REVIEW ESSAY: A CRITIQUE OF FLYNN'S PARAMETER-SETTING MODEL OF SECOND LANGUAGE ACQUISITION

ROBERT BLEY-VROMAN CRAIG CHAUDRON
University of Hawai'i at Manoa

This paper is a critique of the parameter-setting model of second language acquisition proposed by Suzanne Flynn. Flynn has applied the notion of a universal head direction parameter to the prediction of acquisition of English adverbial clause structure and anaphoric relations by L1 Spanish, Japanese, and Chinese learners. In two studies involving elicited imitation and comprehension of target language sentences, Flynn argues that the head direction parameter explains the pattern of results. One proficiency group of Spanish learners revealed a greater ease in production of postposed (forward) pronoun anaphora over preposed (backward) anaphora, while Japanese and Chinese learners showed no differences. Flynn claims that the match in head direction between Spanish and English favors this outcome and related results, while Japanese and Chinese learners have difficulty because of the mismatch. This critique raises serious questions as to the adequacy of Flynn's methodology and model to investigate the issue or explain the results.

1. INTRODUCTION

How are young children able to acquire—with essentially uniform success—systems of such apparent abstract complexity as human languages? And how is this ability related to second language acquisition? In recent years, much work in language acquisition theory from a linguistic perspective has been devoted to an exploration of a *parameter-setting model of language acquisition*. (See Roeper and Williams 1987 for representative work.) The concepts of this promising approach to child native language development have also been adopted in recent work in second language acquisition (Liceras 1985, White 1985a, 1985b, Hilles 1986).

In a series of publications, Suzanne Flynn (1987a, and see further references below) has reported an investigation into the importance of what she calls the "head-direction" parameter in second language acquisition. This research is significant for at least three reasons. First, it appears to bear on the question of the adequacy of a parameter-setting model for second language acquisition, and thus indirectly on whether the same psychological mechanisms underlie both first and second language acquisition (Clahsen and Muysken 1986, Long

University of Hawai'i Working Papers in ESL, Vol. 7, Number 1, March 1988, pp. 67-107.

1987, Bley-Vroman in press). Second, Flynn's work attempts an empirical test of this model in a substantial study involving contrasting first languages. Third, her study investigates the structural constraints common to both second language production and comprehension. It is therefore important to explore the validity and implications of Flynn's research.

In the following paper we will review Flynn's study in depth, in order to determine whether her version of a parameter-setting model, or some other explanation, best accounts for her data. Both the conceptual framework for her study and various methodological aspects are complex, requiring detailed analysis. Our principal conclusions will be that Flynn's study is theoretically not well founded, it is flawed in numerous methodological ways, and the clearest results are subject to more plausible interpretations than those offered by Flynn.

The paper is divided into two main sections. The first deals with the theoretical rationale for Flynn's experiment. We explore the head direction parameter in linguistic theory and its relationship to adverbial clause position and anaphora direction, which are the focus of Flynn's research. We also outline Flynn's hypotheses and their basis in language acquisition research. The second main section analyzes the experimental methods and results. We consider separately results of production and comprehension tests; within each of these subsections we explore both the relationship between the hypotheses and the particular experiments, and the adequacy of Flynn's explanations of the results.

2. THEORETICAL FOUNDATION

Basic to the parameter-setting model is the observation that the properties of languages "cluster" on certain typological parameters. A language of one type will have one cluster of properties; a language of a different type will have another. This clustering in effect eases the burden on the learner of a language. The learner need only observe that the language is of a certain typological character, in order to deduce the existence of the various properties that cluster in that type of language. As Flynn (1987a:87) puts it, the property clusters are "sets of deductive consequences resulting from setting parameters for a certain value". (Chomsky 1985, especially chapter 3, provides a general discussion of the major concepts and their justification.)

2.1. THE HEAD DIRECTION PARAMETER

One way in which languages differ is the order of the head of a construction and its complements or modifiers. English is said to be a predominantly head-initial language because the verb (the head of a verb phrase) precedes its object (its complement); the head noun in a noun phrase precedes its relative clause (a modifier); and a preposition precedes its object (its complement). Japanese, on the other hand, is said to be head-final because the verb follows its object, the head noun in a noun phrase follows its modifiers, and a preposition follows its object.

