INX 199 – Assignment 3
Looking for Patterns of Overregularization

1 Introduction

How does a child learn to use *apples* to refer to two apples standing on a table, but *apple* when referring to the single apple she is holding in her hands? How does a child learn that she should use the form *kicked* (instead of *kick*) when telling her mother that she just kicked a ball? There are two possibilities. One is that children learn different forms of a word by imitation: they hear people around them referring to two apples using the plural form and to one apple using the singular form, and they just learn it. One important question arising from this theory is how children learn the plural form of a word when they have not heard it before. What would a child who knows the word *bike*, but has never heard the plural form *bikes*, say when referring to two bikes?

Another possibility is that children do not simply learn by imitation, rather they extract rules from what they hear. The idea is that, for example, after a child hears a sufficient number of words in both singular and plural forms, she extracts a rule that says "pluralize a word by adding an *-s* to the end of it". One side effect of this learning mechanism is what psycholinguists refer to as **overregularization** or **overgeneralization**. After learning a rule, a child may apply it to cases where the rule is not applicable. For example, the child may use the wrong form *foots* instead of *feet*, or may say *I eated*, or even *I ated*, instead of *I ate*. In fact, strong evidence provided in favor of this theory is that children have been observed to utter sentences that contain exactly these types of overregularization errors.

In this assignment, you will be looking for patterns of overregularization in real sentences uttered by three children: Adam, Eve, and Sarah. In other words, you will analyze the utterances of these children at different ages to see what kinds of overregularization errors they make. You will also look for patterns of error-making over the course of a child's language development.

2 Data

As in Assignment 2, the data that you will be working with is part of the CHILDES collection, available at *http://childes.psy.cmu.edu/* (MacWhinney 1995). For this assignment, you work with a number of utterance files from Adam, Eve, and Sarah. (The format of the CHILDES utterance files is described in Assignment 2.) You can copy the files from the following directories on CDF: *~t3fazlya/inx199/PUBLIC/Adam*; *~t3fazlya/inx199/PUBLIC/Eve*; *~t3fazlya/inx199/PUBLIC/Sarah*. Each directory also includes a file called Mappings.<ChildName> (e.g., Mappings.adam) which, for your convenience, gives you the age of the child at the time each file was recorded. It is important that you keep the files for each child separate. You can do this by creating subdirectories named Adam, Eve, and Sarah on your home directory.

3 Your Tasks

3.1 Task I (30 marks)

For this task, you will be looking at Adam's utterances only. You are asked to find different kinds of overregularization errors Adam makes, across different syntactic categories (such as nouns, verbs, etc.). You can do this through the following steps (see below for what you hand in for this task):

1. Use the **freq** command to find instances of errors (and their frequency) in all Adam's utterance files. All erroneous words end in the symbol **@n**. For example, *maked@n*, *saidn't@n* are some of the errors that Sarah makes.

   Remember from Assignment 2 that you can use the **+s** option to look for all words matching a particular pattern. Here you are looking for all words ending in **@n**, so you can use **+s**"*@n".

   You can run **freq** on more than one utterance file at once; for example, the following usage of **freq** looks for the word *mommy* in all child speech in Adam's utterance files in the given directory:

   freq +t*CHI +s"mommy" <*dir*>/Adam/adam*.cha

   where <*dir*> is your home directory (or wherever you created the subdirectory Adam).

   This command gives you the frequency of occurrence of *mommy* in each of Adam's files. Here is an excerpt of the output:

   From file <<*dir*>/Adam/adam53.cha>
   29 mommy

1	Total number of different word types used
29	Total number of words (tokens)
0.034	Type/Token ratio

   From file <<*dir*>/Adam/adam55.cha>
   48 mommy

1	Total number of different word types used
48	Total number of words (tokens)
0.021	Type/Token ratio

   For this task, you only need the words and their frequencies, so you can ignore the Type/Token information. You should store the output of this stage in a file for future reference.

2. Examine the erroneous nouns you have found in Step 1 to see what kinds of overregularization errors Adam makes on nouns. Note that simply listing the erroneous nouns is not an acceptable answer. Record a short description of the error (no longer than a sentence) along with one or two examples of each kind to hand in (see below for what you hand in).

3. Repeat Step 2 for verbs.

4. Repeat Step 2 for other syntactic categories (e.g., pronouns, adjectives, prepositions). State the category of the word(s) in the description of each kind of error.

Note (Steps 2 through 4): You need to provide a short description for each kind of error that you find in Steps 2 through 4 above. To provide an appropriate description of an error, you may need to consult the original utterances (sentences) the error appears in. You may also need to consult the original utterance files to decide what syntactic category an instance of a word belongs to. Note also that there might be some errors marked by **@n** that are not due to overregularization.

