Consequences of lexical variability in toddlers' vocabularies

Caitlin Richter
ricca@seas.upenn.edu

(Very) Large-scale word acquisition studies

MacArthur-Bates Communicative Development Inventory (**CDI**, Fenson et al. 1994, Frank et al. 2017) and cross-linguistic adaptations list children's vocabulary contents

- Comparable to other methods of early vocabulary estimation (Bleses et al. 2008, Naigles & Hoff-Ginsberg 1998)

CDI variation studies focus on:

- Age vs. vocabulary **size** (Bleses et al. 2008; Fenson et al. 1994, Trudeau & Sutton 2011)
- Proportional composition of lexicon in **high-level categories** – parts of speech, semantic classes (Braginsky et al. 2015, D'Odorico et al. 2001, Wehberg et al. 2008)

➢ Notable exception: Mayor & Plunkett (2014) estimated lexical variability

- Substantial **production and perception variability** for all CDI age ranges

This study **directly compares sets of words children know**

- 12 language varieties: Danish, English (Australia), English (USA), Korean, Mandarin (Beijing), Mandarin (Taiwanese), Norwegian, Portuguese (EU), Russian, Slovak, Spanish (Mexican), Turkish
- 1000+ children for each language, most 16-36 months
- 546–771 words per language, average 667

Jaccard similarity

Overlap in two children's word inventories:
Ratio of items both children produce to total number of items at least one of them produces

$$J(A,B) = \frac{A \cap B}{A \cup B}$$

- Accurate in smaller CDI vocabularies
- Overestimated in larger vocabularies, where CDI itself is less accurate

Variability for a group is represented by **average pairwise Jaccard similarity** for pairs in that group (Kim et al. 2018)

Example calculation:

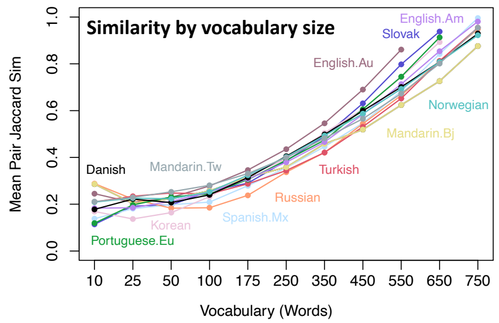
Child A	Child B
ball	
big	
book	book
	cow
	fish
house	
mommy	mommy
	no
	phone
pink	
sheep	sheep
star	
3 shared words 12 total words	
Jaccard similarity: 3/12 = 0.25	

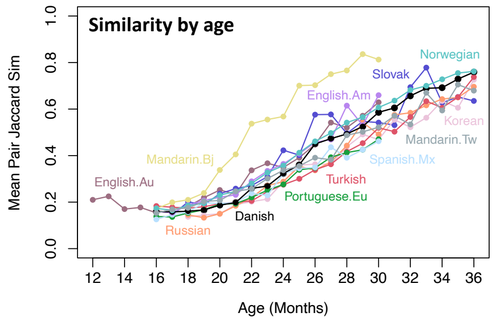
References: Bates & Goodman 1997 On the inseparability of grammar and the lexicon, Bleses et al 2008 Early vocabulary development in Danish and other languages, Braginsky et al 2015 Developmental Changes in the Relationship Between Grammar and the Lexicon, Brown 1973 A first language: The early stages, Cazden 1968 The acquisition of noun and verb inflections, D'Odorico et al 2001 Vocabulary development in Italian children: A longitudinal evaluation of quantitative and qualitative aspects, Fenson et al 1994 Variability in early communicative development, Ferguson & Farwell 1975 Words and sounds in early language acquisition, Fisher Gleitman & Gleitman 1991 On the semantic content of subcategorization frames, Fourtassi et al 2020 The growth of children's semantic and phonological networks, Frank et al 2017 Wordbank: An open repository for developmental vocabulary data, Gleitman et al 2005 Hard words, Kidd et al 2018 Individual differences in language acquisition and processing, Kim et al 2018 Differences of Early Semantic Relatedness between Late Talkers and Typically Developing Children, Naigles & Hoff-Ginsberg 1998 Input to verb learning: Evidence for the plausibility of syntactic bootstrapping, Schneider et al 2015 Large-scale investigations of variability in children's first words, Stoel-Gammon 2011 Relationships between lexical and phonological development in young children, Tardif et al 2008 Baby's first 10 words, Trudeau & Sutton 2011 Expressive vocabulary and early grammar of 16-to 30-month-old children acquiring Quebec French, Vihman 2019 Phonological templates in development, Wehberg 2007 Danish children's first words, Yang 2016 The price of linguistic productivity.

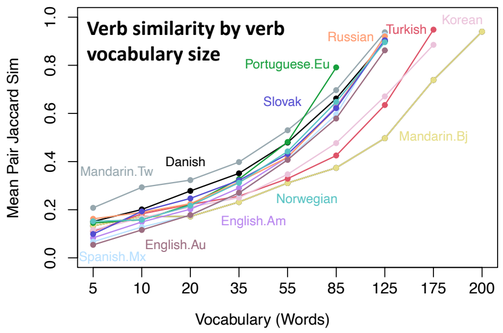
There is no such thing as the order, or age, of word acquisition for a language

When children know the **same number** of words, to what extent do they know the **same words**?