Flynn proposes that the concept of head direction should be generalized to include the position of adverbial adjuncts, and that the preferred direction for pronoun interpretation is related to a language's head direction. As her arguments are rather dispersed, we must draw from several discussions in Flynn (1984, 1987a:Chapter 3 and elsewhere), in order to summarize the principal relationships that she proposes should hold between these three properties.

- A language which is principally head-initial will
  a. Prefer adjunct adverbial clauses in sentence-final position
  b. Prefer pronouns to follow their antecedents
- A language which is principally head-final will
  a. Prefer adjunct adverbial clauses in sentence-initial position
  b. Prefer pronouns to precede their antecedents

If these properties are indeed linked, then there ought to be interesting consequences for second language acquisition. For the learner, working out the position of adverbial clauses and the interpretation of anaphora should depend on identifying the head direction of the language being acquired. There may also be effects depending on whether L1 and L2 head direction match. Flynn proposes to investigate the relationship of these linked properties in second language acquisition, and especially, the effect of L1/L2 differences.

Before we consider Flynn's hypotheses about L2 acquisition, we will discuss the concept of head direction itself, then the presumed link to adverbial clause position, and finally the relationship to pronominalization direction. We will show that Flynn's linking of these in a head direction parameter finds virtually no support in current linguistic theory.

2.1.1 The concept of principal head direction.

It is often difficult to say that a language has a particular head direction. Japanese is very uniform, and it seems right to say that it is head-final. However, things are seldom so simple. English is only "predominantly" head-initial: adjectives precede their head nouns. And in Chinese, NP objects follow the verb (head-initial), and the object of a PP follows the preposition; but the complement to an N in an NP precedes. Chinese has prepositions (a head-initial characteristic), but relative clauses usually precede their head nouns (a head-final characteristic). Cases like Chinese show that the concept of unitary head direction for a language is misconceived, and that "languages may parameterize on both the type and the level of category" (Huang 1982:40)[1]. Thus, the language learner cannot assume that there is a single value head direction, from which the configurations of all types and levels phrases can be deduced. It is ironic that Flynn should have chosen Chinese as one of the languages that demonstrate an L1/L2 differential in head-direction (Flynn and Espinal 1985).

Although particular languages do tend to have predominant head directions, there is no consensus in linguistic theory as to whether or in what way these tendencies constitute a single parameter. Flynn does not espouse any theory of head direction. Flynn and Espinal (1985) merely comment in a footnote, "The exact nature of this parameter is at issue both theoretically and empirically" (p. 110). This failure to provide any explicit theoretical grounding deprives Flynn's work of much of its interest.

2.1.2 Head direction and adverbial clause placement.

Flynn suggests that in head-initial languages like English, adverbial clauses tend to follow main clauses. In head-final languages, on the other hand, adverbial clauses are suggested to come first. Thus *Mary laughed when John*

[1] Huang says specifically of Chinese that it uses the head-initial rule for the lowest level of expansion, but requires the head-final rule for all higher levels and that, furthermore, noun phrases never involve the head-initial rule at any level. Initially, Flynn herself pointed out the problems in deciding head direction. "The precise definition of a PBD [Principal Branching Direction—later replaced by Head Direction] is an empirical issue both theoretically and empirically. In this research, languages are chosen for experimental purposes which are basically consistent in BD (e.g., English, Spanish, and Japanese)." (Flynn 1984:86) Later she abandoned this rationale and decided to study Chinese, despite its inconsistent head direction.

spilt the milk is thought to be more usual than *When John spilt the milk, Mary laughed.* We know of no version of current linguistic theory from which such a connection between head-complement direction and adverbial clause placement could be derived. Whether such a connection even exists is very unclear. After all, in English, adverbial clauses are highly mobile, depending on pragmatic and discourse factors. Complements, on the other hand, are much more rigidly fixed in position: English objects in the verb phrase never precede verbs and relative clauses never precede nouns, no matter what the discourse context. In Japanese, the order of adjuncts is also largely affected by pragmatics and discourse, although a general requirement that the main verb be final in the sentence limits the possibilities.

