A Network Approach to Lexical Growth and Syntactic Evolution in Child Language Acquisition

Matt Adamo
Samantha Boylan

SI508 Networks
December 15, 2008
Child language acquisition presents an interesting opportunity for network analysis. In basic terms, children go from making sounds, to producing a few words, to simple utterances, and then to more complex, adult-like sentences. It follows that networks built from longitudinal child language data would grow from just a few nodes early on to a dense hairball in the later stages. What would this growth look like? How do certain network metrics change over time, and how do they compare to adult speech? This project is an attempt to answer these questions as well as demonstrate how network analysis can be used in child language acquisition research.

Building on some previous network analyses of natural language, and taking an approach similar to Ke and Yao, we analyze the language development of one child from her first utterances as a one-year old until the age of four. We examine the structure of the network, measure degree centralization, and calculate average outdegree for the approximately 3-year span of the corpus data. We also analyze her mother's network properties as a basis for comparison. We then present our results and ideas for possible future research.

**Related work**

There has been some work relating networks and language in general such as Ferrer i Cancho & Sole (2001), Sole et al (2005), and Motter et al (2002). This work has mostly centered around a proposed small-world structure for language graphs, as well as exploring the different ways of creating networks from language data such as co-occurrence networks, where words are connected if they appear next to each other in a sentence; syntactic networks, where the network can be built up based on connections between sentence components (determiner->subject->verb->object); semantic networks, where the network maps out different semantic relations like isa, part-whole, or binary opposition (Sole 2005).

Applying network analysis to child language acquisition appears to be a fairly new idea. Ke and Yao (2008) use a network approach to analyze the data of 12 English-speaking children ranging in age from 1;8.22 (read: one year; eight months-twenty-two days) to 2;0.25. They look at network size over time, calculate average degree, and examine how the role of the determiners "a" and "the" change over time. While their analysis is interesting, they do not apply centrality measures to their data or identify connected components and lone nodes. We hope to expand upon their research by exploring different methods of network analysis.

**Data and methodology**

Our data comes from the Providence corpus, a longitudinal corpus of transcribed parent-child interactions publicly available as part of the CHILDES database (MacWhinney 2000). Each child was video recorded at home with his/her primary caregiver for about an hour every other week. The recordings began with each child’s first words around the age of 1 and continued for two years. The researchers who developed the Providence corpus decided to extend the data collection period for two
children, Lily and Naima.

We chose to look at the data from Lily. She has about twice as many transcript files as the other children in the corpus, and since our goal was to look at language acquisition over time, it made sense to look at data that spanned as much time as possible. Lily's transcript files begin with her earliest words at age 1;01.02 (read: one year; one month; two days) and end at age 4;00.02. A sample transcript segment can be found in the Appendix. The transcript filenames follow the convention LILxx, where “xx” is the data collection session number. For example, LIL01 is the first data collection session, and LIL80 is the final one. We use this convention in our findings below.

The transcripts are XML encoded, so we used a Perl script to extract Lily’s utterances from each transcript and to format them to be readable by the network analysis tools Pajek and Guess. We extracted her mother’s utterances the same way. Due to an error we were unable to resolve, our Perl script failed to parse the transcript file LIL73. So, in total we ended up with a network file for each transcript except LIL73, for a total of 79 network files.

In the resulting networks, the nodes are the words from the transcript. The edges are directed and indicate a word order relationship. For example, the phrase "a book" would have one edge originating at "a" and directed toward "book." See Figure 1 below for an illustration of the sentence “do it on the floor.”

![Figure 1. Example graph of the sentence "Do it on the floor."](image)

Once the networks were constructed we began collecting data using Pajek. Because we were interested in what the network growth looks like, we counted the nodes and edges for each network file. Building on the work of Ke & Yao, we measured indegree/outdegree centralization in addition to average outdegree, to see if and how those measurements change over time. We also compare average outdegree development to mean length of utterance (“mlu,” a common way of assessing linguistic development) to see how average outdegree changes as Lily’s speech becomes more sophisticated.

