University at Buffalo
The State University of New York

Comparing Semantic Networks of Early Vocabulary across Languages

Mengyang Qiu
Nichol Castro, Ph.D.

Department of Communicative Disorders and Sciences

Background and Motivation

- Network-based approaches have provided important insights into the organization and acquisition of early vocabulary in children (e.g., Hills et al., 2009; Beckage et al., 2011)
  - Word association norms (e.g., USF and SWOW) and the CHILDES database are commonly used to infer the (static) semantic structure of early vocabulary
  - Norms from the MacArthur-Bates Communicative Development Inventory (MB-CDI) are used to estimate age of acquisition and compare different lexical growth models: preferential attachment, preferential acquisition, and lure of the associates

Background and Motivation

- Network-based approaches have provided important insights into the organization and acquisition of early vocabulary in children (e.g., Hills et al., 2009; Beckage et al., 2011)
  - Word association norms (e.g., USF and SWOW) and the CHILDES database are commonly used to infer the (static) semantic structure of early vocabulary
  - Norms from the MacArthur-Bates Communicative Development Inventory (MB-CDI) are used to estimate age of acquisition and compare different lexical growth models: preferential attachment, preferential acquisition, and lure of the associates
- Can we generalize the findings to other languages?
- Are there any structural differences across languages?

Background and Motivation

- The growth of semantic networks of early vocabulary from 10 languages (Fourtassi et al., 2020)
  - Used the English free association norms as referent
  - Substantial semantic similarities are shared across languages (Youn et al., 2016): celestial phenomena (SUN, MOON), landscape features (SKY, SEA), and natural substances (STONE, WATER)

Background and Motivation

- The growth of semantic networks of early vocabulary from 10 languages (Fourtassi et al., 2020)
  - Used the English free association norms as referent
  - Substantial semantic similarities are shared across languages (Youn et al., 2016): celestial phenomena (SUN, MOON), landscape features (SKY, SEA), and natural substances (STONE, WATER)
- Large free association norms or corpora of child-directed speech are often not available for many languages
  - In CHILDES, $\sim$1.5M utterances from parents in English (up to 30 months) vs. 82,450 in Chinese (Mandarin + Cantonese)
  - SWOW: 17 languages

Background and Motivation

- The growth of semantic networks of early vocabulary from 10 languages (Fourtassi et al., 2020)
  - Used the English free association norms as referent
  - Substantial semantic similarities are shared across languages (Youn et al., 2016): celestial phenomena (SUN, MOON), landscape features (SKY, SEA), and natural substances (STONE, WATER)
- Large free association norms or corpora of child-directed speech are often not available for many languages
  - In CHILDES, $\sim$1.5M utterances from parents in English (up to 30 months) vs. 82,450 in Chinese (Mandarin + Cantonese)
  - SWOW: 17 languages
- The structure of free association networks changes across the life span (Dubossarsky et al., 2017)

Network Estimation Directly from MB-CDI

- A valid and efficient tool to assess early comprehensive/expressive language
- MB-CDI has been adapted to many other languages
- Data available on Wordbank (Frank et al., 2017)

Network Estimation Directly from MB-CDI

Full Child-by-Word Data

Language

English (American)

Form

Words & Sentences

Get Data

Caution: can take a while

Download Data

Show 25 entries

Search:

data_id	age	sex	mom_ed	value	item_id	type	category	definition
All			All		All		All	All
129242	27	Female	Some Secondary	produces	item_1	word	sounds	baa baa
129242	27	Female	Some Secondary		item_2	word	sounds	choo choo
129242	27	Female	Some Secondary		item_3	word	sounds	cockadoodledoo
129242	27	Female	Some Secondary	produces	item_4	word	sounds	grrr
129242	27	Female	Some Secondary		item_5	word	sounds	meow
129242	27	Female	Some Secondary		item_6	word	sounds	moo
129242	27	Female	Some Secondary	produces	item_7	word	sounds	ouch