Flynn's efforts to link head direction and adjunct placement results in her generalizing the terminology of "head" and "complement" to include the case of main clause and adjunct adverbial. This terminological innovation obscures the well-known differences between adjunct adverbials and the complements (arguments) of predicates. The traditional distinction between complements and adjuncts is central to all current linguistic theories. For example, in generative grammar, verbs are subcategorized for arguments, but not for adjuncts. In X-bar theory, the distinction relates to level of attachment.[2] Furthermore, the distinction between subcategorized and non-subcategorized elements is essential to the Projection Principle, which to a great extent determines the architecture of the grammar in Government-Binding theory—the theory of syntax to which Flynn apparently subscribes. If head-complement order is a consequence of direction of case and/or theta-role assignment (Travis 1984; Koopman 1985—sources cited by Flynn), then order of main clause and adjunct adverbial cannot be unified with the order of head and complement, since main clauses surely neither assign case nor theta-role to the adjunct adverbial.[3] Some reasons to reject the identification of adjuncts with complements (reasons involving government relations, extraction possibilities, and mobility) are in fact acknowledged by Flynn and Espinal in a footnote (1985:111, footnote 5) but dismissed without argument.[4]

---

[2] Thus Huang's observation, cited above, that a given language can have different head directions depending on bar-level, amounts to an assertion that there is no necessary deductive link between the position of adjuncts and the position of complements, as Chinese itself so clearly shows.

[3] We are indebted to Lynn Eubank for this observation.

[4] In another footnote, Flynn and Espinal admit that the concept of the head direction

2.1.3 Head direction and anaphora direction.

Children acquiring a head-initial language have been observed to prefer pronouns occurring later in the sentence than the antecedent (*forward preference*), while the reverse has sometimes been claimed for child learners of head-final languages. For example, child speakers of head-initial English find it easier to deal with forward pronominalization (*After John spilled the milk, he wiped it up*) than backward pronominalization (*After he spilled the milk, John wiped it up*). Lust (1981) calls the connection between head direction and pronominalization preference the "Directionality Constraint" or "Directionality Principle".[5] Both in logic and experimental design, Flynn closely follows the work of Lust (1981, Lust and Wakayama 1979; Lust 1986:44-54 provides an excellent summary), who supervised the dissertation on which Flynn's published reports are largely based.

What evidence is there for the Directionality Principle? Many scholars have shown that there is a preference for forward pronominalization in English speaking children, especially in structures which in principle allow both directions (Lust 1986:44-52). However, a preference for backward pronominalization in head-final languages is not well-established, having been suggested only in the research of Lust and her colleagues. The studies of Lust and others on which this assertion is based (Lust and Wakayama 1979, and Lust and Mangione 1983 on Japanese, Lust and Chien 1984 on Chinese) have recently come into question. These studies are largely based not on pronominal anaphora per se but on other structures, in particular coordinate conjunction, which is obviously not the same phenomenon (Reinhart 1986). Lust and Chien[6] (1984) analyze their L1 Chinese coordination data in a way

parameter applies "by implication rather than argument to adverbial subordinate clauses." (1985:111) We are not certain what is meant by "implication". Obviously not logical implication—which would be the only relevant sense. As far as we know, no linguist has ever "implied" a unity of complement and adjunct, even without argument.

[5] In Lust's original formulation, the Directionality Principle related pronoun interpretation not to head direction but to a slightly different concept—Principal Branching Direction. Flynn (1984) in fact uses the term Principal Branching Direction rather than head direction. In more recent work, Flynn's terminology has changed. While the two concepts would seem to be different, Flynn provides no argument in favor of one over the other and indeed makes no attempt to distinguish them.

