INX 199 – Assignment 2
The Make-up of Early Child Vocabulary

Due November 9, 2005 10:10am

1 Introduction

Nouns and verbs have different functions in language, and hence children learn them differently. Nouns often refer to objects in the outside world, e.g., *apple*, *ball*, *dog*. To learn a noun, a child needs to establish a correspondence between the word and its referent in the outside world. Verbs, on the other hand, are more complex: they express states, actions, or events, e.g., *is*, *eat*, *fall*. For learning the meaning of a verb, a child has to learn all the components involved in the state or the event described by the verb. For example, to learn the verb *eat*, a child should learn that it requires two participants, one performing a particular action on the other. The child should also learn that the former participant appears before the latter in a sentence, e.g., *John ate an apple*.

Psycholinguists often assume that since learning nouns is less complex, the rate of acquisition of nouns is faster than that of verbs for almost all children. To verify this assumption, we can look at the syntactic category of the words that a child uses at different periods of time. If the assumption is true, we expect that the child uses more nouns than verbs at an early age. As the child gets older, s/he starts using more and more verbs.

In this assignment, you are asked to verify this assumption by looking at the patterns of word usage in real sentences produced by children. In other words, you will analyze the words that a child uses at different time periods to see whether you observe a difference in the proportion of nouns and verbs.

2 Type and token frequency

A type refers to a different word, while a token is a particular instance of a word. Given a particular text, type frequency is the number of different words used in the text, while token frequency is the total number of words appearing in it. For example, in "*A dog is chasing the* *cat that is chasing another dog.*", there are 8 word types (i.e., type frequency is 8), but 11 word tokens (i.e., token frequency is 11). Note that the word *dog* counts as a single type, no matter how many times it occurs in the text. We can also determine the type and token frequency of words from different syntactic categories (e.g., verbs and nouns) in a given text. In the above text segment, noun type frequency is 2 (*dog* and *cat*), noun token frequency is 3 (2 instances of *dog* and one instance of *cat*); verb type frequency is 2 (*is* and *chasing*), and verb token frequency is 4 (2 instances of each verb).

3 CHILDES

For this assignment, you will be working with real child language. The data you will be using is part of CHILDES (Child Language Data Exchange System), available at *http://childes.psy.cmu.edu/* (MacWhinney 1995). CHILDES contains a collection of child language transcripts, as well as a number of Child Language ANalysis programs (CLAN). You will use some of the CLAN programs to analyze utterances produced by a child in order to make inferences about the child's vocabulary. The following subsections describe the data and the tools in more detail.

3.1 Data files

The CHILDES transcript collection consists of a number of files, each containing utterances produced by a child at a particular age. The name of each file has three components: the name of the child whose utterances are recorded in the file, a sequence number, and the extension *.cha*; e.g., *adam01.cha* and *sarah011.cha*. For this assignment, you work with two utterance files produced by Adam at different ages, **adam01.cha** and **adam20.cha**. You can copy the files from the following directory on CDF: *~t3fazlya/inx199/PUBLIC/*, or download them from the course web page: *http://www.cs.toronto.edu/~suzanne/199/assignments.html*.

The general format of an utterance file is shown here:

PART I.	Lines starting with **@**	Personal information about the session, and the participants.
PART II.	Lines starting with *****	Utterances produced by the child or a caretaker, e.g., the mother.
	Lines starting with **%**	Extra information about the scene.

From PART I, you need only the child's age, marked as **@Age of CHI**. Age is given in this format: **y**;**m.d**, in which **y** shows the year, **m** shows the month, and **d** shows the days. For example, if you look into the utterance file **adam20.cha**, the age is **3;0.10**, meaning Adam was 3 years and 10 days old when recording the set of utterances in this file. From PART II, you will only be looking at the utterance lines. Utterances are marked by a ***** followed by a code identifying the speaker, such as **CHI** for the child, and **MOT** for the mother.

3.2 CLAN programs

CLAN contains programs for the analysis of child language transcripts in CHILDES. The motivation behind creating CLAN was to automate analyses that were widely used by researchers in the field. Examples of the CLAN tools include programs for counting items, such as **freq**, and programs for the computation of some statistics over the transcripts, such as **mlu** for calculating the mean length of utterance (the average number of words in a sentence).

For this assignment, you only need to use the **freq** program. To run a program, you need to include 3 main parts: the command name (here, **freq**), a list of options to use the command in specific ways, and the name of the utterance file (e.g., **adam01.cha**). Here is the syntax of a typical CLAN command:

≺command-name≻ ≺options-list≻ ≺filename≻

Note that you can include more than one option for a command. Each option has 3 parts: **+/-** for turning the option on or off, respectively; the option's name that is often a single letter, such as **t** or **s**; and the option's specific value. Here are some of the useful options for the **freq** command, along with some examples on how to use them:

Option's Name (Option's Value)	Usage Examples	Description
t	+t*CHI	Looks into child's utterances only.
(speaker code)	-t*CHI	Looks into all but child's utterances.
s	+s"dog"	Looks only for the word *dog*.
(what to search for)	+s"wh*"	Looks for all words beginning with *wh*.
	-s"wh*"	Looks for all words but *wh*-words.
	+s@*myfile*	look for words listed in the file named *myfile*.

