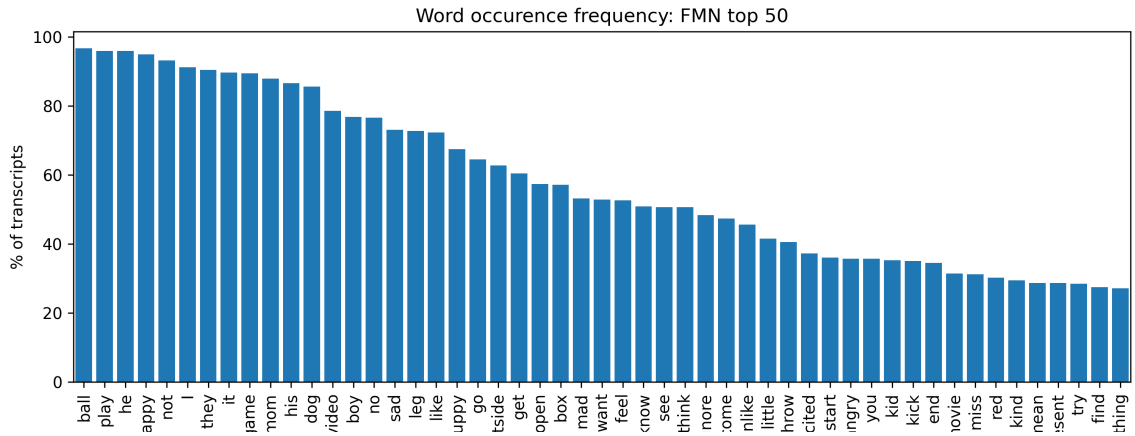
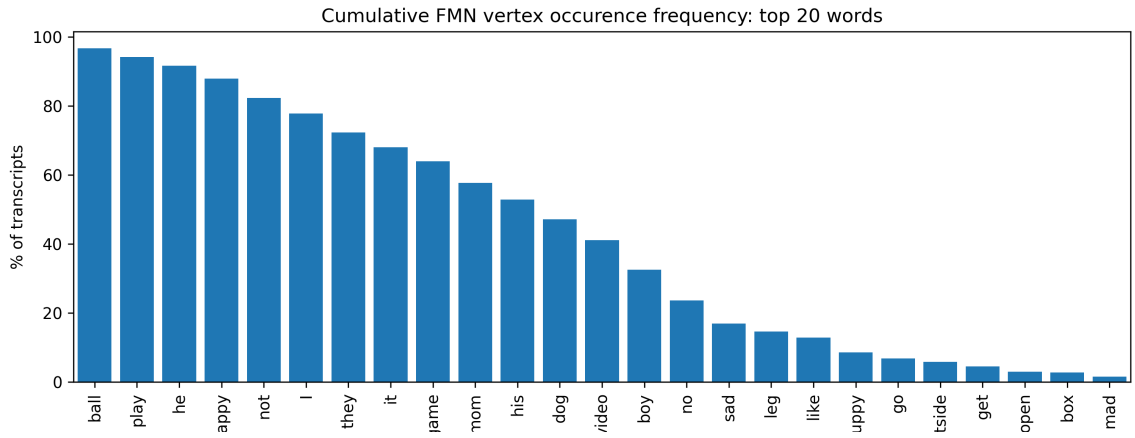
Transcript analysis report

Mutually included words:

No words occur consistently across all transcripts. And when we consider only the words which are useful in forming formamentis networks (FMN), ignoring non-informative words such as link-words, the most commonly occuring words across all transcripts is as folows:



Selecting a subset of these words which mutually appear in all transcripts quickly reduces the number of usable transcripts. The following plot shows the cumulative effect of including the most common words on the % of usable transcripts:

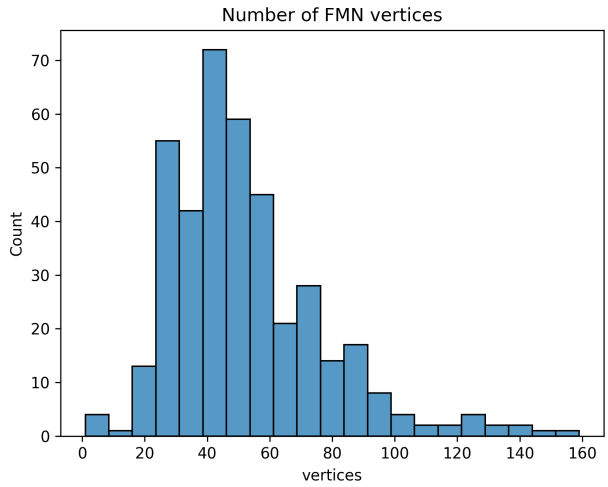


It therefore seems that a model which requires mutual inclusion of specific words will not be viable, as using even the 5 most popuar words reduces the usable transcripts by ~20%, recuding the number of usable transcripts from 389 -> 318.