[6] Due to lack of space, we will not elaborate on the flaws in Lust and Chien's analysis. Suffice it to say that they average their results for non-anaphoric forms with anaphoric ones to obtain a score for *Directionality* in coordination. This method tends to obscure any potential directionality preference.

that would obscure a forward directionality preference, and evidence for the claimed backward preference is weak. O'Grady, Suzuki-Wei, and Sook Whan Cho (1986) have performed additional studies with Japanese and Korean subjects, focussing on anaphora rather than coordination, and using structures closely analogous to those used by Flynn in her study of L2 English. They have shown that the Japanese have a FORWARD preference in these cases, and speculate that this is the same for all languages.

In addition to the failure of empirical studies of L1 acquisition to establish a link between head direction and pronominalization, the Directionality Principle finds little theoretical support. In no formulation of the Binding Principles (those principles of syntax which deal with pronominals and anaphora) is there any function for head direction, nor for adverbial adjunct placement. This is true especially of the recent theory of binding associated with Chomsky (1981, 1982), which is completely non-directional. It is also true of theories which propose different conditions on forward and backward pronominalization (for example the "precede or command" theories based on Langacker 1969). Reinhart, the principal architect of the currently most widely accepted theory of binding (that based on c-command) has pointed out that the Directionality Principle is "not related at all" to the Binding Principles. "So this picture of the grammar could not yield the directionality effect of English. If the PBD [Principal Branching Direction] principle is correct, this means that the child makes no use at all, at the relevant stage, of the BP [Binding Principles], but rather operates by an altogether independent parameter" (Reinhart 1986:141). Flynn nowhere addresses the evident contradictions between the Binding Principles and the Directionality Principle.[7]

[7] Lust has responded to Reinhart's observations as follows: "There is nothing intrinsically contradictory between the PBD [Principal Branching Direction: i.e. the Directionality Principle] and the BP [Binding Principles]. In fact, since both are defined with regard to 'command' relations they may be intrinsically related. While one (PBD) describes a parameter of grammatical organization, the other (Binding Principles) describes a set of principles which apply within a specific language grammar. The domain of the first is grammar. The domain of the second is specific sentences." (Lust 1986:70) It is difficult to know what to make of this response. We see that the BP and PBD both involve "command" relationships, in the loose sense that they both have reference to structure. But hardly anything in language does not. We do not understand the claim that the principles apply in different "domains", or why their applying in different domains should constitute an argument that they are "intrinsically related".

2.2 HYPOTHESES RELATING TO SECOND LANGUAGE ACQUISITION

Even though Flynn does not provide any clear theoretical basis for her study of parameters in second language acquisition, it must be said that at our current stage of understanding no proposal for a parameter is uncontroversial. Even parameters supported by linguistic theory have many debatable characteristics. The fact that Flynn's parameter is without theoretical underpinnings thus does not rob her study of all interest.

What hypotheses does Flynn derive about second language acquisition from her concept of a head direction parameter? The basic question is the role of the first language—specifically the potential effects of a match or mismatch of L1 and L2 head direction on the acquisition of adverbial clause placement and anaphora. Within a parameter-setting model, there are two general possibilities with respect to second language acquisition. First, it is possible that an adult learner would proceed exactly as a child and be guided simply by the head direction of the language being learned, showing the same parametric clustering of head direction with adverbial clause placement and pronominalization direction as the child. Second, it is possible that the adult learner's initial stages would be significantly affected by head direction of the native language. Flynn takes this second possibility as most promising. It yields the prediction that there will be a significant difference in the acquisition of the relevant structures depending on head direction of the native language. Learners of a L2 with head direction matching L1 have an advantage: they can rely on the L1 parameter setting in the L2. If head direction of L1 and L2 do not match, learners will have to set a new value for the parameter; this will make acquisition of the structures linked by the parameter more difficult.[8]

The following are two formulations of Flynn's hypotheses for her study:

> [I]f acquisition of a second language involves this essential language faculty [UG], then second language acquisition should involve this principle of PBD [Principal Branching Direction—since retermed head direction] in some way .
>
> . . . [W]e would predict that the first and second language PBD

[8] One might presume that the idea of parametric differences causing difficulty is simply contrastive analysis in new clothing. In a sense this is correct: work in this vein might reasonably be termed "UG-based CA". However, Flynn (1987a:85-87) correctly emphasizes certain aspects of the UG approach which distinguish it at least from habit-formation views of contrastive analysis.

