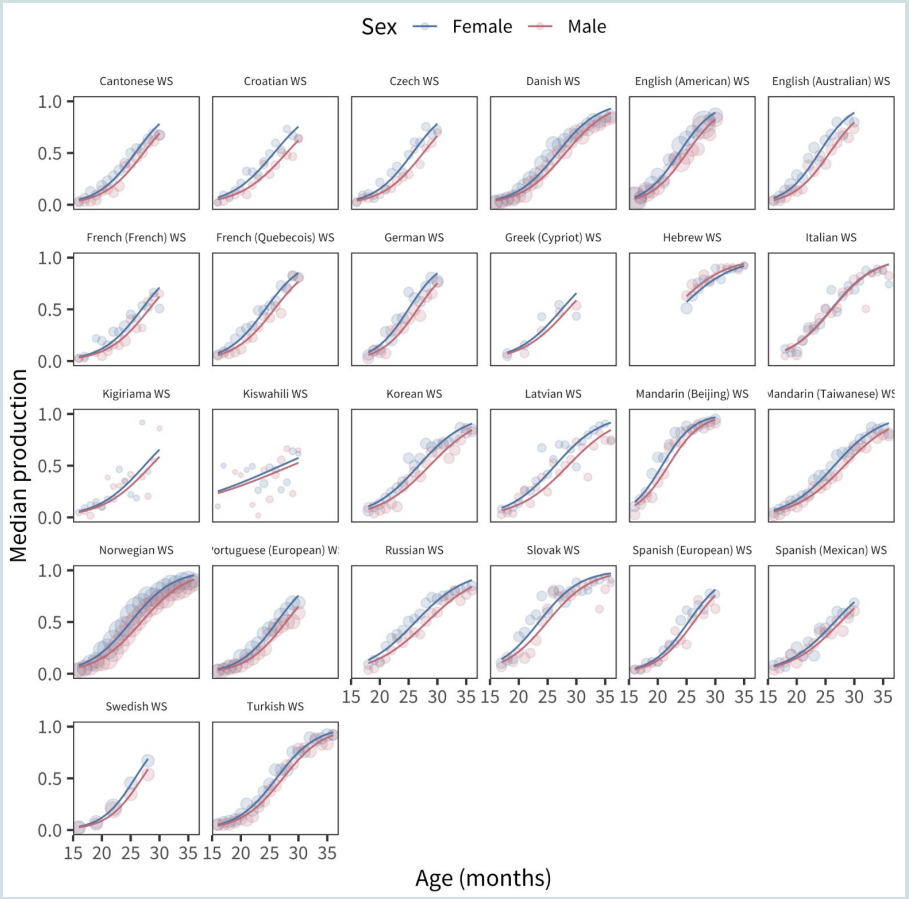
Consistency in early vocabulary across languages

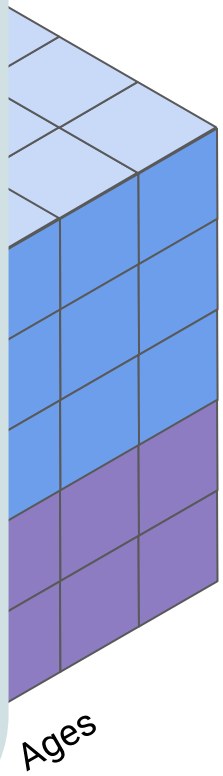
Frank et al. (2021)