We also compared Lily’s networks her mother’s. Because we did not expect Lily’s mother’s speech to vary much over time, we calculated her indegree/outdegree centralization and average outdegree for every fifth network file beginning with LIL05.
All data was stored in a spreadsheet for analysis.

**Findings**

**Node and edge growth**

The growth of nodes and edges in the networks shows some interesting patterns in Lily’s language development. The number of nodes remains below 10 until age 1;4.25. At that point, the number of nodes follows a rough pattern of doubling every few weeks. The most striking increase is between age 1;8.14 and 1;8.28, where the number of nodes goes from 83 to 152. The rapid node growth continues until about age 2, when the number of nodes begins increasing at a slower rate. The number of nodes varies between 300-400 for the rest of the data points.

The edge growth follows a similar pattern. The first time we see any edges is at age 1;4.25. The number of edges grows quickly, with 30 edges by age 1;6.20, 70 edges by 1;8.28, and then doubling at age 1;9.26 to 184 and doubling again at 1;10.09 to 372. The number of edges continues to grow as Lily gets older, even as the addition of nodes somewhat levels out.

![Figure 2. Chart of node and edge growth](image-url)
It's interesting to note the rapid growth of nodes followed closely by the rapid growth of edges. The data also show that node growth somewhat levels off at age 2, but edge growth continues throughout the data collection period. These two observations demonstrate a different way of examining the relationship between lexical growth (increase in nodes) and linguistic complexity (increase in edges), an advantage of our network approach. Future work might look for these patterns in other children.

**Indegree & outdegree centralization**

Centralization measures can give us some finer-grained insight into the structure of the network. Measuring indegree and outdegree centralization shows us how uniformly indegree and outdegree centrality are distributed across the network. A higher centralization means indegree/outdegree centrality is concentrated in a few nodes; a lower centralization means it is spread more evenly among each node in the network. Looking at how this measurement changes over time, we can see if patterns of word usage change as Lily gets older.

![indegree & outdegree centralization](image)

*Figure 3. Indegree & outdegree centralization for Lily and her mother*

The data show that indegree centralization stays relatively low throughout Lily's language development. In fact, after a brief initial period of very low centralization, Lily’s indegree centralization reaches a level similar to her mother’s.
Lily’s outdegree centralization shows a more interesting pattern. It starts off low, increases rapidly to a point at 1;11.07, and then decreases over time until it falls into a range similar to her mother’s outdegree centralization. A look at the network graph (Figure 4 below) for this time point demonstrates why we see the sharp increase at 1;11.07.

![Network graph of LIL21, age 1;10.09](image)

**Figure 4. Network graph of LIL21, age 1;10.09**

A higher outdegree centralization means that there are a few nodes with high outdegree and many others with low. The graph shows that words like “and,” “a,” “its,” and “is,” are acting like hubs within the network. Since these are the only hubs within the network, outdegree centralization increases. As Lily acquires additional function words over time (“the,” “this,” “that,” “in,” etc…), the number of nodes with high outdegree increases, causing the decrease in centralization that we observe.

This finding raises some questions for future research. The first question is whether the spike and subsequent decrease in outdegree centralization is a feature of this particular data, or if it is a pattern we can expect to find with most children. The second question is
whether we see this pattern only in English-learning children or if it occurs with other languages as well. If children learning languages other than English do show this pattern, what words form the hubs? This would be especially interesting in languages such as Japanese that lack determiners like "a" and "the," common hubs in our data.

**Increasing average outdegree & mlu**

Average outdegree increases as Lily gets older, which is consistent with other types of networks that grow over time (Leskovec 2005). Average outdegree also follows a pattern of growth similar to mean length of utterance (mlu).

![avg outdegree & mlu](image)

*Figure 5. Average outdegree and mlu*

We see a sharp increase in Lily's average outdegree between 1;9.12 and 1;10.23, after which it continues to increase. This suggests that this might be a time period worth examining more closely, an insight that the plot of mlu does not reveal.

Lily's mother's data show that average outdegree is variable, but generally falls between 2.5 to 4. Lily's average outdegree enters this range at about age 2;2.22.

