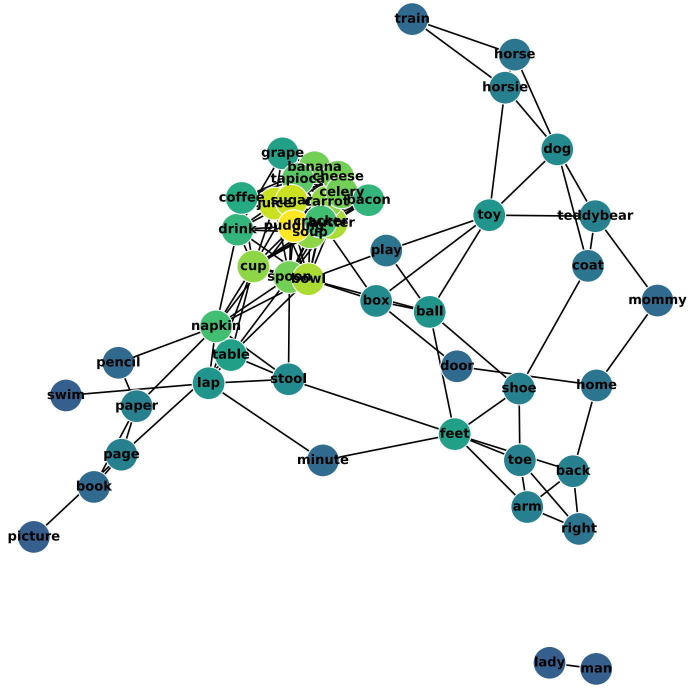
Semantic network growth of children from different socio-economic backgrounds

Man Ho Wong

April 19, 2022



Background

- Vocabulary development of a child has been linked to mother's educational background and socio-economic status
- Children from families of higher socio-economic status have been shown to have larger vocabulary size in early ages.
- The goal of this project is to study the relationship between vocabulary development and child-directed speech (CDS) among native American English speakers
- Instead of vocab size, I am particularly interested in the growth of semantic network

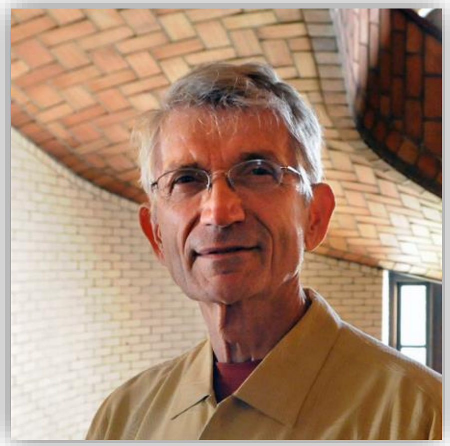
1 Data curation

Goal: To search for and download corpora relevant to the project

GitHub: Child-Vocab-Development/code/data_curation.ipynb

Data source: CHILDES of TalkBank

- **TalkBank** – a multilingual speech corpus directed by our neighbor, Brian MacWhinney at Carnegie Mellon University
- **CHILDES** (**Chi**ld **L**anguage **D**ata **E**xchange **S**ystem) /tʃaɪldɪs/?
  - child language component of TalkBank
- Organized by different languages and clinical conditions:
  - English (NA, AAE, UK), Korean, etc.
  - Language/developmental disorders
  - Bilingual children
  - and more
- Creative Commons License (CC BY-NC-SA 3.0)
- More rules and guidelines for using data on their website
- You can also contribute data from your research participants or (future) children



Brian MacWhinney
Professor of Psychology and Modern Languages at Carnegie Mellon University
Picture: https://www.cmu.edu/dietrich/modlang/about-us/filter/affiliated/brian-macwhinney.html

Thank you, Brian!

Example: Brown Corpus

CHILDES English Brown Corpus

Roger Brown (1925-1997)
Psychology and Social Relations
Harvard University
website

Participants:	3
Type of Study:	naturalistic
Location:	USA
Media type:	no longer available
DOI:	doi:10.21415/T5HK5G

Not all corpora provide audio/video files

Browsable transcripts

Download transcripts

Transcript in .CHAT format

```
*CHI: don't    CLITIC    dog       .
%mor: mod|do   neg|not   v|dog     .
%gra: 1|3|AUX  2|1|NEG   3|0|ROOT 4|3|PUNCT
%com: Adam repeated these utterances several times
```

Citation information

Brown, R. (1973). *A first language: The*
Harvard University Press.

In accordance with TalkBank rules, any use o
accompanied by at least one of the above references.

Project Description

This subdirectory contains the complete transcripts from the three participants Adam, Eve, and Sarah who were studied by Roger Brown and his students between 1962 and 1966. Adam was studied from 2;3 to 4;10, Eve from 1;6 to 2;3, and Sarah from 2;3 to 5;1. Brown (1973) summarized this research and provided detailed documentation regarding data collection, transcription, and analysis.

116-page long manual for transcription format

Tools for Analyzing Talk

Part 1: The CHAT Transcription Format

Brian MacWhinney
Carnegie Mellon University

January 24, 2022
https://doi.org/10.21415/3mhn-0z89

When citing the use of TalkBank and CHILDES facilities, please use this reference to the last printed version of the CHILDES manual:

MacWhinney, B. (2000). The CHILDES Project: Tools for Analyzing Talk. 3rd Edition.
ssociates.

he programs and data systematically through

Rules for transcribing recording

In addition to these cliticizations, other common assimilations include forms listed in this table.

Assimilations

Assimilation	Standard	Assimilation	Standard
dunno	don't know	kinda	kind of
dyou	do you	sorta	sort of
gimme	give me	whyntcha	why didn't you
lemme	let me	wassup	what's up
lotsa	lots of	whaddya	what did you

Unlike the mod:aux group, further types of assimilations are nearly limitless. Some of the most common assimilations are listed in the v-clit.cut file in MOR. However, it is not possible to list all possible assimilations or to assign them to particular parts of speech. Moreover, these other assimilations need to be treated as two or more morphemes. To do this, you should use the replacement notation, as in

```
*CHI: lemme [: let me]
```

If you do this, MOR and the other programs will work on the material in the square brackets, rather than the *lemme* form. An even simpler way of representing some of these forms is by noting omitted letters with parentheses as in: “gi(ve) me” for “gimme,” “le(t) me” for “lemme,” or “d(o) you” for “dyou.”

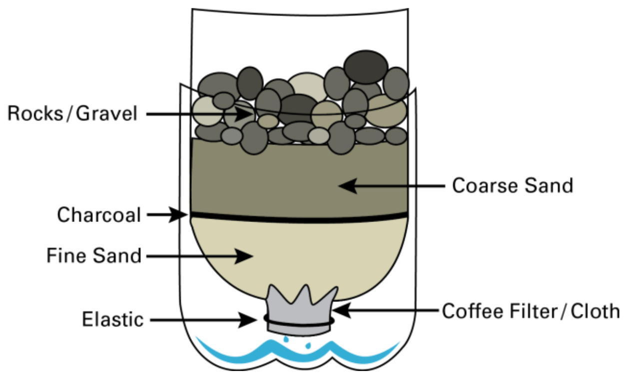
Which corpora to use?

- There are **47** NA English corpora in CHILDES.
  - e.g. Brown, MacWhinney, NewmanRatner etc.
- Impractical to evaluate each file in each corpus…
- A more efficient strategy:

  Search for relevant data in **three phases**, where each phase narrows down the scope of search, and each phase uses more specific search criteria

  - 1. **Identify** relevant corpora fitting a set of basic criteria
  - 2. **Screen** for CHAT files containing the information we need
  - 3. **Refine** the dataset by filtering CHAT files with more specific criteria
    - done on the fly during data analysis

Filtering data like water…



Picture: https://prepperpete.files.wordpress.com/2014/04/diagram.png

Search criteria

Phase 1:

- Check if **ANY CHAT file** in the corpus match the following criteria (not inspected every file!)
  - **Participants:** data should include child ('CHI') or mother ('MOT')
  - **Child info:** data should contain child age, sex and socioeconomic status (SES) info (if not included in mother's info)
  - **Mother info:** data should contain SES or education info
- **Result: 13 corpora** matching the criteria

This is a bit more manageable than 47 corpora…

Phase 2:

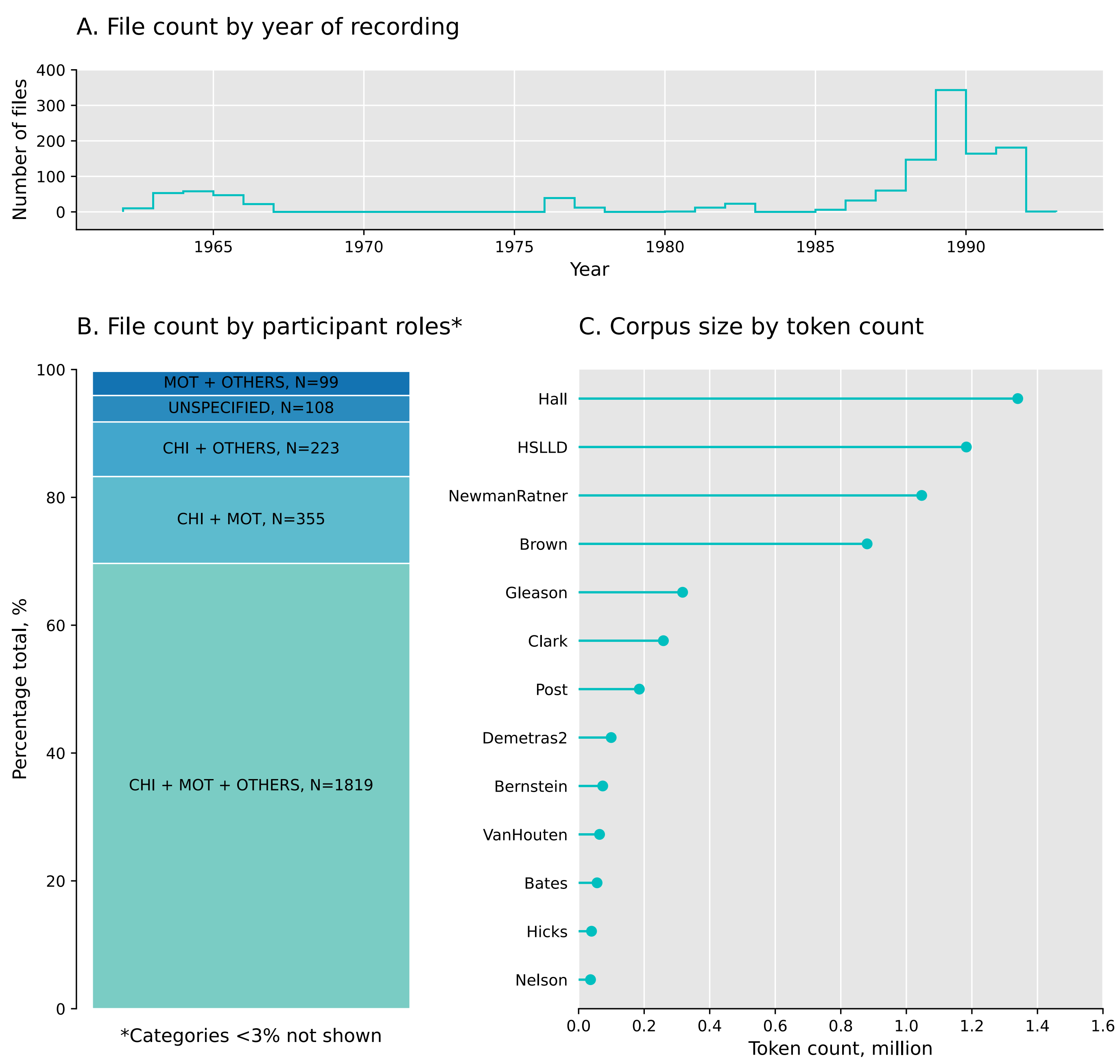
- Not all CHAT files were inspected in Phase 1 - not sure whether all files in each of the 13 corpora match the search criteria
- Phase 2: **Screen all CHAT files** with the same criteria as in Phase 1
- Additional criterion: children **younger than 6 years old**

A Pandas DataFrame, `data_idx`, was also created to store file info. It serves as an index to the files in all the corpora.


```
{'UTF8': '',
  'PID': '11312/c-00015633-1',
  'Languages': ['eng'],
  'Participants': {'CHI': {'name': 'Adam',
    'language': 'eng',
    'corpus': 'Brown',
    'age': '2;03.18',
    'sex': 'male',
    'group': 'TD',
    'ses': 'MC',
    'role': 'Target_Child',
    'education': '',
    'custom': ''},
   'MOT': {'name': 'Mother',
    'language': 'eng',
    'corpus': 'Brown',
    'age': '',
    'sex': 'female',
    'group': '',
    'ses': '',
    'role': 'Mother',
    'education': '',
    'custom': ‘’},
...
  'Date': {datetime.date(1962, 10, 22),
datetime.date(1962, 10, 23)},
  'Time Duration': '15:00-16:00',
  'Types': 'long, toyplay, TD',
  'Tape Location': '646'}
```


Header of each CHAT file contains basic info about the recording, e.g. languages, participants, etc.
(Can be accessed with PyLangAcq Python package)