What you hand in for Task I

I.A. (20 marks) Organize the information you gathered for Adam into a table; here is a sample:

Task I: Adam

Syntactic Category (e.g., Nouns)		
Error Kind 1	description of the error	examples
Error Kind 2	description of the error	examples
...		
Syntactic Category (e.g., Verbs)		
Error Kind 1	description of the error	examples
Error Kind 2	description of the error	examples
...		
...		

I.B. Write down your brief answers to the following questions. Keep your answers concise and provide convincing evidence for your claims.

(i) (5 marks) Are there any qualitative differences in the kinds of error you observe across different syntactic categories?

(ii) (5 marks) At which age does Adam start marking overregularization errors? Why do you think Adam does not make such errors before this time? Is there a time when Adam starts recovering from such errors? If yes, when is that? If no, what do you think the reason is?

3.2 Task II (45 marks)

For this task, you will be looking at utterances of Eve and Sarah (separately for each child). You will be searching for patterns of error-making for a particular kind of overregularization error, i.e., incorrect past tense forms of irregular verbs. The verbs (in base form) are { *come*, *do*, *fall*, *go*, *make*, *take* }.

Repeat the following for each child:

1. Use the **freq** command to calculate the frequency of occurrence of each verb, in all correct and incorrect past tense forms, in each of the child's utterance files.

   NOTE: Recall from Task I that many of the erroneous words are marked with a **@n**. When looking for an error, e.g., *saidn't*, you should thus look for the word with and without the marker **@n** at the end, e.g., *saidn't* and *saidn't@n*.

   Also recall from Task I that you can run **freq** on a set of files. In addition, **freq** can be used to look for a list of words at the same time. This can be done by creating a file containing one word per line, and using the option **+s@***myfile*, where *myfile* is the name of the file. A sample file is given here:

   ```
   my
   mine
   me
   ```

2. Calculate the percentage of correct usages (p_c) of each verb at each particular age. Note that each utterance file corresponds to a session in which the utterances are recorded for a child. Hence it is possible that two or more utterance files for a child are recorded at the same age. In your calculations, you must sum up the frequencies across two or more files with the same age for the child. Use the Mappings.sarah and Mappings.eve files to make it easy to match up the files with the same age.

   Note also that you should calculate the percentage of correct usages of a verb only when the verb is used (either correctly or incorrectly). When the frequency of all the correct and incorrect forms are zero, the percentage is not defined. When all frequencies are zero, it means that there has been no usage of the particular verb, and hence we cannot deduce anything about the percentage of correct usages.

What you hand in for Task II

II.A. (30 marks) Organize the information you gathered for Eve and Sarah into two tables, one for each child, as shown below. In each table, list the frequency and percentage information for each verb, given in alphabetical order from left to right. Replace <ChildName> in the sample table with Eve or Sarah as appropriate, and replace <correct> and <incorrect> in the sample table with the correct and incorrect past tense forms you used for each verb in your searches.

Task II: <ChildName>

	Verb: *come*			···	Verb: *take*		
Age	<correct>	<incorrect>	p_c	···	<correct>	<incorrect>	p_c
<2;3.19>	2	0	100%	···	0	0	——
<2;4.0>	0	0	——	···	1	1	50%
<2;6.10>	5	2	71%	···	4	2	67%
<2;8.0>	2	2	50%	···	0	0	——
<3;0.10>	0	2	0%	···	4	0	100%
···							

For your own use, you should sketch the percentages you calculated for each verb (and each child) into a bar graph. A sample graph is shown in Figure 1 below. The percentages are taken from the sample table shown above, for the verb *come* for <ChildName>.

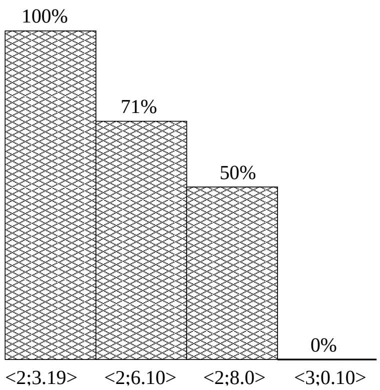

Figure 1: Verb *come* for <ChildName>

Note: **Only turn in the tables of information; do not hand in the graphs**. However, you will need the graphs for answering the following questions.

II.B. Write down your brief answers to the following questions. You should consult your tables and/or bar graphs to answer the questions. As for Task I, your answers should be concise, and your claims should be backed up with convincing evidence and/or reasoning.

(i) (5 marks) Is there a verb (or verbs) among those given to you for which Eve and Sarah mostly use the correct past tense form? Which verb(s) and why? (You should be able to answer this question in two or three sentences.)

(ii) (5 marks) Within the data for each child, do you observe a difference in the learning curves (reflected in the bar diagrams) of different verbs? Discuss possible reasons for your findings for each child.

(iii) (5 marks) Do you observe a difference in the learning curves across the two children? Discuss possible reasons for your findings.