

# **A Grammatical Coding and Analysis System for Language Data from Normal and Brain-Damaged Children**

*Susan Curtiss*<sup>1</sup>, *Jeannette Schaeffer*<sup>1</sup>,  
*Tetsuya Sano*<sup>2</sup>, *Jeff MacSwan*<sup>1</sup>, *Todd Masilon*<sup>1</sup>

---

<sup>1</sup> *UCLA Psycholinguistics Laboratory, 2210 Campbell Hall, Los Angeles, CA 90095, USA*  
<sup>2</sup> *3-2-16 Kanaoka, Higashiosaka-shi, Osaka 577, JAPAN*

KEYWORDS: language-development, coding, analysis

AFFILIATION: <sup>1</sup>UCLA, <sup>2</sup>Meiji-Gakuin University

E-MAIL:           ibenajq@mvs.oac.ucla.edu  
                      schaeffe@biology.ucla.edu  
                      sano@tsc.tohoku.gakuin.ac.jp  
                      macswan@ucla.edu  
                      masilon@humnet.ucla.edu

FAX NUMBER:   (310) 206-5743

PHONE NUMBER: (310) 825-8111

## **A Grammatical Coding and Analysis System for Language Data from Normal and Brain-Damaged Children**

### **1. Background**

This abstract will describe a grammatical coding and analysis system devised to study language development in normal and brain-damaged children, but which can be used to analyze mature as well as developing grammars.

The research project for which this system was developed investigates the lateralization and localization of language in its development. This research program studies language development in children who have catastrophic; i.e., medically intractable epilepsy for which they undergo surgical resection of the diseased tissue. The surgeries range from unilobar resections (e.g., temporal lobectomy), to multilobar resections (e.g., temporal-parietal-occipital lobectomy or, in more extreme cases, hemidecortication (removal or disconnection of the entire cortex, often referred to as hemispherectomy)). The effects of disease and removal of different parts of the left and right hemisphere at different ages are examined and compared – left vs. right, one age vs. another, and, importantly, brain-damaged child vs. normal developing child.

The research focuses on several questions, including: 1) the capacity of each hemisphere alone to subserve lexical and grammatical development (a comparison of left-hemispherectomized and right-hemispherectomized children with each other and with normally developing children), 2) the development of lateralization and localization of grammar (specifically, syntax and morphology) as opposed to lexicon, 3) the effects of localized brain damage on specific functional subsystems of the grammar; namely, the D(eterminer)-system, the I(nflectional)-system, and the C(omplementizer)-system, 4) the effects of localized brain damage on lexicon – the establishment of a mental dictionary of content words and their interrelations – as opposed to syntax and morphology, and 5) maturational constraints on grammar acquisition, again syntax and morphology. The research is part of a multi-disciplinary investigation involving the study of psychopathology, neurophysiology and neuropathology as well as linguistics, and other research questions are directed at better understanding any relationship between linguistic development and function in other areas. These research questions include whether there is a direct relationship between thought disorder and developmental linguistic deficits and whether we find an association between specific patterns of linguistic delay or anomalies and specific neuropathology.

As can be seen, many of our research questions can only be answered in relation to a comparison with normal language development. However, the patterns and range of what is attested in lexical development and the acquisition of morphology and syntax in normally developing children have not yet been fully established. Our study thus involves considerable data collection and analysis of language from normally developing children to enhance the data base which we can then use as an appropriate metric against which the brain-damaged children in our study can be compared.

The study is a five year, longitudinal study, designed so that we can document language development; i.e., change over time, not just language performance at a single point in time. For the study, documentation and assessment of linguistic performance for each “surgical” child is carried out pre-operatively, at 6-months post-surgery, 12-months post-surgery, and then at yearly intervals. For each of these data points, data from a normal, matched control are collected.

Both formal test performance and observational data are collected. However, the major data source for our evaluation of the language performance of each child in the study is the language sample. It is for the analysis of these spontaneous speech samples that our coding and analysis procedures were developed.

## 2. Coding System

Below is a brief description of our coding and analysis conventions. These conventions are theoretically grounded in and motivated by Government and Binding theory as outlined by Chomsky, 1981, 1986; Pollack, 1989; and others. They arose from the need to capture morphological and syntactic distinctions and generalizations than are afforded by the present CLAN commands of the CHILDES database. Insofar as our coding system uses the existing CHAT transcription format and is amenable to existing CLAN commands, we offer it as a useful extension to existing CHILDES capabilities.

Our coding scheme adds to the speaker tier of the CHILDES transcript system three additional tiers for morphological, syntactic and lexical coding.

### 2.1 The Morphological Tier

On the morphological tier (%mor), morphemes related to the functional heads C, I and D are coded as such and labeled with codes identifying the specific structures involved. See, for example, (1) and (2) below.

- (1) *she* is a **Subject PRONoun**, its nominative case-marking received via and so related to the functional head **I[nfl]**, and is thereby coded: **IPROS|she**.
- (2) *my*, the **POSsessive Determiner** related to the functional head **D[et]** is coded: **DPOSD|my**.

We also code for phonetically overt bound morphology on this tier. Stems and affixes are divided by “-”, as in (3).

- (3) *Comes* is coded: **IF|come-s**, where **IF** corresponds to a **Finite** form of the verb related to functional head **I**.

So the sentence *She comes* would be coded as in (4).

- (4) \*CHI: *she comes* .  
%mor: IPROS|she IF|come-s

Other free grammatical morphemes are also coded by syntactic category on this tier (e.g., **Preposition**, **PRONoun**, etc.), and utterance length is generated on the %mor tier as well. Thus, all and only the utterances and words and morphemes within them to be included in such counts are put on this tier. Errors related to all structures coded are identified by “=”. Omissions, misselections, and overinsertions of morphemes are all captured through this scheme as illustrated by (5) – (7):

- (5) Omissions receive the code “=0x” where x is obligatory but omitted.
- (6) Misselections receive the code “=y” if y is there instead of x.
- (7) Overinsertions are coded as “-x=-x” if there are two markings of x instead of one.

It is important to note that, although this tier is labeled the “morphological” tier, its label should not be taken literally. Rather, our morphological tier should be understood as a place for coding syntactic as well as morphological information as specified above, designed to meet the principle objectives of including but going beyond the coding of bound and free morphemes, to code a) the functional categories C, I, D and their subcategories, and b) errors related to these functional categories and their subtypes (i.e., omissions, misselections and overinsertion).

### 2.2 The Syntactic Tier

The syntactic tier (%syn) is designed as the place for coding constituent structure (including types of embedding and internal phrasal and clausal structure), constituent order (capturing linear order and movement), and related errors. Each syntactic phrase is labeled with category and grammatical function labels as illustrated by (8) - (10):

- (8) **SNP = Subject Noun Phrase** (category = NP, grammatical function = Subject)
- (9) **MOD = Modal**
- (10) **CADJ = ADJunct clause, complementizer in C** (projects up to C)

Embedded clauses are enclosed in parentheses ( ) and codes for embedding types in square brackets ( [ ] ). Verbal morphology is also included on this tier, using the same codes as used on the morphological tier. An example coding of a well-formed utterance is shown in (11).

- (11) \*CHI: *She can't sing if he's in the room.*  
%syn: SNP MOD NEG V ( [CADJ] SNP IS PP )

Errors (e.g., order errors, omission errors) are also coded, using “=” to signal errors as on the morphological tier. (12) illustrates the coding of an omission error.

- (12) \*CHI: \**He going*  
%syn: SNP IS|=0 V-ing

Our system for coding syntactic structure on this tier allows for analysis of the cooccurrence of grammatical structures (e.g., null subjects and finite verb marking).

### 2.3 The Lexical Tier

The lexical tier (%lex) should be understood as the tier for coding types, tokens, and errors of the major lexical category words (N's, V's, ADV's, and ADJ's) in each utterance of a speech sample. All of the morphemes which are part of the same lexical entry are placed, without spaces between them, in the same listing. An “=” is again used to signal an error. The lexical coding is illustrated in (13) below.

- (13) \*CHI: My father poured dinner for us.  
 %lex: N|father V|=pour N|dinner

This tier allows us to examine the structure, size and productivity of the speaker's lexicon. Omitted words would be captured on one of the other tiers defining the resultant grammatical error.

This coding system is still undergoing modification. We are forced to add to or modify our current system to accommodate data which deviate in unexpected ways from the normal, adult grammar. By designing our system to describe and analyze disordered as well as normal language, however, we have developed a system which is useful for linguistic analysis of all kinds of speakers, including normal children, children with language disorders, adults with acquired aphasia and, of course, normal mature speakers.

### 3. Analysis System

Freq.exe, one of the CLAN programs associated with the CHILDES Project<sup>1</sup>, provided the basis for an elaborated computer analysis system which we designed to answer specific research questions regarding various subsystems of the grammar. While freq.exe is a powerful tool for counting frequencies of words, morphemes or codes, research questions posed by our project required that a variety of very different morphological, syntactic, and lexical codes be counted as evidence regarding a single subsystem of the grammar. With respect to the I-system, for instance, our coding system marks a subject pronoun *she* as IPROS|she, and errors in IPROS as IPROS|=she. Counts of IPROS in the language samples could be obtained using the standard freq.exe command shown in (14); however, the error counts required two freq.exe commands, as shown in (15).

(14) freq +t%mor -t\* +s "\*"IPROS|\*" acoded.ext  
 >output1.ext

(15a) freq +t%mor -t\* +s "IPROS|=\*" acoded.ext  
 >output2.ext

(15b) freq +t%mor -t\* +s "=IPROS|\*" acoded.ext  
 >output3.ext

(15a) is an error with respect to the functional category itself, while (15b) is an error with respect to the element within that functional head.

For the I-system alone, occurrences of fifteen different codes and their errors needed to be counted individually and sorted for inspection. Because we code for a variety of errors in each category, two or three freq.exe commands would be necessary for each error subtotal, depending upon the range of errors coded in the data for each category. Occurrences of error obtained in freq.exe searches, such as those in (15), would have to be added together in order to account for the total number of errors in each category under analysis. In all,

forty-seven freq.exe searches would be required for each data point. Besides amounting to an arduous and time-consuming task, the conventional freq.exe method left us with the need to subtotal error counts for each category by hand, thus introducing an opportunity for miscalculations in our results.

Therefore, in order to simplify and streamline analyses for the I-system, we created a program called i-sys.exe which performs four basic operations, as outlined in (16).

- (16) Each time i-sys.exe is run, it
- (a) calls freq.exe forty-seven times to conduct relevant I-system searches on coded files, creating a separate freq.exe file containing search results for each iteration;
  - (b) reads the forty-seven files created in (a) and extracts the "total number of words" (codes, actually) in each; and
  - (c) computes results obtained in (b) to compile a final report called *i-report.ext*.
  - (d) calculates the number of morphemes, number of utterances, and mean length of utterance (MLU)

The forty-seven freq.exe output files are all compiled into a single file called freq-out.ext and deleted from the disk. The analysis procedure is completed in a run time of approximately 13 seconds under DOS, on a 80 MHz. Pentium computer.

The collection of reports generated allows easy inspection of various aspects of the I-system, including the proportion of errors in each. Comparable "reports" are generated for the D- and C-systems, as well as closed class items.

### 4. Critical Assessment and Conclusions

This methodology and these analysis tools free us from the alienating labor of manually counting tokens in our data sets and provide us with a rich and accurate basis for interpreting our data. However, although this system is appropriate for our research objectives and of great assistance to us in our work, as theoretical insights and the data demand, we will continue to modify and improve it.

#### Notes

<sup>1</sup> See Brian MacWhinney, *The CHILDES Project: Tools for Analyzing Talk* (New Jersey: Lawrence Erlbaum, 1991).