And here is a sample usage of the **freq** command:

freq +t*CHI -s"wh*" sarah011.cha

This command returns the frequency of all words except those starting with *wh* (-s"wh*"), that are used by the child (+t*CHI), in the specified file, *sarah011.cha*. Note that the spacing is important: you are required to separate the command name, different options, and the file name by a space; but you should not put spaces between the parts of a single option. Note also that it is possible for some options to be omitted. For example, the following usage of **freq** returns the frequency of all the words in *sarah011.cha* that are used by the child:

freq +t*CHI sarah011.cha

4 Your tasks

4.1 Task I

For each utterance file, **adam01.cha** and **adam20.cha**, do the following:

1. Find Adam's age at the time of recording the utterances.

2. Find the frequency of all words that Adam uses, and store them in a file for future reference. Recall that you can calculate the frequency of occurrence of the words in an utterance file using the **freq** command. Remember also that you should only look for words used by the child. Note that **freq** considers different word forms as different words (e.g., "go" and "going" are two different words). You should follow the same rule in this assignment.

   Record your usage of the **freq** command, for reporting to us. (See table in Task III below.)

3. From the words found in 2., find those that the child uses at least 10 times. You can use the Linux **sort** command to make this task easier.

   Adam uses words from all syntactic categories, such as nouns, verbs, adjectives, adverbs, pronouns, etc. The focus of this assignment is on nouns and verbs only; hence you should only consider words from these two categories. To decide whether a word is a noun or a verb, you can consult a lexicon (**LEXICON**) that we provide for you. **LEXICON** contains an alphabetically-ordered list of nouns and verbs, marked by their syntactic category (N for nouns and V for verbs). You can copy **LEXICON** from CDF or download it from the course web page.

   Note: For steps 4 through 7, consider only nouns and verbs with frequency of 10 or greater.

4. Separate the words into Nouns and Verbs, and find out the number of times each word is used (its frequency).

5. Calculate noun token frequency (child_noun_token), and noun type frequency (child_noun_type).

6. Calculate verb token frequency (child_verb_token), and verb type frequency (child_verb_type).

7. Calculate the percentage of noun types (%child_noun_type), and the percentage of verb types (%child_verb_type). These values show what percentage of the child's productive vocabulary (at a particular age) is composed of each category.

4.2 Task II

Repeat steps 2 through 7 above, but this time you should look into patterns of noun and verb usage in the child-directed speech (CDS), i.e., utterances produced by the child's mother, father, or any other adult. As before, record your usage of the **freq** command (step 2 above).

Choose new names for the percentage of noun types, and the percentage of verb types in the child-directed speech (step 7 above), i.e., %CDS_noun_type and %CDS_verb_type, respectively. From this Task, you only provide the syntax of the **freq** command you used, and the values of %CDS_noun_type and %CDS_verb_type. (see Task III.)

What you hand in for Tasks I and II

Organize the information you gathered so far (in Task I and Task II) into two tables, one for each utterance file. A sample table is shown here:

file: **sarah001.cha**	
child: Sarah	
age: 2;1.0	
I. syntax of **freq** command used for Task I.	
II. syntax of **freq** command used for Task II.	
Nouns	Verbs
(ordered alphabetically, one per line)	(ordered alphabetically, one per line)
baby: 18	*eat*: 14
ball: 23	*go*: 10
dog: 12	
child_noun_type: 3	child_verb_type: 2
child_noun_token: 53 (18+23+12)	child_verb_token: 24 (14+10)
%child_noun_type: 60%	%child_verb_type: 40%
%CDS_noun_type: 50%	%CDS_verb_type: 50%

4.3 Task III

What you hand in for Task III

Give brief answers (one or two paragraphs) to the three questions on the following page.

Note: Answer the questions in order, and in answering each question, only draw on the evidence explicitly mentioned up through that question.

Q1: From the tables, find out what percentage of Adam's productive vocabulary is composed of each category (N or V) at each age. Relate your findings to class discussions on the early acquisition of verbs and nouns.

Q2: Do you observe a difference in the percentages of noun types and verb types for the child and the child-directed speech? Do you see any meaningful correspondence between the composition of the child's lexicon and the patterns of word usage by the caretakers? What does this tell you about the child's language development?

Q3: Across the two ages, Adam uses some of the same words (N and V types), and some different words.

- How many noun types occur in both lists of noun types across the two ages? List the nouns in common on the two lists.

  Answer the same questions for verb types: How many verb types occur in both lists of verb types across the two ages? List the verbs in common on the two lists.
- Consider that Adam knows more words than he uses in any particular situation (recording session). Give a possible explanation for the different amounts of overlap you see for noun types and verb types. Draw on what you know in particular about the learning and use of nouns among children.
- What does this pattern of overlap say about the issue of early noun and verb learning, and your answer to **Q1**? Does this new data change your conclusion above?

5 Working at home

You can work at home, and on any operating system other than Linux (what we use on CDF). However, we cannot provide you with any help if you decide to do so. You can download the CLAN programs from the CHILDES web page.

6 References

MacWhinney, B. (1995). The CHILDES project: tools for analyzing talk. Hillsdale, NJ: Lawrence Erlbaum.