> mismatch would affect the acquisition of anaphora in particular. Specifically, we would predict more anaphora errors where the first language does not match the second language in PBD given the principle of first language acquisition. Where the PBD of the second language matches that of the first language, one might expect the pattern of acquisition of anaphora in the second language acquisition process to be similar to that in the first, but one would not necessarily predict this where there was a PBD mismatch. (Flynn 1984:78)

> Evidence should be found clearly indicating that L2 learners consult the configuration determined by this parameter [HD "head direction"] in organizing other aspects of complex sentence formation in the L2, namely, sentence embedding [in this case, adverbial clause placement] and anaphora.
>
> . . .When L1HD ≠ L2HD, L2 learners must assign a new value to the parameter, and we would expect acquisition patterns to correspond to early L1 acquisition patterns for this parameter.
>
> . . .When L1HD = L2HD, L2 learners need not assign a new value to the HD parameter. These L2 learners can rely upon the L2 value to guide their hypotheses about other aspects of the L2 grammar, such as anaphora. Acquisition patterns of these complex sentence structures should correspond to later stages in the L1 acquisition of these structures. . . . L2 acquisition of complex sentence structures and anaphora should be significantly easier than when L1HD≠L2HD. (Flynn 1987:84)

These statements are at a very high level of generality. Flynn (1987a) proposes more specific hypotheses relative to several of the test sentence sets, which we will discuss in sections describing those tests.

There seem to be two major points in these hypotheses. They are that a mismatch in L1/L2 head direction should lead to (a) greater DIFFICULTY in acquisition of the L2 forms, and (b) acquisition of the L2 forms in a PATTERN similar to that shown by L1 learners. Some clarification of these points is needed.

Regarding difficulty, Flynn has hypothesized that acquisition of the structures in question would be "significantly easier" with a match and that there would be "more anaphora errors" with a mismatch. Flynn's experiment

tests this hypothesis by means of evaluation of degree of target-like performance and by means of analysis of error types.

The second point is more difficult to discern as a clear hypothesis. We note first that Flynn's 1987 formulation is different from the 1984 one, although the source of both of them is the same study. The 1984 hypothesis predicts that in case of an L1/L2 match, the pattern of acquisition will correspond to the child L1 pattern, with an ambiguous prediction ("not necessarily predict") in the case of a mismatch. The 1987 hypothesis proposes that in BOTH conditions there will be similarity to child L1 acquisition patterns, but that there will be a difference in TIMING of the appearance of the patterns. However, Flynn nowhere addresses the question of what constitutes "early" and "later stages" of L1 acquisition. Flynn's eventual proposal of a more specific hypothesis in one experimental test case does not clearly derive from her minimal exposition of L1 "patterns" of acquisition. Therefore, we will see that the evidence accumulated in her experiment is very difficult to relate to this hypothesis

3. THE EXPERIMENT

The experimental approach is, in broad outline, to study the acquisition of the parametrically related structures by three different groups of learners. One group were native speakers of Spanish—a head-initial language. The other two groups had supposedly head-final native languages: Japanese and Chinese. The data which Flynn reports derive from two studies: Flynn's (1983) dissertation research on L2 production and comprehension by Japanese and Spanish subjects, and a replication of it with a set of Chinese subjects. Flynn (1987a) presents the most complete set of results for Japanese and Spanish production and comprehension, and Flynn and Espinal (1985) presents the Chinese production results, comparing these with selected results for the Japanese (which had not been published at that time). Other reports focus only on certain details and discussion, with Flynn (1984) reporting selected results on the Japanese and Spanish production tasks, and Flynn (1986) selecting results only from the Spanish production and comprehension tasks.[9]

[9] See the bibliography of Flynn (1987b) for a list of the published work derived from these two studies. Our review concentrates on Flynn (1984, 1986, 1987a) and on Flynn and Espinal (1985), although we make occasional references to other articles.