Composition of curated data



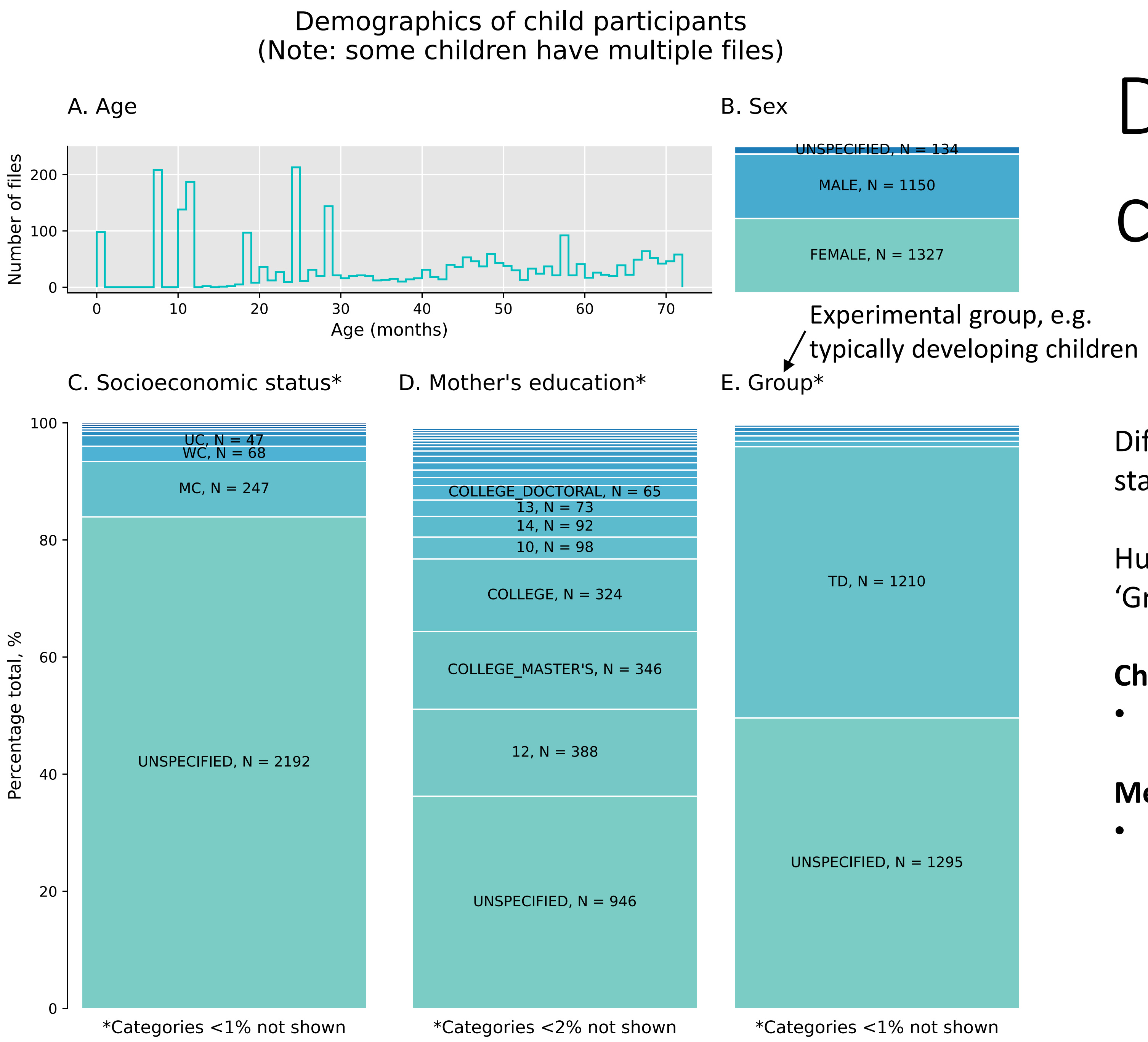
Oldest kids are > 60 years old now

Fun fact: a lot of recordings were made in Greater Boston area/ New England

2 Data preprocessing

Goal: To prepare clean and well-integrated data for analysis

GitHub: Child-Vocab-Development/code/data_preprocessing.ipynb



Data needs some cleaning…

Different corpora use different annotations/ standards

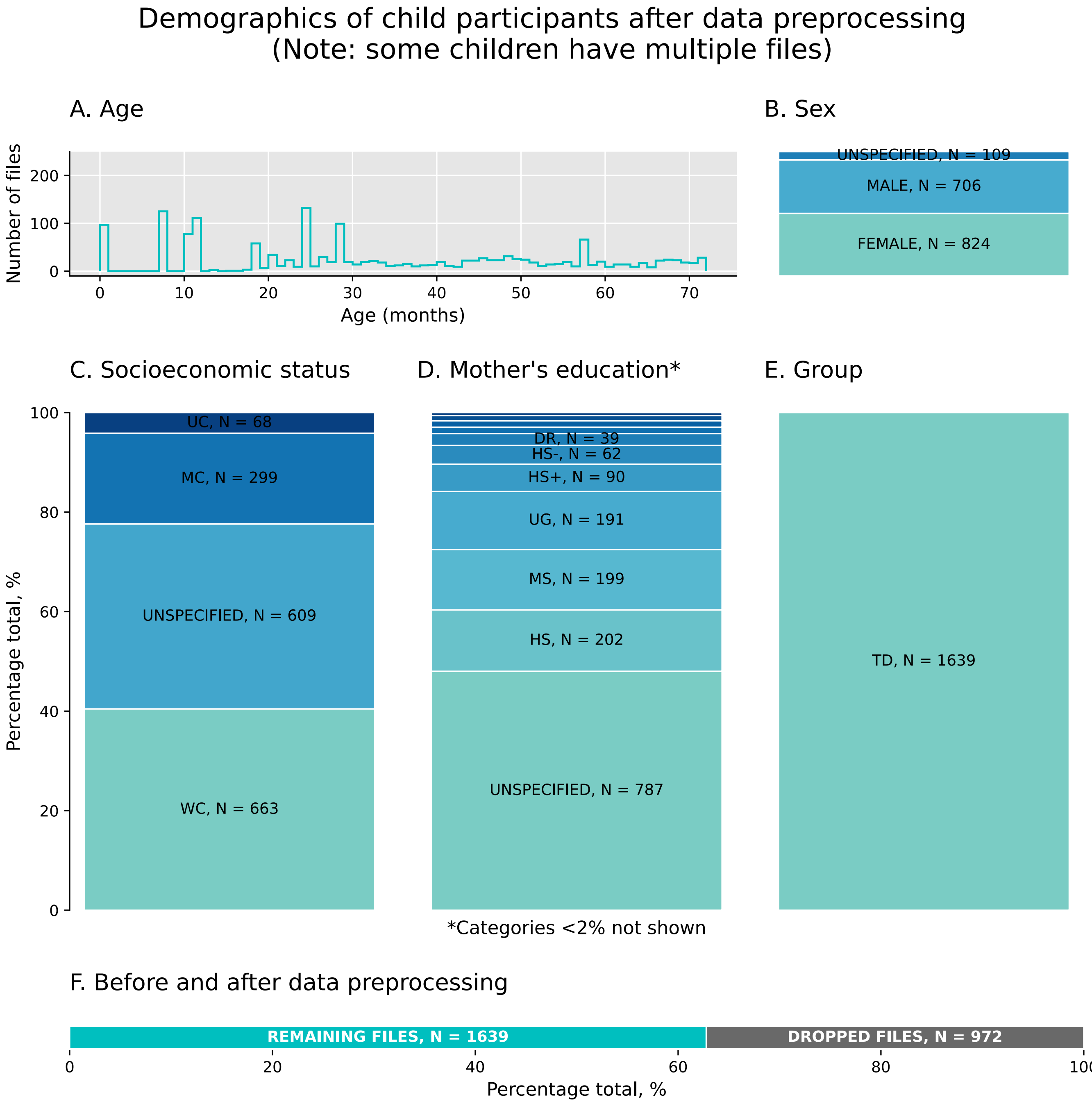
Humans make mistakes (e.g. SES info entered in the ‘Group’ field)

Check for missing info

- e.g. ‘`unspecified`’ labels?

Merge labels

- e.g. mother’s education level: ‘`12`’ should be merged with ‘`high_school_diploma`’ and ‘`GED`’ into one category – ‘`HS`’ (high school level)



After data preprocessing…

Missing info were filled with info available on corpus's homepage; unfilled entries were dropped

Labels with same definition were **merged**

Entries with both SES and mother's education info missing were **dropped**

Entries with these keywords in the '`type`', '`situation`' or '`file_path`' field were **dropped**:

- Read, book, story, elicit, explanatory, magnet, interview (non-CDS)
  (i.e. files recorded in less naturalistic situation, e.g. book reading and elicited data where the discourse was more or less 'planned')

3 Exploratory analysis

Goal: To evaluate the quality and limitations of the processed data

GitHub: Child-Vocab-Development/code/exploratory_analysis.ipynb

Evaluating the processed data with MLU

- **MLU** – **M**edium **L**ength of **U**tterance
  - A common metric to assess child's linguistic productivity during development

$$MLU = \frac{Number\ of\ morphemes\ (or\ words)}{Number\ of\ utterances}$$

- In general, accuracy depends on:
  - Correctly parsed speech
  - Sample size (total number of utterances)

Age range (in years)	MLUw	MLUm
2; 6–2; 11	2.91	3.23
3-3; 11	3.57	3.95
4-4; 11	4.19	4.66
5-5; 11	4.42	4.92
6-6; 11	4.63	5.14
7-7; 11	4.82	5.33
8-8; 11	5.03	5.59

Rosselli M, Ardila A, Matute E, Vélez-Uribe I. Language Development across the Life Span: A Neuropsychological/Neuroimaging Perspective. Neurosci J. 2014;2014:585237.

Other useful metrics:
- Type-to-token ratio (TTR)
- Noun-to-verb ratio (NTVR)
- Variation set

MLU was chosen for this project for its simplicity and versatility (can be used to assess both children and caregivers)

Evaluating the processed data with MLU

Overall data quality

- Are the numbers close to those in published papers?
- If not:
  - Speech **correctly parsed**?
  - **Sample size** sufficient?
  - **Sample distribution** balanced?
  - Can we really **combine** different corpora?

Limitations (more practical considerations…)

- Which factors (e.g. SES, mother's education) show an effect on MLU?
- Age range of data?
- Is PyLangAcq package sufficient to extract data from CHAT files?

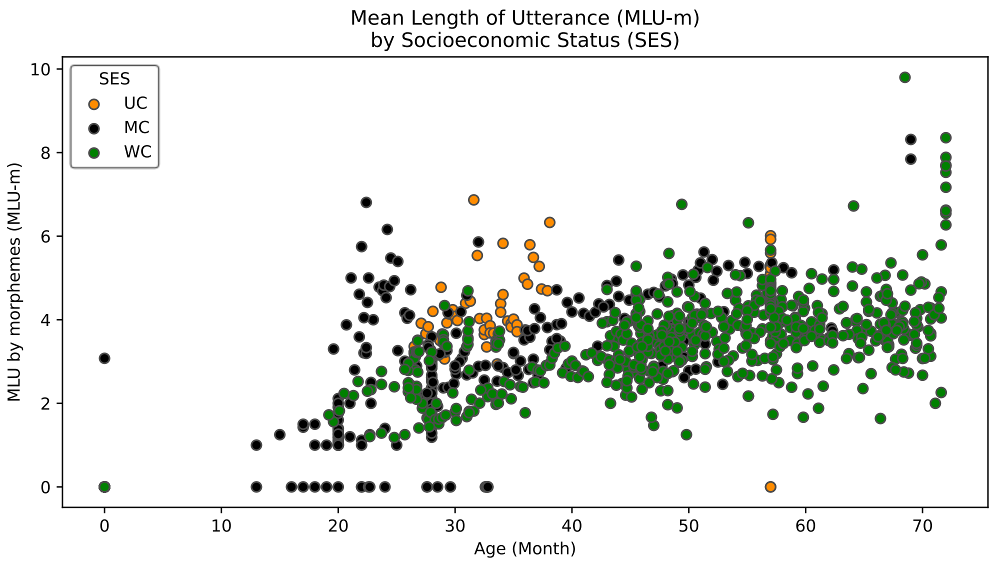
PyLangAcq's MLU function counts period (.) and empty string as a word or morpheme as long as they are annotated!

→ Will need to develop custom function for this project

Example:

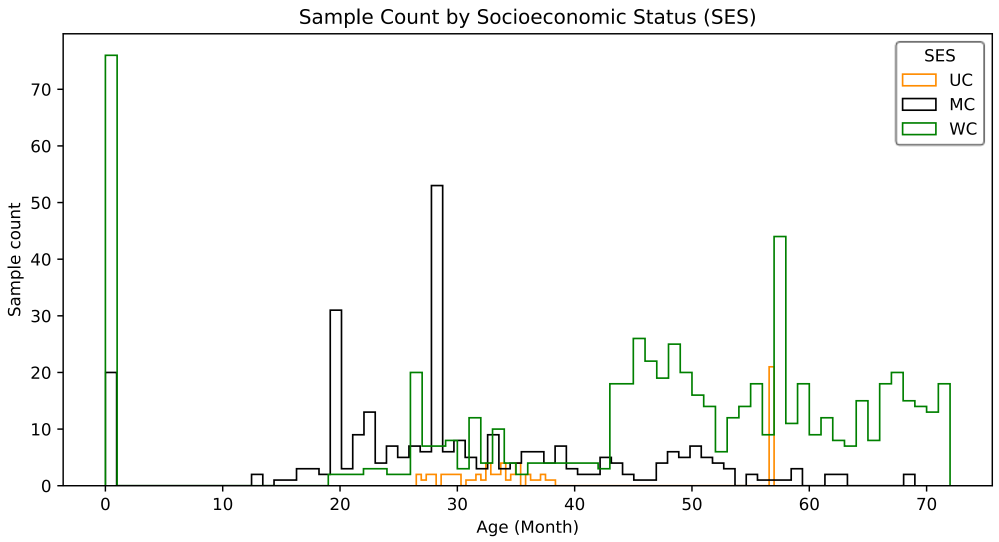
```
Mother: do you want some?  # 4 morphemes (correct)
Child: .                   # Period counted as a morpheme
Mother: yes?
Child: (nodding)           # Gesture annotated as 'none'
                           #   and counted as a morpheme

# Both child's utterances have 1 "morpheme"
```



MLU from processed data look reasonable

- MLU matching the published data: about 2 to 4.5 morphemes/utterance from age 2 to 6
- As expected, children from higher SES background showed higher MLU
- (Statistical tests not done yet, but the trend is clear)



However…

- **Sample distribution** is not balanced across ages in each SES
- UC has small **sample size**
- Some age groups have an unusually large sample size – likely come from corpora where participant age was well-matched, and from repeated recordings within a month.
- Only **20 to 42 months** have adequate sample size across SES
    - Will look at this age range (also matching most studies)

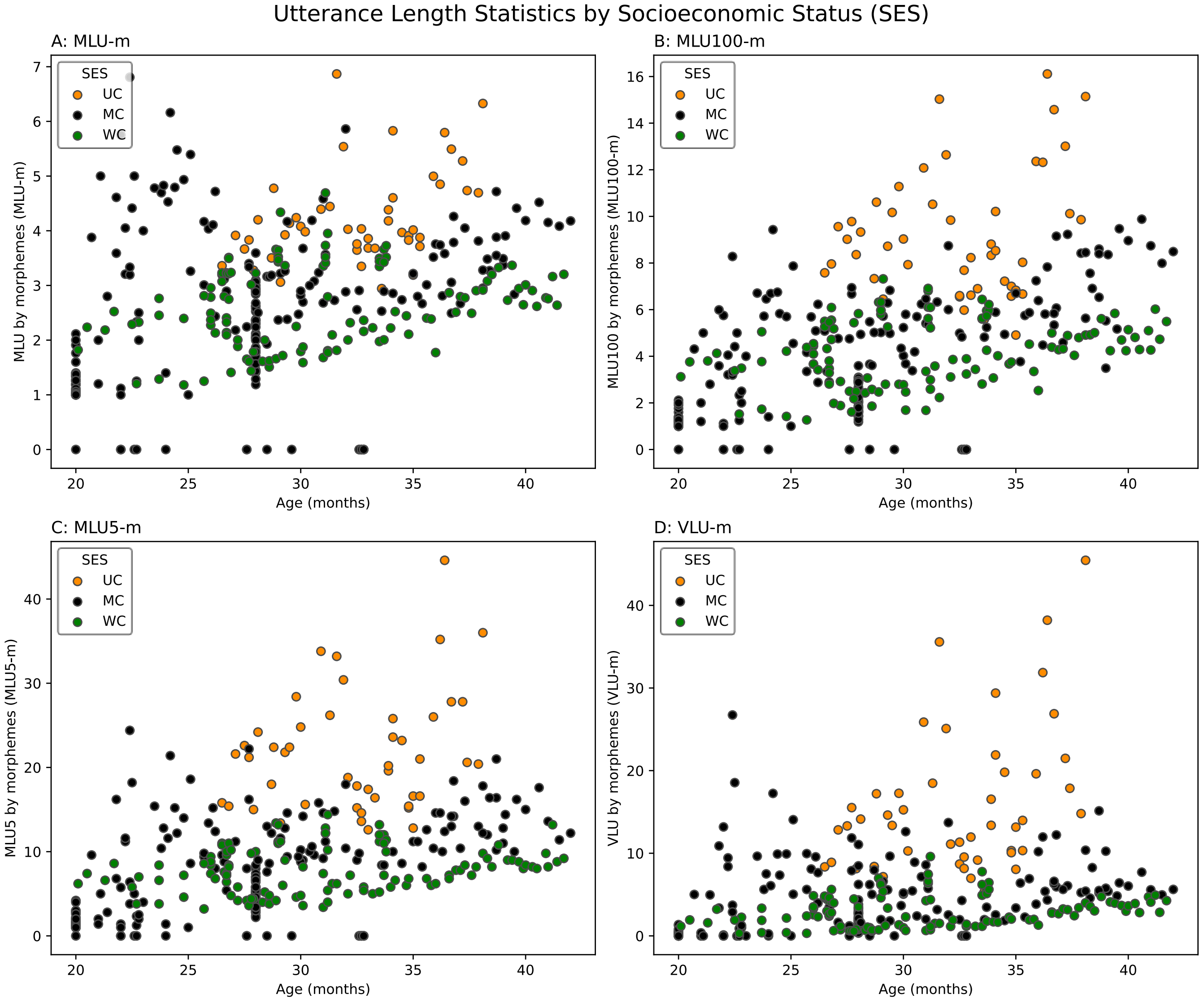
Confirming effects of SES with other measurements of utterance length

Samples have different lengths of recording
Need to confirm the robustness of MLU

Evaluate data with different lengths of measurement
MLU-m: Mean Length of Utterance by morpheme
MLU100-m: MLU-m of first 100 utterances
MLU5-m: MLU-m of first 5 utterances
VLU-m: Variance of utterance length by morpheme

Same observation in all four measurements:

- Effects of SES are detected regardless of lengths of measurement
- Data is ready for further analysis



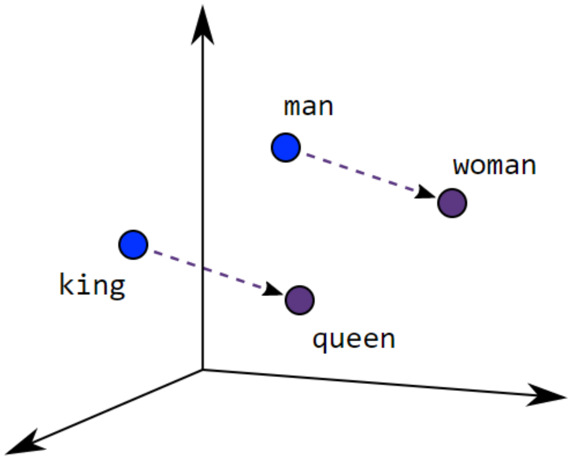
4 Vocabulary analysis

Goal: To characterize semantic network properties across SES

GitHub: Child-Vocab-Development/code/vocabulary_analysis.ipynb

Semantic network: Relationship between words

- Relationship between words in a lexicon/vocabulary can be represented by how similar they are in their semantic meanings
- Many metrics to measure **word-to-word similarity**
- metrics can be derived from two different sources:
  - human-annotated datasets (e.g. WordNet)
  - word associations learned by machine algorithms (e.g. word2vec)
- This project will use machine-generated word associations because this is…
  - more flexible than human-annotated datasets
    e.g. getting different word associations by using different training data probably less prone to human biases (depending on training data)
  - Besides, human-annotated datasets are based on formal taxonomies of words (such knowledge is unlikely to be present in a young child's world)



Source: [Embeddings: Translating to a Lower-Dimensional Space | Machine Learning Crash Course | Google Developers]

ConceptNet

Two main types of machine-learning models to generate word associations:

- count-based and prediction-based models

This project will use word embeddings based on a semantic network called ConceptNet

- a network built by both count-based and prediction-based models
- The most unique feature:
  unlike other semantic networks, it is concept-based*
  rather than word-based

  *Concepts based on an open commonsense database, Open Mind Common Sense (OMCS)

Catherine Havasi, one of the developers of ConceptNet, grew up in Pittsburgh
Source: Wikipedia

ConceptNet Numberbatch

GitHub: ConceptNet5

ConceptNet

For mapping word relations in a young child's lexicon, concept-based models are probably more suitable than other word embedding models:

- word meanings are closely related to the concepts that the child is acquiring at the same time

Word/concept similarities derived from ConceptNet:

