Quantifying the bilingual (dis)advantage in vocabulary acquisition

Yining Hua, Harvard College, Harvard Medical School, yining_hua@hms.harvard.edu

Joshua Hartshorne, Boston College, joshua.hartshorne@bc.edu
Authors

Ning Hua
Affiliated Graduate Student
Harvard Medical School
Brigham and Women’s Hospital

Joshua Hartshorne
Department of Psychology and Neuroscience, Boston College
Hypothesis

>> Monolingual children: 2*time hearing the same language

Twice better in language efficiency tests?
Hypothesis

But...

(Bialystok et al., 2010)

>> Most words appear way more frequently than others?

2 year-olds only know high-frequency words.

>> The extra words monolinguals hear aren’t necessary?

Distribution of PPVT standard scores in English for monolingual (n = 772) and bilingual (n = 966) children.
Analysis 1

>> Simulate the learning process
   1,000 monolingual and 1,000 bilingual children

>> The big picture
1. Generate a daily conversation corpus for each child
2. Find a way to measure the learning outcome -> 680 English CDI
3. Simulate learning process on the selected words

CDI: The MacArthur-Bates Communicative Development Inventories (CDIs) are parent-report instruments for data-gathering about early language acquisition.
Pipeline

(1) Build the Phrase Corpus

TalkBankDB → Query → Qualified CHAT files → Phrase corpus → Simulated daily conversations for 1000 kids (age=30)

(2) Simulate daily conversations

Word frequency of the 680 CDI words → Find learned vocabulary distribution of CDI words → By-child by-word learning rate at 30
The corpus

- Selecting transcripts (CHAT files)

- Build the corpus from the transcripts based on the CHAT manuscript:
  E.g. [/] denotes a repetition of the last word
Word frequencies

We chose 30 month as the limitation because Wordbank only has info up to 30 month

- 30-month-old monolinguals ➔ ~18.75 millions of input words ~\( N \) sentences (varies with language)
- Randomly sample with replacement for \( N \) (\( N/2 \) for bilinguals) sentences
- Calculate word frequencies of the CDI words in each of the daily conversations

#input_words = 0.625*month (million), Mid-low SES

Number of words addressed to children in the three SES groups by age
Source: Emma Kelty-Stephen, adapted from Hart & Risley, 1995
Pipeline

Test 1: Estimate how many times a kid must hear a word in order to learn it?

(1) Build the Phrase Corpus
- TalkBankDB
  Query → Qualified CHAT files → Phrase corpus

(2) Simulate daily conversations
- Word frequency of the 680 CDI words
- Simulated daily conversations for 1000 kids (age=30)

(3) Calculate learning rates of CDI words
- By-child by-word learning rate at 30
- By-month acquisition rate of 680 CDI words
- English Monolingual CDI-WS data → By-month acquisition rate of 680 CDI words
Learning rates

1 - \text{pbinom}(0, \text{num_occurence}, \text{learning_rate}) = \text{acquisition_rate}
Simulation alg.

For each of the children:
  For each of the 680 CDI words:
    Repeat word_frequency times:
      r <- random float <- (0, 1);
      If r in range(0, learning rate):
        The child learned the word;
      The child didn’t learn the word in the learning process;
Results - English

Child Vocabulary Distribution

Ratio: 296/415 = 0.71325301204
Results - Norwegian

Ratio: $\frac{186}{256} = 0.7265625$
Results - Mandarin

Child Vocabulary Distribution

Vocabulary sizes

Number of kids

Ratio: 465/588 = 0.79081632653
Goal

Distribution of PPVT standard scores in English for monolingual \(n = 772\) and bilingual \(n = 966\) children.
What about the adults?

Bilinguals and monolinguals **adults** have similar language efficiency.
Possible reasons

>> We assumed that monolinguals and bilinguals have the same acquisition rate at month 30.
What if the bilinguals learned way faster and they had higher acquisition rate?

>> Real data?
Bilingual CDI data

472 children

age = 12 - 48 with missing months

1. English-Hebrew: 40 children, age = 24, 26, 28, 29, 30, 31, 32, 34, 35, 36, 40, 41, 42, 43, 45, 47
2. English-Spanish (dataset 1): 147 children, age = 22, 25, 30, 36, 42, 48
3. English-Spanish (dataset 2): 161 children, age = 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36
4. Maltese-English: 9 children, age = 12, 16, 20, 24, 25, 26, 30
5. Irish-English: 48 children, age = 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36
6. French-English: 68 children, age = 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 28, 29, 30, 31, 32
Analysis 2 & 3

2. Vocabulary acquisition ~ age*bilingualism + (1 | childID) + e

3. Vocabulary acquisition ~ #input_words*bilingualism + (1 | childID) + e

#input_words: English_exposure * estimated input words (Hart & Risley, 1995)

1. Proportion of time exposed to English
2. Language households

Related data: e.g. information of the main caregiver was used by taking the average of the values reported.

* For those data with no english exposure information at all, we excluded them in this analysis.
Results - age

Few data points at the end -> noise

Monolinguals

Bilinguals
Results - #input words
By word type Analysis

- CDI words are classified into types of words (e.g. animals, action words, etc.)

Vocabulary acquisition ~ \#input_words*bilingualism + 
\textbf{wordType}*bilingualism + 
(1 | childID) + e
As compared to action words
1. **Action words** are the hardest to learn in accordance with Gleitman, L. (1990). The structural sources of verb meanings. Language acquisition, 1(1), 3-55.

2. Bilingual children have most advantage in learning **connecting words**, followed by **question words**, and **connection verbs**.

3. Bilingual children have the least advantage (but still advantage) in learning **household** words and descriptive words, followed by names of **body parts** and **food drinks**.
Conclusion

1. The difference between monolingual and bilingual performance in early-age language efficiency tests did not (only) come from the less heard words.

2. Monolingual and Bilingual performance regarding language efficiency are similar in adults.

3. The number of words learned are significantly associated with age, the number of input words, bilingualism.
   a. Bilinguals learn slower when we look at the age (in early ages)
   b. Bilinguals learn way faster when we look at the number of input words (in early ages)
      i. Bilingual children have most advantage in learning connecting words, followed by question words, and connection verbs.
      ii. Bilingual children have the least advantage (but still advantage) in learning household words and descriptive words, followed by names of body parts and food drinks.
Thanks to our contributors

Dr. Odelya Ohana & Dr. Armon-Lotem Sharon
Bar Ilan University

Dr. Daniela R. Gatt
L-Università ta’ Malta

Dr. Erika Hoff
Florida Atlantic University

Dr. Jacqueline Legacy
Concordia University, Texas

Dr. Ciara O’Toole
Concordia University, Texas

Dr. Diane Poulin-Dubois
Concordia University
Make a contribution

We'll help you get the data into Wordbank's format!

Wordbank
An open database of children's vocabulary development

Wordbank contains data from 75,144 children and 82,983 CDI administrations, across 29 languages and 56 instruments:

- 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 is an open database of children's vocabulary growth, featuring data from contributors around the world. Wordbank archives data from the MacArthur-Bates Communicative Development Inventory (MB-CDI), a family of parent-report questionnaires and enables researchers to browse these data in interactive analyses and access them via the wordbankr R package.
Bilingual CDI data:


Others: