SIMULATING CHILDREN'S NULL SUBJECTS: AN EARLY LANGUAGE GENERATION MODEL

Carole T. Boster
Department of Linguistics, Box U-145
University of Connecticut
Storrs, CT 06269-1145, USA
tenny@uconnvm.uconn.edu

Abstract

This paper reports work in progress on a sentence generation model which attempts to emulate certain language output patterns of children between the ages of one and one-half and three years. In particular, the model addresses the issue of why missing or phonetically "null" subjects appear as often as they do in the speech of young English-speaking children. It will also be used to examine why other patterns of output appear in the speech of children learning languages such as Italian and Chinese. Initial findings are that an output generator successfully approximates the null-subject output patterns found in English-speaking children by using a 'processing overload' metric alone; however, reference to several parameters related to discourse orientation and agreement morphology is necessary in order to account for the differing patterns of null arguments appearing cross-linguistically. Based on these findings, it is argued that the 'null-subject phenomenon' is due to the combined effects of limited processing capacity and early, accurate parameter setting.

1 THE PROBLEM

It is well known among researchers in language acquisition that young children just beginning to speak English frequently omit subjects, in linguistic contexts where subjects are considered mandatory in the adult language. Other major structural components such as verbs and direct objects are also omitted occasionally; however, the frequency at which children omit mandatory object NPs tends to be much lower than the rate at which they omit subjects. For example, P. Bloom's (1990) analysis of early speech transcripts of Adam, Eve and Sarah (Brown, 1973) from the CHILDES database (MacWhinney and Snow, 1985), indicates that these children omitted subjects from obligatory contexts 55% of the time on average, whereas obligatory objects were dropped at rates averaging only 9%. But by around age 2 1/2, or when the mean length of utterance (MLU) exceeds approximately 2.0 morphemes, the percentage of null subjects drops off to a level about equal to the level of null objects.

The reason for the so-called null-subject phenomenon in early child English has been widely debated in the literature. Different theories, though they vary greatly in detail, generally fall into two broad categories: processing accounts and parameter-setting accounts. The general claim of those who favor a processing account is that the phenomenon (in English) is caused by severe limitations in the child's sentence-processing or memory capacity. It is known that young children's utterances are much shorter on average than adults', that their sentence length increases steadily with age, and that other components of a sentence are also routinely omitted, which could be evidence of processing limitations. Yet some who argue for a strictly grammatical explanation (including Hyams (1986), Hyams and Wexler (1993)) claim that the differential patterns of null subjects over null objects cannot be accounted for by any existing processing account, and instead take this as evidence that the 'unmarked' setting for the relevant parameter(s) related to null subjects is (+pro-drop); various accounts are offered for how children learning languages that do not permit null subjects ultimately make the switch to the correct parameter value.

Others, including Valian (1991) and Rizzi (1994) have noted differences in the frequency of early null subjects depending on their position in a sentence; they tend to be omitted in matrix but not embedded clauses, and in sentence-initial position but not after a moved wh-element. This observation has been used to argue for a different grammatical explanation of the null-subject stage. Both Lillo-Martin (1991) and Rizzi (1994), for example, argue that the initial value of the parameters is set to (- pro-drop); Lillo-Martin claims that the matrix subject is outside the domain where the pro-drop parameters are applied initially, while Rizzi claims that the matrix CP is considered optional at an early stage in acquisition. Further evidence which may support either this approach or a 'combined' processing and parameters account includes the higher percentages and different patterns of pro-drop and topic-drop found in the speech of children learning Italian, a pro-drop language (Valian, 1991) and Chinese, which allows 'topic-drop' (Wang et. al.,

1992), as compared to English-speaking children of the same age and MLU. Processing constraints should remain the same for children around the globe, so it is not clear that processing alone can account for the different distributions of nulls exhibited by 2-year olds learning English, Italian, and Chinese. However, the crosslinguistic differences also argue against the claim that all children start out with the relevant parameter(s) initially set to (+pro-drop).

2 THE MODEL

FELICITY, a sentence generation model that emulates early child language output, has been designed in order to determine whether the 'null-subject' phenomenon in early child language can best be accounted for by an incorrect initial setting of certain parameters, by processing limitations, or by an interaction between parameter setting and processing. FELICITY assumes a modular approach, following Garrett (1975), in which the intended message goes through three processing modules to yield three levels of output: semantic, then syntactic, then phonetic. The model incorporates several standard assumptions of Principles-and-Parameters theory including X' structure-building capacity (Chomsky, 1981), head-complement ordering parameters, and several parameters currently thought to be relevant to the null-subject phenomenon. Following the Continuity Hypothesis (Pinker, 1984), the model has the potential capacity for producing a full clausal structure from the beginning; the structure-building mechanism is presumed to be innate. It is also assumed, following the VP-internal Subject Hypothesis (Koopman and Sportiche (1988) and others) that the subject is initially generated within the VP. An algorithm controlling processing capacity, similar in principle to that proposed by Gibson (1991) to account for processing overload effects in adult sentence processing, will limit structure-building and dictate maximum 'holding' capacity before a sentence is output. The lexicon will initially include all words used productively in transcripts of an English-speaking child at age 1;7; lexical entries will include information about category, pronunciation, obligatory and optional complements, and selectional restrictions on those complements. All parameters will be binary. They can be assigned either value initially and can be reset; reference to any given parameter can also be switched on or off. The processing capacity of the model can also be adjusted, and the lexicon can be updated.

The model will be able to produce a sentence with a specific meaning or intent (as children presumably do), if it is given certain data about the intended proposition; this data will comprise a semantic representation containing a verb, its theta-grid (i.e. agent, experiencer, goal and/or theme), information about time frame or tense, person and number, mood, negation, and whether or not arguments have been identified previously in the discourse. When making direct comparisons of the model's performance with children's actual utterances, the data that is input to the model will be coded on the basis of inferences about what the child 'intended' to say based not only on actual transcribed output but also from the situation, prior discourse, and possibly caregiver's report (cf. L. Bloom (1970) on 'rich interpretation' of children's utterances).

Syntactic processing proceeds as follows: Begin structure-building at the level of the matrix CP, but via a recursive phrase-building process. Phrase-building begins by merging a complement phrase with its X^0 head (after the complement phrase has been built) to form an intermediate or X' level of structure. This unit is then combined with its specifier to form a 'maximal' phrase or XP. Lexical items are inserted as soon as the appropriate X^0 heads (or XPs, for pro-forms) become available. Each time a structural unit is built, and each time a lexical entry is inserted, the processing load is incremented; when the maximum load is exceeded, the model abandons processing and outputs the words currently in the buffer.

3 INITIAL APPLICATION

FELICITY's output will be compared to actual output from a longitudinal sample of several English-speaking children's early utterances, using transcripts available on the CHILDES database. The initial lexicon will be constructed based on the productive vocabulary of a given child from her first transcript. The 'processing limit' will be set at a given maximum, such that the model's MLU approximates that of the child in the transcript; the algorithm will be fine-tuned to determine how much relative weight or processing 'cost' should be assigned to (a) lexical lookup to get subcategorization information for the verb; (b) building of a structural unit; and (c) retrieval of phonological information. The sentence-generation procedures will be run under two conditions, once with parameter-checking enabled and then with parameter-checking disabled. Additional runs will try to emulate the child's output patterns during subsequent transcripts, after augmenting the model's lexicon with new words found in the child's vocabulary and adjusting the processing limit upward so that the output matches the child's new MLU. Statistical comparisons will be made between the model's and the children's performance (at

comparable MLU levels) including percentages of null subjects and null objects in the output, percentages of overt nominal subjects (full NPs) vs. overt pronominal subjects, percentages of other sentence components omitted, and amount of variability in utterance lengths.