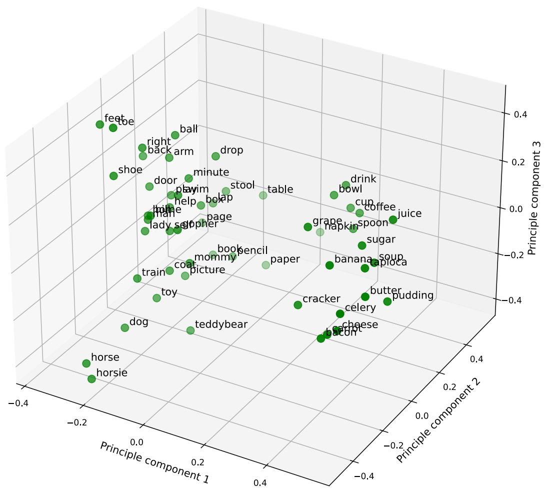
```
>>> wordvec.most_similar('pittsburgh’)
[('university_of_pittsburgh', 0.9756660461425781),
 ('pittsburgher', 0.9552338123321533),
 ('carnegie_mellon_university', 0.9533942937850952),
 ('yinzer', 0.9217094779014587),
 ('pittsburghese', 0.8789857029914856),
 ('benjamin_franklin_bridge', 0.8256341218948364),
 ('philadelphia_county', 0.8219272494316101),
 ('independence_hall', 0.8207418918609619),
 ('philadelphia', 0.8077220320701599),
 ('walt_whitman_bridge', 0.7923399806022644)]
```

Words/phrases similar to ‘pittsburgh’

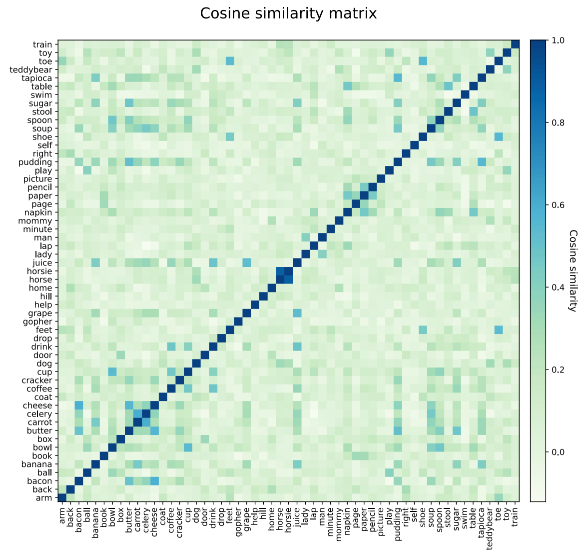
Cosine similarity between two words

Steps to generate a semantic network

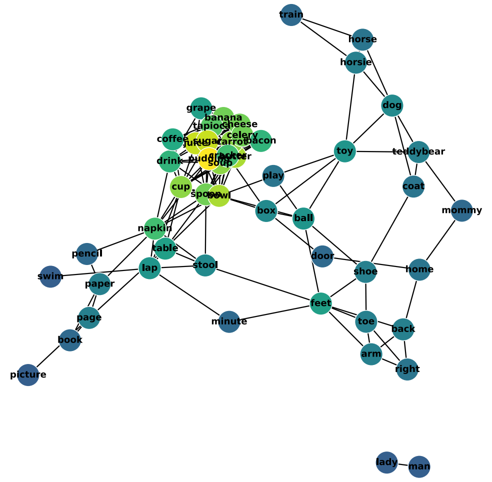
1. Get the word vectors



2. Calculate pairwise similarities between word vectors



3. Generate edges between nodes using similarity as weight



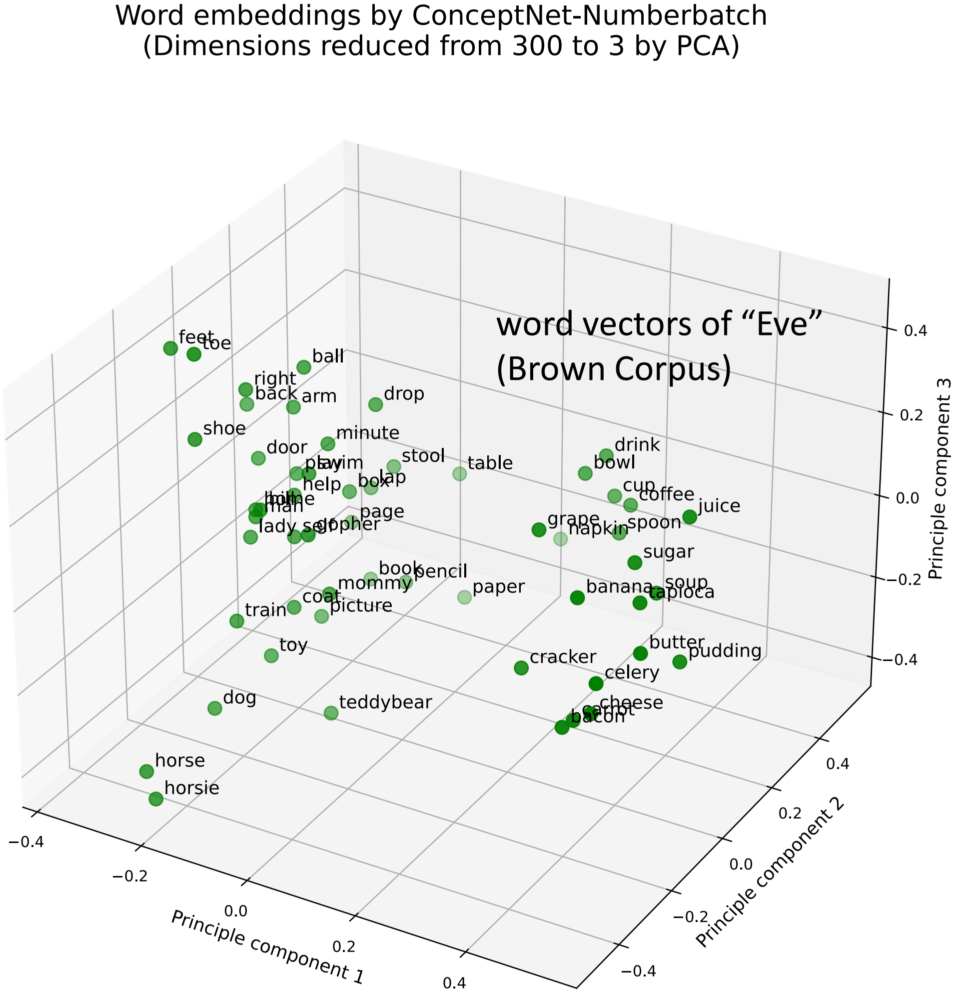
1. Get the word vectors

I. Extract tokens from a CHAT file
only interested in nouns this time

II. Import ConceptNet-Numberbatch
a set of word vectors from a pre-trained model based on ConceptNet

III. Map each word in the word list to its word vector in ConceptNet-Numberbatch
Out-of-vocabulary words are excluded

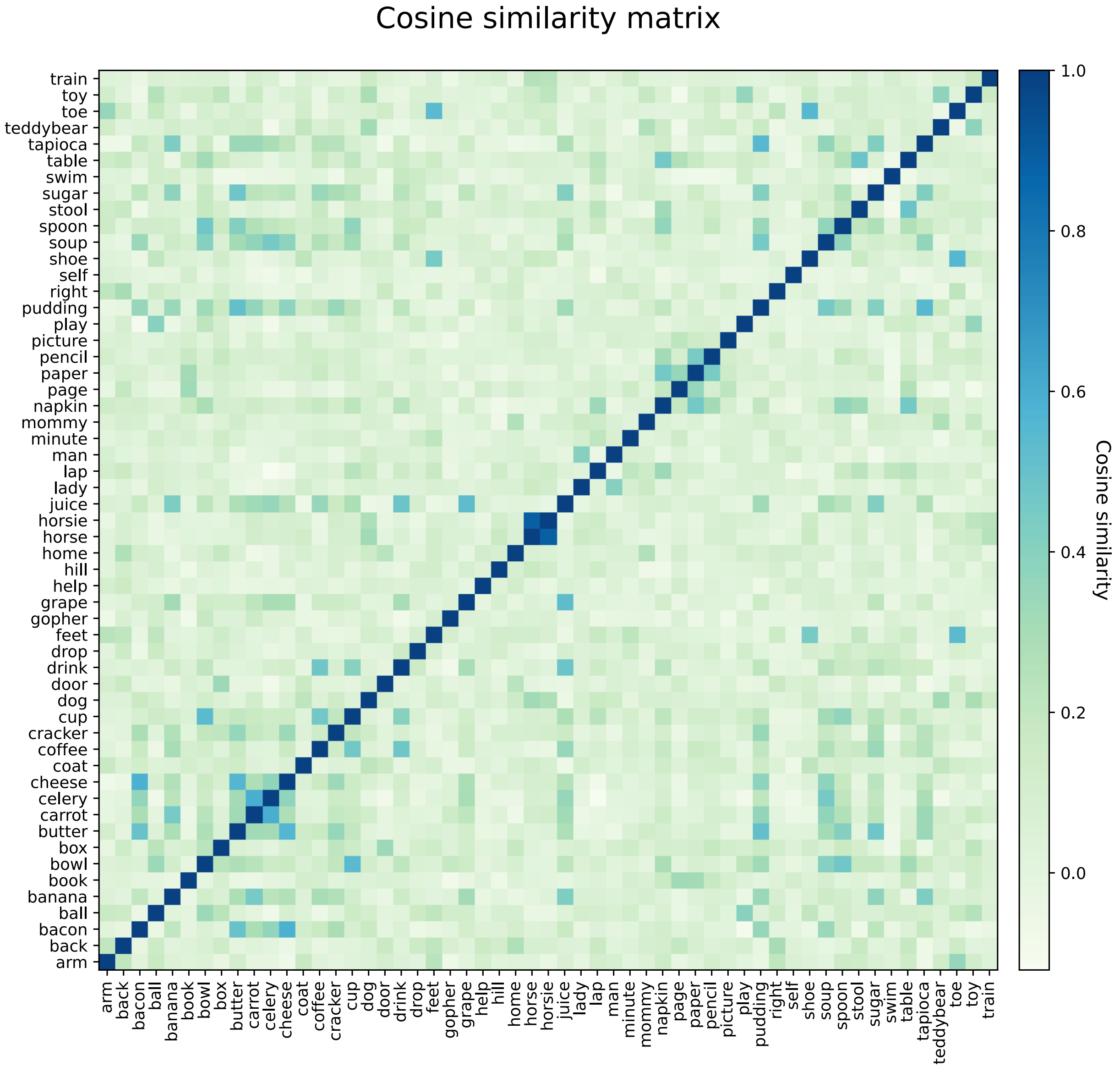
Can be done with **Gensim**
(a Python library for training of vector embeddings)