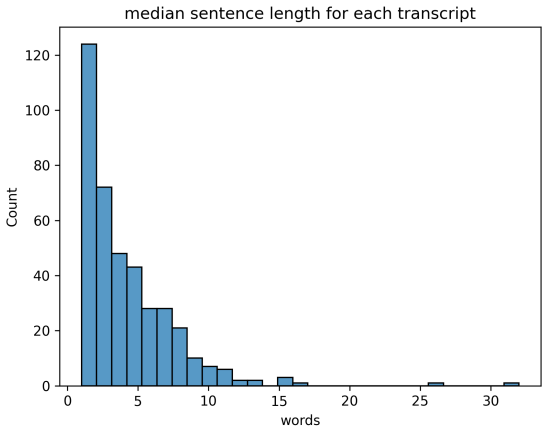
This would also cause the predictive ability of the model on new data to be dependent on in clusion of the specified words.

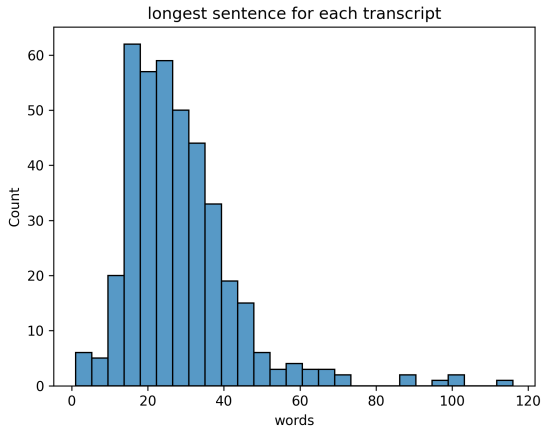
General data statistics:

The average number of FMN usable words in each transcript is 53, with the a small number of transcripts containing very few words:



Most sentences are short, with fewer than 5 words in them:

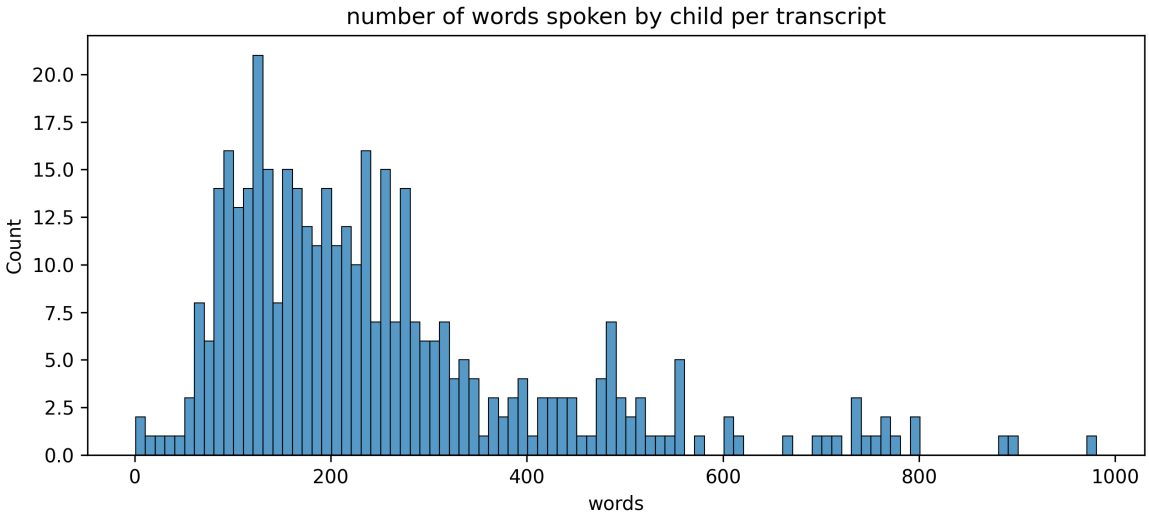




But most children managed at least one sentence of reasonable length.

Transcript lengths:

The follwing histogram shows the total number of words spoken by each child in their entire transcript:


number of words spoken by child per transcript
Count
words
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
0
200
400
600
800
1000