An increasing average degree has several implications that could make it a useful measurement of linguistic development. As Ke & Yao mention, a high average
outdegree implies increasing lexical variety. Figure 6 below shows the graph of LIL45 with the nodes sized according to outdegree. We see that there are several prominent nodes, including the determiners “a” and “the.” This means they have a high outdegree, and therefore connect to many different nouns. We also see that the pronoun “I” is prominent, meaning it must connect to a variety of verbs.

Figure 6. Network graph of LIL45, age 2;5.27, mlu = 2.846

It is not taken for granted that longer sentences (higher mlu) imply a higher outdegree. Even if, for example, this sentence were converted to a graph, the average outdegree would only be 1. Average outdegree considers how all of the nodes in the network interact, and it only increases when nodes connect to larger amounts of unique nodes. We interpret this to be a useful measure of linguistic sophistication.

Conclusion

Our findings show that network analysis can be a useful way of interpreting longitudinal child acquisition data. Most significantly, our findings show how a network analysis can pinpoint interesting time periods during the acquisition process. We also show how a visualization of a network can provide an understanding of child speech different from word counts and mlu calculations.
This project only considered one child, so it’s unreasonable to draw any conclusions about language acquisition from our results. However, our results are interesting enough to merit further research, and they suggest that our analysis techniques could be useful tools in future child language acquisition research.
References


Appendix

Sample transcript segment – LIL45

@Begin
@Languages:  en
@Participants:  CHI Target_Child, MOT Mother, OPE Camera_Operator
@ID:  en|providence|CHI|2;5.27|female|||Target_Child||
@ID:  en|providence|MOT|||||Mother||
@ID:  en|providence|OPE |||Camera_Operator||
@Birth of CHI:  13-JAN-2001
@Date:  10-JUL-2003
@Coder:  Marisa Colas

*CHI:  I'm gonna [: going to] put this xx +...  %mov:"lil45"_0_5423
%pho: ˈæmɡəʊŋəʊ φΙς
*MOT:  hey Amanda do you happen to know +..?  %mov:"lil45"_5421_6826
*MOT:  Manuela had mentioned to me +...  %mov:"lil45"_6940_8696
*CHI:  xx .  %mov:"lil45"_8695_9641
*MOT:  oh .  %mov:"lil45"_9641_13413
*CHI:  yy .  %mov:"lil45"_13413_15643
%pho: ɪ̬jiks
*CHI:  yy .  %mov:"lil45"_15643_17785
%pho: ˈsɪvn
*CHI:  seven eight .  %mov:"lil45"_17785_20878
%pho: ˈsɛvˈɛst
*CHI:  nine .  %mov:"lil45"_20878_24420
%pho: ˈnɔn
*MOT:  hey Lily .
*CHI:  nine ten eleven twelve .  %mov:"lil45"_24420_31158
%pho: ˈnɔnˈtənˈlɛnvˈtwɛv
*MOT:  do you want, do you want to tell Amanda how you count to twelve in Spanish?
*MOT:  uno .  %mov:"lil45"_31156_32823
*CHI:  dos `tres@u [: tres] `kwatro@u [: cuatro] cinco seis
   si`etes@u [: siete] `otfo@u [: ocho] nu`cuet@u [: nueve]
   +...  %mov:"lil45"_32823_41461
%pho: ˈdɔs `twɛs`kwɔntsɪko`sɛs`sə`ɛtɛɾ`otfɔ`nwevɛɾ
*CHI:  xx .  %mov:"lil45"_41461_43491
*MOT:  diez .
*CHI:  once .  %mov:"lil45"_43491_46423
%pho: ˈənset
*MOT:  once .
*MOT:  once ?  %mov:"lil45"_45368_46308
*CHI:  `doseʃ@u [: doce] .  %mov:"lil45"_46306_50123
%pho: ˈdæʃɛɾ
*MOT:  yay .
*OPE:  good job .
*CHI:  one .  %mov:"lil45"_50123_51728
%pho: ˈwʌn
*MOT:  Dora the explorer .