4 PRELIMINARY FINDINGS

Initial trials indicate that, once the processing-complexity algorithm is tuned appropriately, FELICITY can approximate the null-subject output patterns found in English-speaking children with no reference to parameter values. Indeed, because the model builds complements before specifiers, it produces a much higher incidence of null subjects than null objects using a processing-overload metric alone. Furthermore, it yields a higher incidence of nulls in matrix sentences than in embedded clauses, and within a clause it only omits subjects in initial position, not after a moved wh-element or topic. However, it appears that the model will also need to reference parameter values if it is to account for the patterns observed in the speech of children learning languages which do allow null arguments; processing constraints alone will not explain the different crosslinguistic distributions of nulls.

5 FUTURE APPLICATIONS

Once FELICITY's processing metric is fine-tuned for English, it can be used to emulate argument omission patterns shown in other languages like Italian and Chinese, to test various parametric theories. If the relevant parameters involved are as given in Lillo-Martin (1991), for example, FELICITY should be able to emulate the relatively high level of null-subject usage by Italian-speaking children reported in Valian (1991) by simply switching certain subparameters related to Null Pronoun Licensing (NPL) and Null Pronoun Identification (NPI) to positive for an Italian child at age 2, while keeping processing constraints at the same levels that were established for English-speaking children. The model should also be able to emulate the higher percentages of null subjects and null objects found in the output of Chinese-speaking children in experiments reported in Wang et. al. (1992) by simply switching the Discourse Oriented (DO) parameter to positive, while leaving the NPL and NPI parameters set at the default (negative) values.

FELICITY can also be used to address theories pertaining to other aspects of language acquisition that appear slightly later in development, such as the appearance of subject-auxiliary inversion in yes/no and wh-questions, and the emergence of Tense and Agreement features. Future enhancements to the model are planned with these applications in mind.

ACKNOWLEDGMENTS

This material is based upon work supported under a National Science Foundation Graduate Research Fellowship. Thanks go to my committee members Diane Lillo-Martin, Stephen Crain, Ted Gibson and Howard Lasnik, and to two anonymous reviewers for helpful comments on an earlier draft.

REFERENCES

Bloom, L. (1970). Language development: Form and function in emerging grammars. Cambridge, Mass.: MIT Press,

Bloom, P. (1990). Subjectless sentences in child language. Linguistic Inquiry, 21, 491-504.

Brown, R. (1973). A first language: The early stages. Cambridge, Mass.: Harvard University Press.

Chomsky, N. (1981). Lectures on government and binding. Dordrecht: Foris.

Garrett, M. F. (1975). The analysis of sentence production. In G. Bower (Ed.), Psychology of learning and motivation (Vol. 9). New York: Academic Press.

Gibson, E. A. F. (1991). A computational theory of human linguistic processing: Memory limitations and processing breakdown [Doctoral dissertation]. Pittsburgh: Carnegie Mellon University.

Hyams, N. M. (1986). Language acquisition and the theory of parameters. Dordrecht: D. Reidel Publishing Company.

Hyams, N., & Wexler, K. (1993). On the grammatical basis of null subjects in child language. Linguistic Inquiry, 24, 421-459.

Koopman, H., & Sportiche, D. (1988). Subjects [Ms.]. Los Angeles: UCLA.

Lillo-Martin, D. C. (1991). Universal Grammar and American Sign Language: Setting the Null Argument Parameters. Dordrecht: Kluwer Academic Publishers.

MacWhinney, B., & Snow, C. (1985). The Child Language Data Exchange System. Journal of Child Language, 12, 271-296.

Pinker, S. (1984). Language learnability and language development. Cambridge, Mass.: Harvard University Press.

Rizzi, L. (1994). Early null subjects and root null subjects. In T. Hoekstra & B. D. Schwartz (Eds.), Language acquisition studies in generative grammar (pp. 151-176). Amsterdam/Philadelphia: John Benjamins.

Valian, V. (1991). Syntactic subjects in the early speech of American and Italian children. Cognition, 40, 21-81.

Wang, Q., Lillo-Martin, D., Best, C. T., & Levitt, A. (1992). Null subject versus null object: Some evidence from the acquisition of Chinese and English. Language Acquisition, 2, 221-254.