2. Calculate pairwise similarities between word vectors

This project uses **cosine similarity**

Can be done with Scikit-Learn's `pairwise.cosine_similarity` function



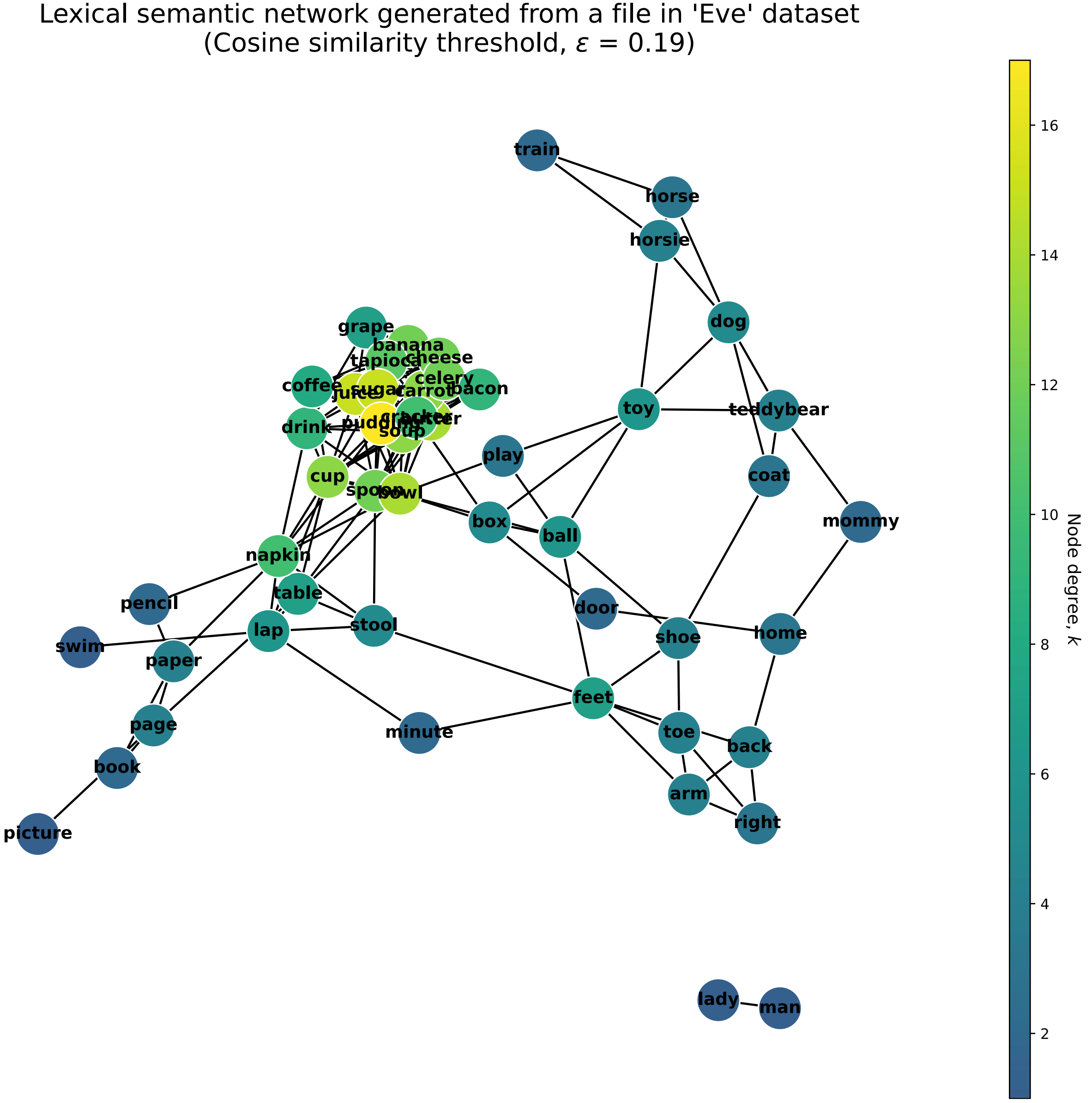
3. Generate edges between nodes using similarity as weight

Need a threshold for cosine similarity, otherwise, all nodes will just be connected

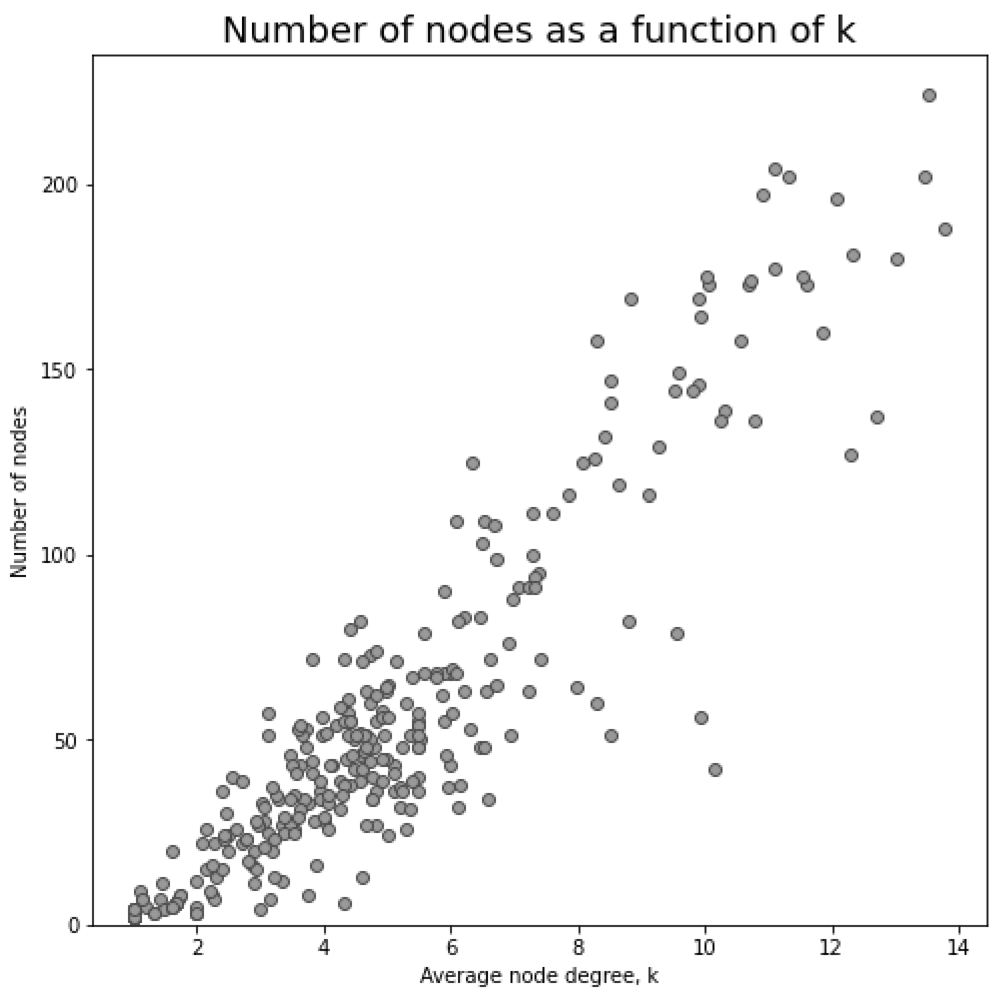
Threshold set at 0.19 according to the literature