Preference for acquiring words with more potential links to their lexicons



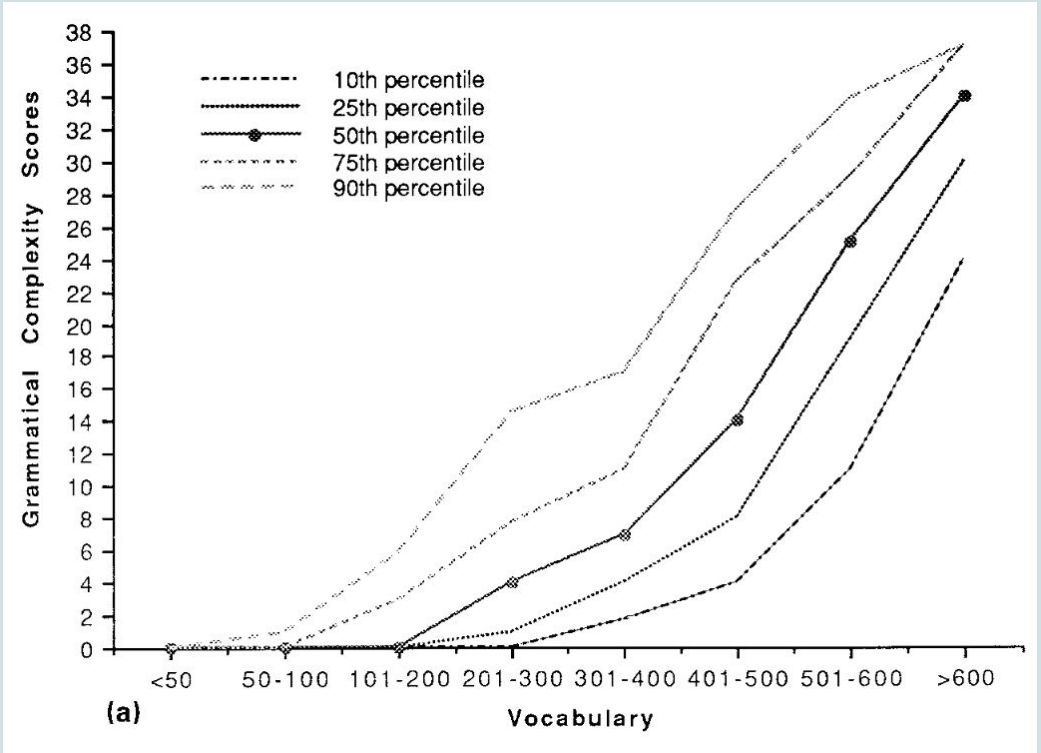
Girls have larger vocabularies than boys

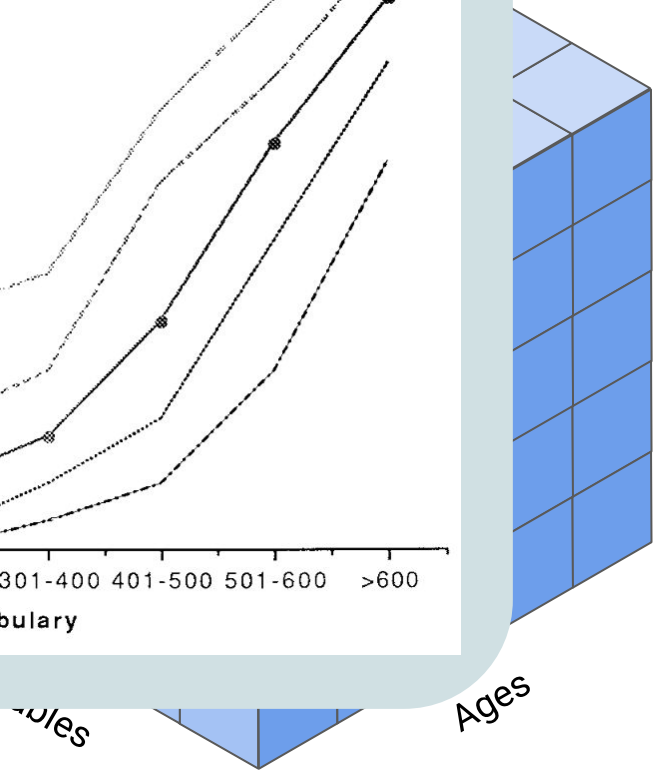




Frank et al. (2021)