Network Estimation Directly from MB-CDI

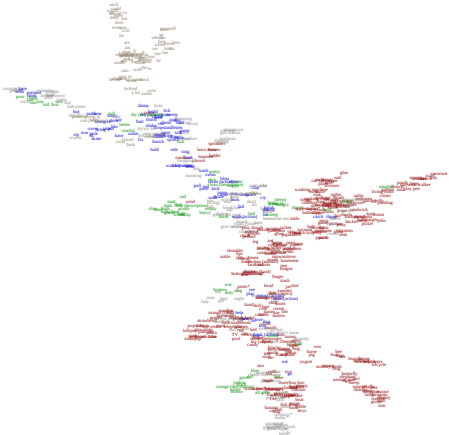
data_id	age	item_100	item_101	item_102	item_103	item_104	item_105	item_106	item_107	item_108	item_109	item_110	item_111	item_112
129242	27	1	1	0	1	1	1	1	1	1	1	0	1	1
129243	21	1	0	1	0	1	0	1	0	1	1	0	1	1
129244	26	0	0	0	0	1	0	1	0	1	0	1	1	1
129245	27	0	0	0	0	0	0	0	0	0	1	0	0	1
129246	19	0	0	1	0	1	0	0	0	1	0	0	0	0
129247	30	1	1	0	1	1	0	1	1	1	1	1	1	1
129248	25	0	0	1	1	1	0	1	1	0	1	1	0	1
129249	24	1	1	1	1	1	1	1	1	1	1	1	1	1
129250	28	1	1	0	1	1	1	1	1	1	1	1	1	1
129251	18	0	0	0	0	0	0	0	0	0	0	0	0	0
129252	29	1	1	0	1	1	1	1	1	1	1	1	1	1
129253	26	0	0	0	0	1	1	1	1	1	1	0	1	1
129254	18	0	0	0	0	0	0	0	0	0	0	0	0	0
129255	24	0	0	0	0	1	0	0	0	1	1	0	1	0
129256	28	1	0	0	1	1	0	1	1	1	1	1	0	1
129257	24	1	1	1	1	1	1	1	1	1	1	1	0	1
129258	26	1	0	1	1	1	0	0	1	0	0	1	0	1
129259	16	0	0	0	0	0	0	0	0	0	0	0	0	0
129260	17	0	0	0	0	1	0	1	0	1	0	1	1	1
129261	30	1	1	1	1	1	1	1	1	1	1	1	1	1
129262	19	1	0	0	0	0	0	0	0	0	0	1	1	1
129263	28	1	1	1	1	1	1	1	1	1	1	1	1	1
129264	26	1	0	0	0	1	1	1	0	1	1	1	1	1
129265	29	0	1	0	0	1	0	0	1	1	1	0	1	1
129266	22	0	0	0	0	1	0	1	0	1	0	0	1	1
129267	23	1	0	0	0	1	0	0	1	0	0	0	0	1
129268	29	1	1	1	1	1	1	1	1	1	1	1	1	1
129269	24	1	1	1	1	1	1	1	1	1	1	1	1	1
129270	25	1	1	1	1	1	1	1	1	1	1	1	1	1
129271	17	1	1	0	0	1	0	0	0	0	0	0	0	0
129272	24	0	1	1	0	1	0	1	1	1	1	0	1	1
129273	25	1	0	0	1	1	1	1	0	1	1	1	0	1
129274	24	1	0	0	1	1	1	1	1	1	1	1	1	1
129275	25	0	0	0	0	1	0	0	0	1	0	0	0	0
129276	16	0	0	0	0	1	0	0	0	0	0	1	0	0
129277	19	0	0	0	0	0	0	0	0	0	0	0	0	0

Network Estimation Directly from MB-CDI

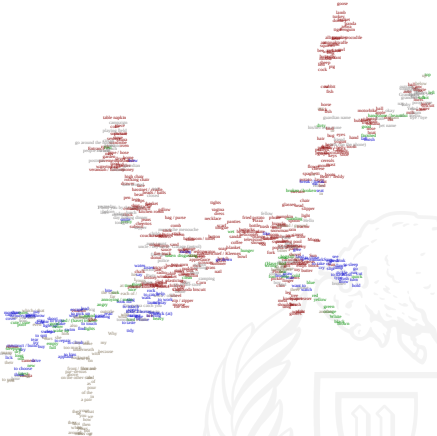
- MB-CDI response matrix: each row represents a child (16 to 30 months), and each cell denotes whether the child produce the word ("1") or not ("0")
- MB-CDI network (at 30 months): nodes represent words in MB-CDI (sentences and the sounds category excluded), and edges represent cosine similarities between word profiles, similar to the correlation based network construction method for verbal fluency (e.g., Borodkin et al., 2016)
- We used the triangulated maximally filtered graph method (Massara et al., 2017) to capture the most important information
- 4 MB-CDI networks: English (American), French (Quebec), Mandarin (Beijing), and Cantonese

Results

English (American)

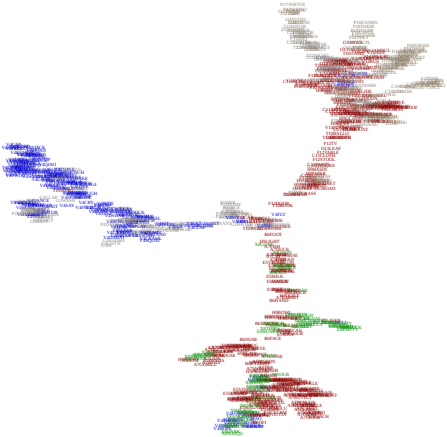


French (Quebec)



Results

Mandarin (Beijing)



Cantonese

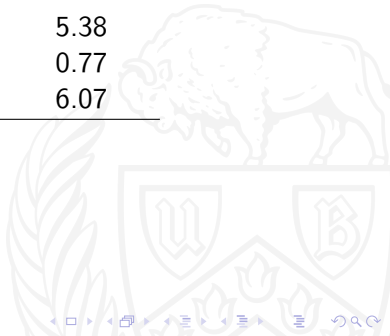
Results

Top 10 highest degree nodes

English	French	Mandarin	Cantonese
leg (58)	fork (89)	**fridge (55)**	**bathroom (69)**
hand (57)	bedroom (58)	**hair (42)**	big (55)
table (57)	**table (44)**	sweater (40)	wash_hands (54)
find (42)	mouth (36)	**sing (36)**	home (46)
hair (41)	**sing (36)**	park (35)	**fridge (44)**
sleep (37)	**find (35)**	window (32)	sit (43)
with (37)	chair (32)	seed (31)	bed (43)
bathroom (36)	baby (30)	stuff/fill (30)	clothes (38)
run (36)	angry (30)	glasses (30)	**hand (37)**
fall (34)	**bathroom (29)**	play (27)	stair (37)