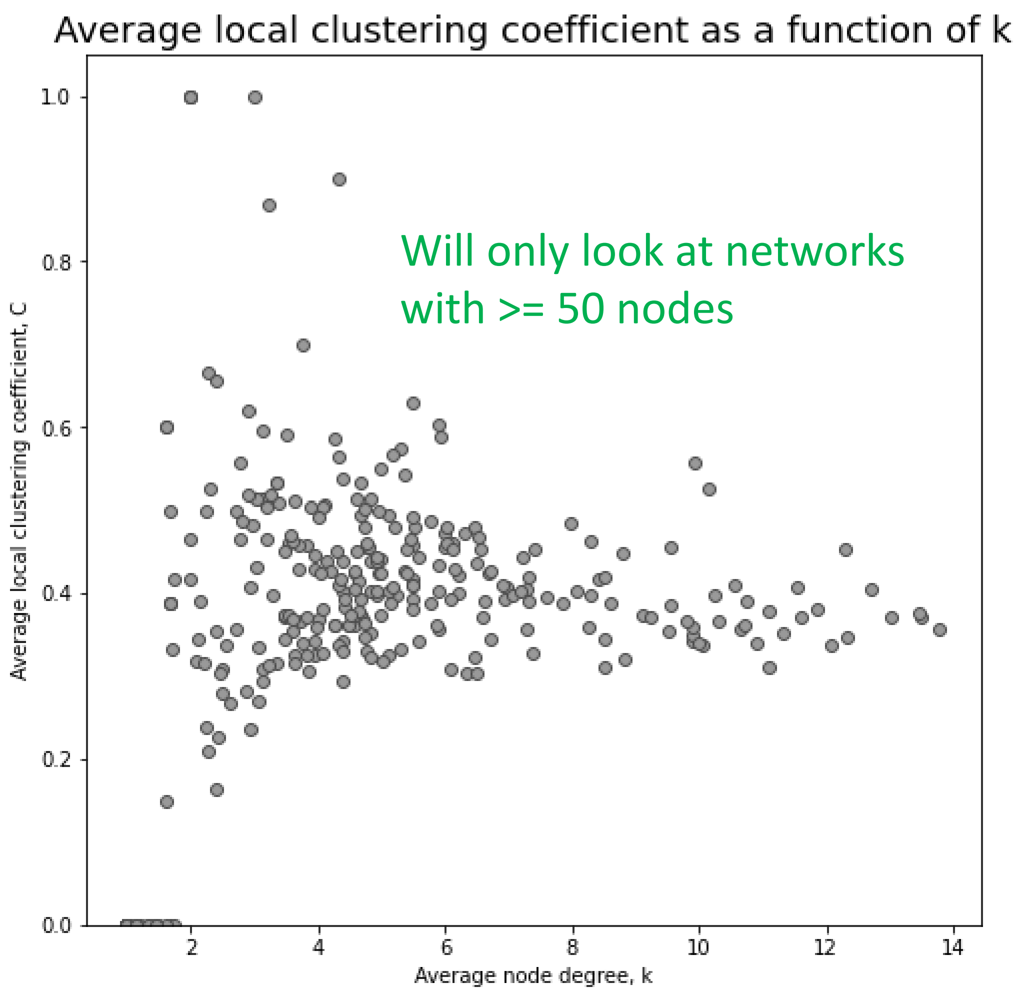
- e.g. largest correlation between AoA of word and its degree was observed in networks generated at this threshold
- May need to try other threshold for best results