Grammar and lexicon co-develop



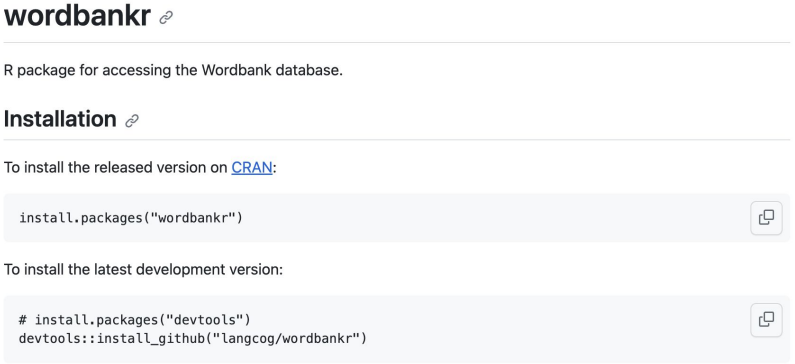


[Bates & Goodman (1997)]

Accessing Wordbank data

- Downloading…

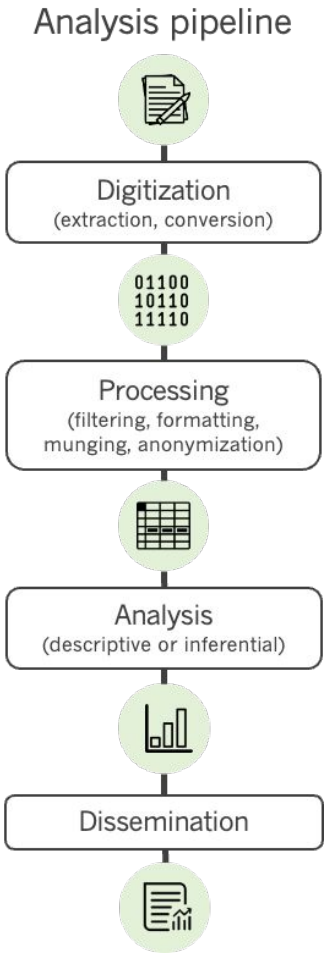
- …or using `wordbankr`
  - Reproducible workflow
  - Version control
  - Flexible updating



Braginsky et al. (2023)

Reproducibility: What

		Data	
		Same	Different
Code	Same	**Reproducible**	Replicable
	Different	Robust	Generalisable



Reproducibility: Why

		Data	
		Same	Different
Code	Same	**Reproducible**	Replicable
	Different	Robust	Generalisable

Provenance, transparency, & verification

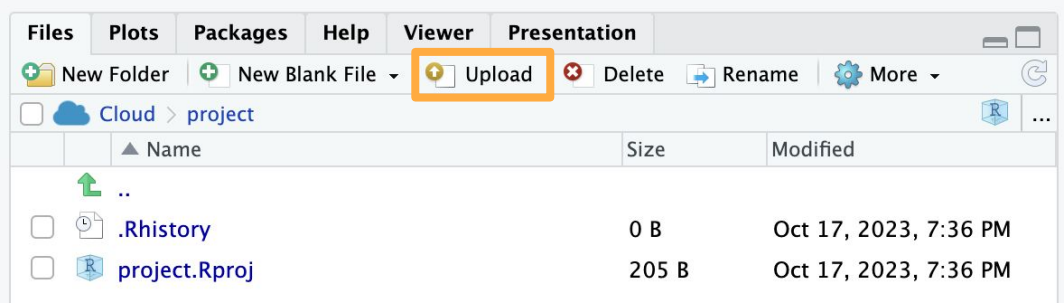
Error detection & mitigation

Efficiency

Reproducibility: How

Reproducible reports: R Markdown (.Rmd) / Quarto (.qmd) documents

- Embed all your processing and analysis code in your report
- Avoid updating / copy-paste errors
- Save pain when changing formats
- Interpretable and transparent
- Quick and ✨pretty✨

Practical session

- Go to https://posit.cloud and register (if you haven't)
- Create a "New RStudio project"
- Download the Rmd file for the practical at https://github.com/alvinwmtan/2023-mpal-wordbank and upload it into your project (the "Upload" button is in the bottom right panel)