Results

	English	French	Mandarin	Cantonese
Nodes	668	652	787	791
Edges	1998	1950	2355	2367
CC	0.26	0.25	0.31	0.26
ASPL	5.32	5.15	8.53	5.38
Modularity	0.76	0.74	0.78	0.77
Smallworldness	5.16	4.90	6.73	6.07

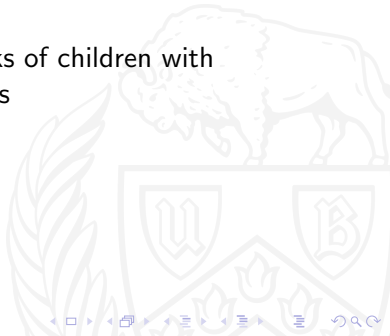


Results

	Animals				Verbs			
	English	French	Mandarin	Cantonese	English	French	Mandarin	Cantonese
Nodes	43	41	49	44	103	109	194	172
Edges	86	96	92	74	246	217	525	301
CC	0.45	0.44	0.53	0.49	0.33	0.38	0.31	0.36
ASPL	2.45	2.50	1.97	2.21	2.98	3.13	3.55	3.22
Modularity	0.49	0.44	0.59	0.46	0.57	0.69	0.67	0.76
Smallworldness	1.00	0.98	1.61	1.50	0.26	1.73	1.18	1.93

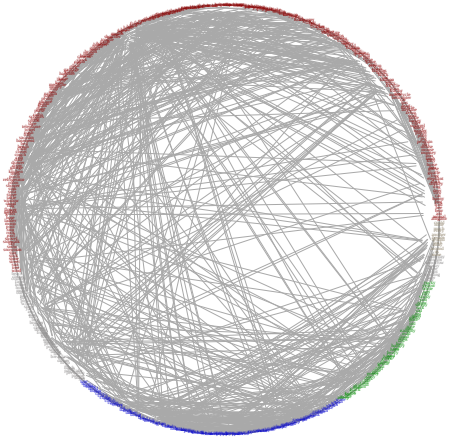
Future Work

- To compare the English MB-CDI network with the free association network and CHILDES network
- To include more languages (uni-lemma)
- To model lexical growth using MB-CDI networks?
- To examine structural/growth differences in MB-CDI networks of children with developmental language disorder / autism / cochlear implants

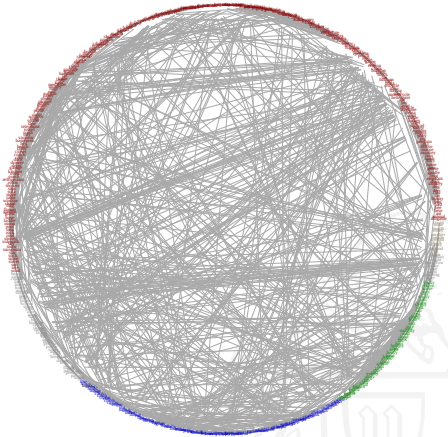


Future Work

English (MB-CDI)



English (CHILDES)



References

Beckage, N., Smith, L., & Hills, T. (2011). Small worlds and semantic network growth in typical and late talkers. *PloS one*, *6*(5), e19348.

Borodkin, K., Kenett, Y. N., Faust, M., & Mashal, N. (2016). When pumpkin is closer to onion than to squash: The structure of the second language lexicon. *Cognition*, *156*, 60–70.

Dubossarsky, H., De Deyne, S., & Hills, T. T. (2017). Quantifying the structure of free association networks across the life span. *Developmental Psychology*, *53*(8), 1560–1570.

Fourtassi, A., Bian, Y., & Frank, M. C. (2020). The growth of children's semantic and phonological networks: Insight from 10 languages. *Cognitive Science*, *44*(7), e12847.

Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (2017). Wordbank: An open repository for developmental vocabulary data. *Journal of Child Language*, *44*(3), 677–694.

Hills, T. T., Maouene, M., Maouene, J., Sheya, A., & Smith, L. (2009). Longitudinal analysis of early semantic networks: Preferential attachment or preferential acquisition? *Psychological Science*, *20*(6), 729–739.

Massara, G. P., Di Matteo, T., & Aste, T. (2017). Network filtering for big data: Triangulated maximally filtered graph. *Journal of Complex Networks*, *5*(2), 161–178.

Youn, H., Sutton, L., Smith, E., Moore, C., Wilkins, J. F., Maddieson, I., ... Bhattacharya, T. (2016). On the universal structure of human lexical semantics. *Proceedings of the National Academy of Sciences*, *113*(7), 1766–1771.