Can be done with NetworkX Library

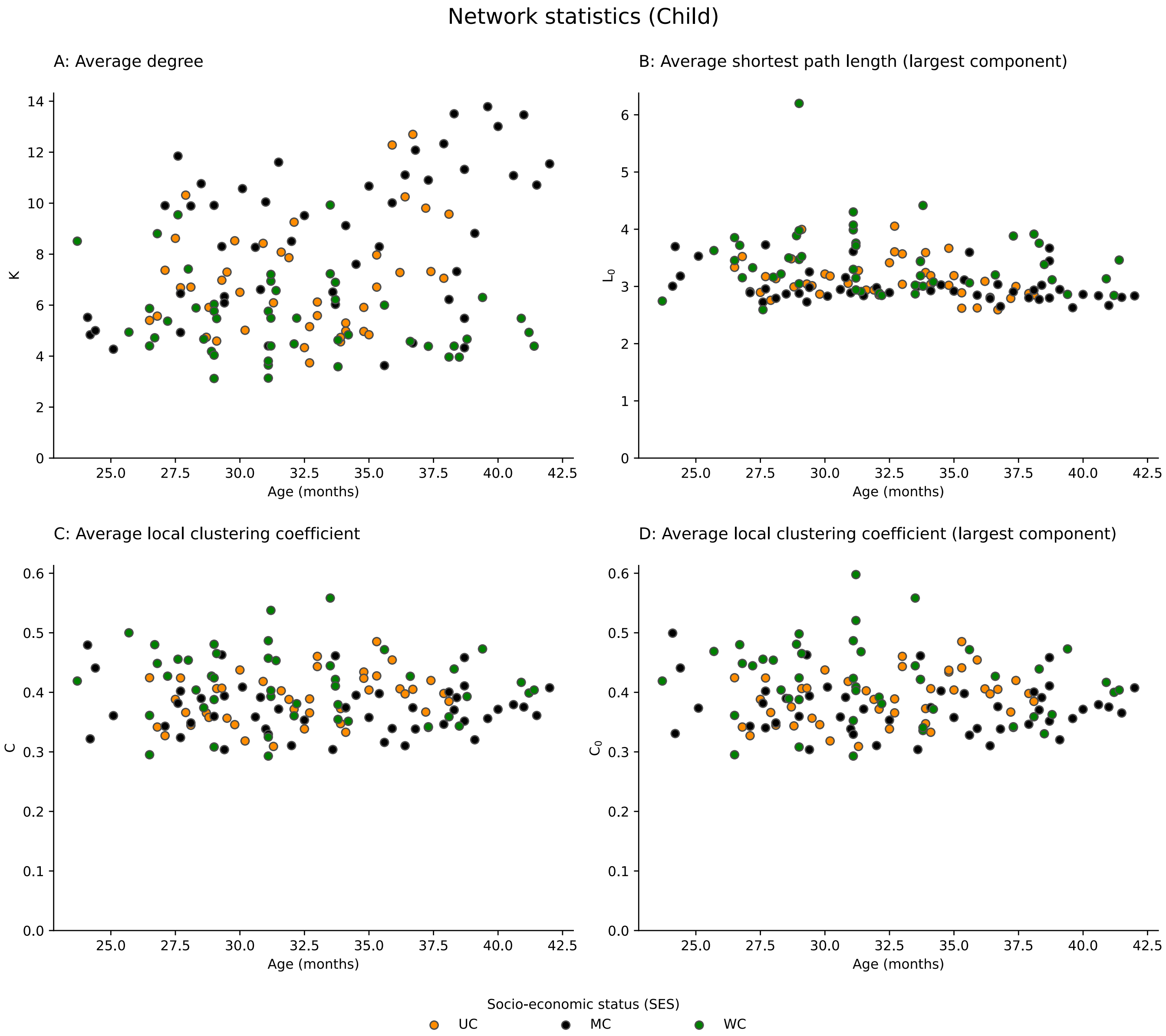


Results so far…

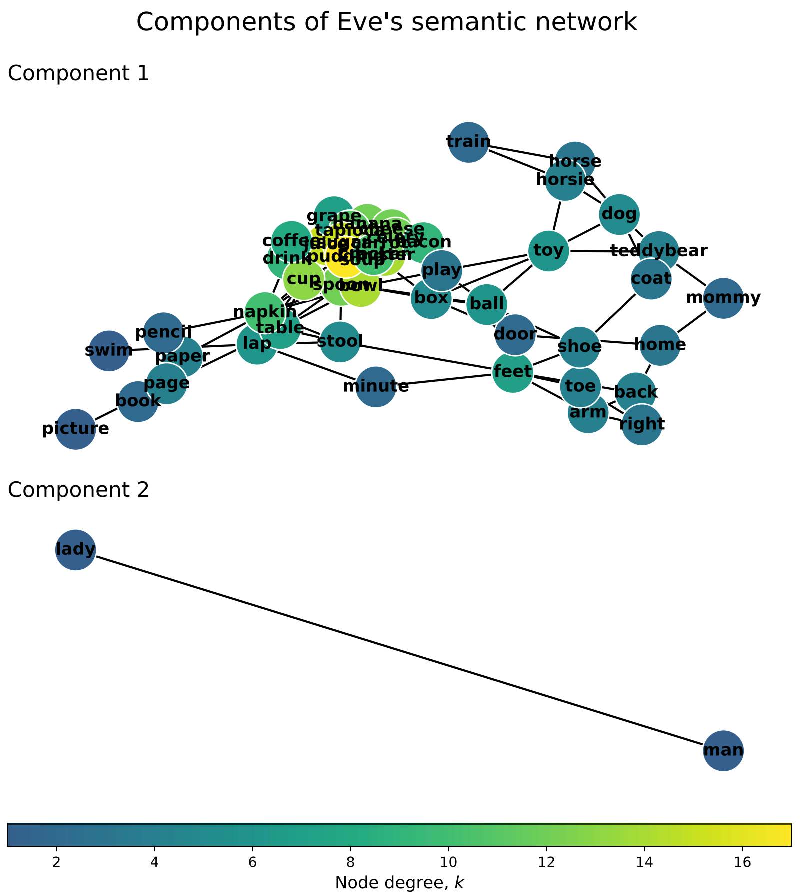




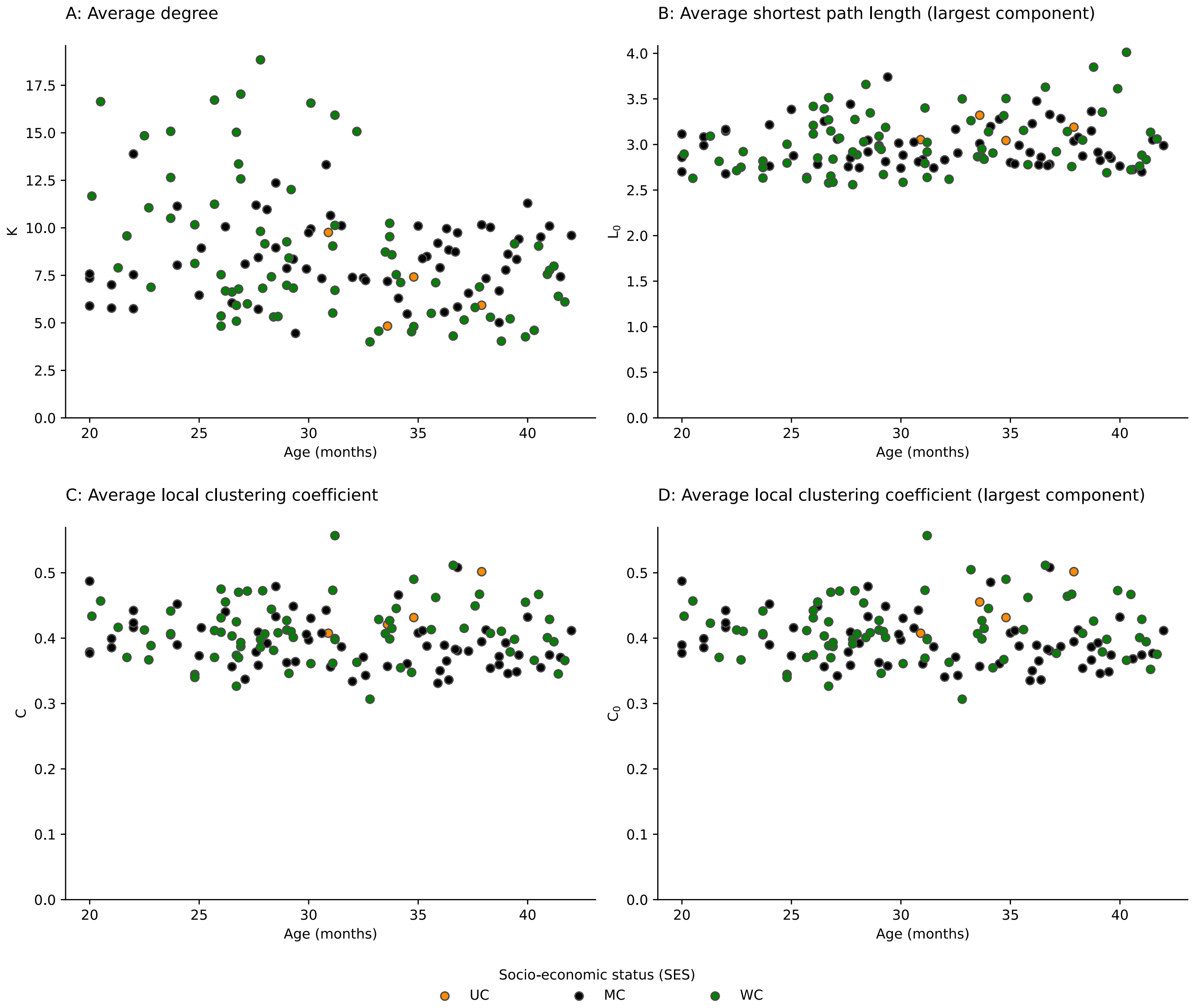
Will only look at networks with >= 50 nodes



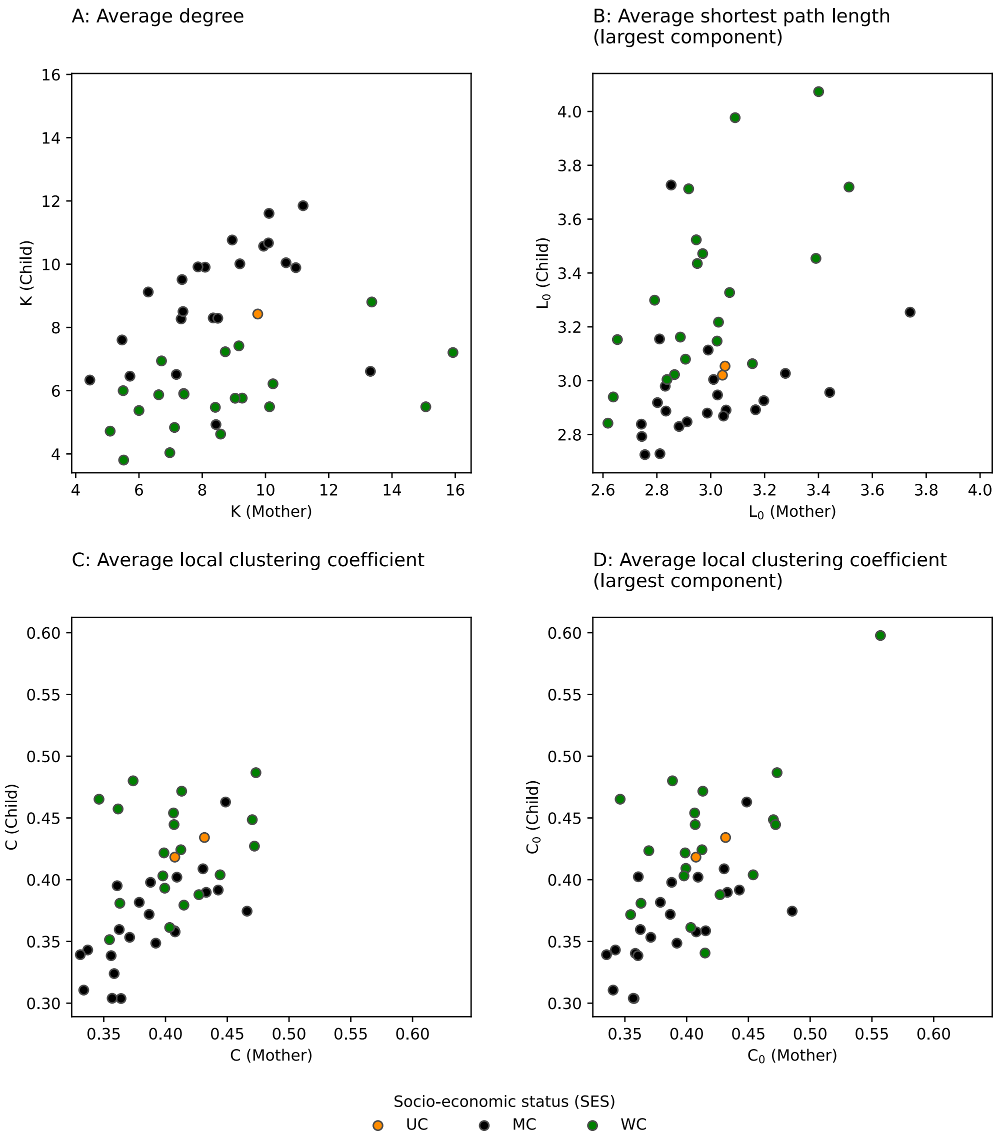
Some networks have more than 1 components



Network statistics (Mother)



Network statistics (Child vs mother, 26-36 months)



Analysis is still on-going…

Question?