What about children at the **same age**?







➢ **Jaccard similarity is only 0.1 – 0.5 during the first 350 words, or first 28 months**

Could variability come mainly from different **learning styles** like noun-dominant or routine-dominant? (D'Odorico et al. 2001, Trudeau & Sutton 2011)

➢ **No:** Order of elements within classes is also highly variable

➢ **Each child learns words in their own order and timing, with substantive differences**

A closer look at first words

US English:

Vocabulary	# Children	Total words	Once	Most common
2	30	28	17	–
5	63	68	32	**mommy (48)**, daddy (44)
10	84	121	49	daddy (72), **mommy (71)**, **uh oh (48)**
15	76	166	66	**mommy (73)**, daddy (69), **uh oh (51)**, **bye (45)**, ball (40)
20	82	221	91	**mommy (76)**, daddy (74), ball (64), **uh oh (55)**, **grrr (52)**, dog (49), **bye (44)**, **no (42)**

Danish:

Vocabulary	# Children	Total words	Once	Most common
2	17	15	6	–
5	39	40	17	mm mm (lækkert) (23)
10	56	85	35	mm mm (lækkert) (39), hej (37), **vov (36)**, av (34), **far (32)**, tak (28)
15	62	131	59	**vov (49)**, **nej (48)**, av (46), hej (46), **far (44)**, mm mm (lækkert) (43), tak (42), **mor (40)**, **hej hej (farvel) (35)**
20	72	169	60	mm mm (lækkert) (66), **vov (61)**, hej (61), **far (61)**, av (60), **nej (55)**, tak (54), ja (48), mad (42), årnnn (bil- lyd) (42), **hej hej (farvel) (41)**, grrr (37), **åh åhh (ups) (36)**

Table columns: (1) Child vocabulary size ±1 word, (2) Number of children in sample with this vocabulary size, (3) Total number of word types for all children, (4) Number of words present in only one child's vocabulary, (5) words present in half or more of children's vocabularies.

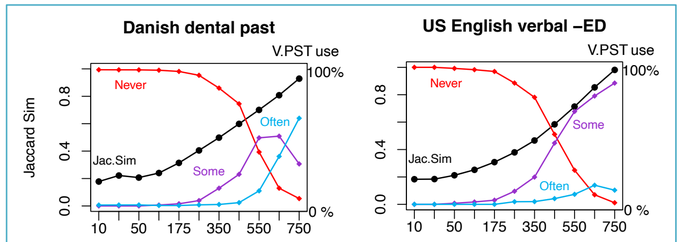
First 1-10 words are described as strikingly **similar across languages** (Schneider et al. 2015, Tardif et al. 2008, Wehberg 2007)

What about different children learning each language?

➢ Top words do recur across languages, but **children's first words are mostly not these** (Ferguson & Farwell 1975, Stoel-Gammon 2011, Tardif et al. 2008)

Comparison: learning morphology

Morphemes acquisition has a **predictable sequence** tied to **vocabulary size** (Bates & Goodman 1997, Brown 1973, Cazden 1968, Trudeau & Sutton 2011)



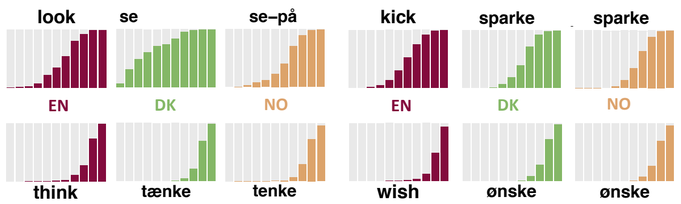
Proportion of children using Germanic dental past suffix Never, Sometimes, or Often; Jaccard similarity reproduced for reference.

➢ **Morphology acquisition course is similar regardless of the identity of the lexical items it is acquired from** (Braginsky et al. 2015, Kidd et al. 2018, Yang 2016)

...but that doesn't mean just anything goes

Syntactic bootstrapping constrains variability (Fisher et al. 1991, Gleitman et al. 2005)

- Some semantic properties are hard to observe but systematically encoded in syntax: mental content, perspectives on events
- Learning is delayed: only after learning syntax↔semantics relations



Mental verbs think, wish in English, Danish and Norwegian. Concrete look, kick, match English frequencies; translations are nearest equivalents on CDI. Bars show the same 11 vocabulary stages used on other graphs.

➢ Important to **check that real individuals follow predictions** (Fourtassi et al. 2020, Naigles & Hoff-Ginsberg 1998)

- Idealisations like Average Age of Acquisition could hide relevant variation

❖**Early word learning is characterised by individual variability**
❖**Specific predictions stand out against this and illuminate word acquisition mechanisms**
❖**Children acquire grammar similarly regardless of lexicon differences**

I thank Charles Yang for collaborative work on this project, as well as the members of the autumn 2020 Distributional Learning seminar it emerged from.

Scripts to reproduce the data are available at https://github.com/catiR/cdi-lexsim