In this section we will outline the method and design of the study, discuss several procedural complications, and follow with analyses of results for the different production and then comprehension tasks. We will see that for a number of Flynn's results there are alternative interpretations, based not on her subjects' imposition of specific grammatical constraints on the stimulus input, but on more general processing constraints. Furthermore, we will point out how the results, for several sets of experimental sentences, suggest both ceiling and floor effects in production by different groups. Such effects seriously limit the sort of conclusions one can draw from her elicitation measures.

3.1 METHOD AND DESIGN OF THE STUDY

Three groups of learners of English as a second language (L1 Spanish, Japanese, and Chinese) were tested. In addition to L1 (*Language*), the other independent factors were learner proficiency (*Level*—Low, Mid, High—also referred to as beginner, intermediate, advanced—1984, 1986, Flynn and Espinal 1985) and two critical within-subject experimental variables: 1) the presence and type of anaphora in a subordinate clause (*Anaphora*, with three values—no anaphora (i.e. full noun), pronoun, and null anaphora), and 2) whether the subordinate clause preceded or followed the main clause (*Directionality*). The dependent variables were several elicited imitation tasks as measures of production, and a sentence act-out task (with geometric figures) as a measure of comprehension. *Task* as a factor did not enter into statistical tests of the outcomes except in one case.

Flynn (1987a) describes the experimental procedures and materials in some detail. She went to considerable lengths to avoid typical biasing factors, by for example, ascertaining that each subject had complete familiarity with the vocabulary used in the different tests. Also, the three proficiency levels were determined for all subjects in the three language groups in the same way—by total scores on the University of Michigan Placement Test for written grammar knowledge and listening comprehension.

3.1.1 Procedural complications

Four potentially serious limitations to confidence in the findings are evident from the description of the test administration procedures. We are not certain whether the first two limitations in fact resulted in any biasing of

the results. The second two, however, are more critical weaknesses in the design of the study, so we will be referring to them at various points throughout the remainder of the discussion.

First, the production and comprehension tasks were administered INDIVIDUALLY to each of the 104 subjects in Flynn (1983) and 60 Chinese subjects in Flynn and Espinal (1985). Moreover, part of the procedure allowed for a further repetition of the stimuli in the event that a subject responded minimally (1987a:112-3). In view of the rather complex nature of the test batteries (27 production items involving sentences of four different types, and 16 comprehension sentences involving two test conditions), such an interactive, individualized presentation has the potential of introducing uncontrolled variability and unexplained error into the elicitation, in the form of changes in rate and prosody in presentation of the stimuli sentences. Recorded stimuli are preferable.

A second limitation concerns the evidently different types of individuals included in the Japanese and Spanish study. The Japanese subjects' background experience with English was somewhat superior to that of the Spanish subjects, and they were quite a bit older (mean 30 years compared to mean 24 years; Flynn 1987a:106-107). Furthermore, a majority (55%) of the 53 Japanese subjects were female homemakers, compared to only 12% female homemakers among the 51 Spanish subjects (Flynn 1987a, Appendix A:197-200). A large number of these Japanese homemakers came from adult education courses offered by the board of education of a New Jersey city, and they appear to be considerably older than the rest of their group or the Spanish group. The majority of the Spanish subjects were students, and the rest of both groups were professionals or other employees studying in intensive language programs. This variability in age and experience factors may have resulted in unanticipated tendencies in the data.

A third, more serious, procedural issue concerns the grouping of subjects into proficiency levels. Flynn uses TOTAL grammar plus listening scores. This measure is not appropriate as a grouping factor for spoken language production, much less as an equation of learners' speaking ability ACROSS LANGUAGE GROUPS. In an attempt to control specifically for differences in productive ability, Flynn employed a set of sentences in the test battery in which the experimental variables were not manipulated. We will see that this control was not only inappropriately applied, but it was later ignored in

several comparisons of results for learner *Level*.

A fourth procedural question concerns the validity of using elicited imitation test sentences as a measure of SPECIFIC syntactic awareness, such as anaphora, or subordinate clause constructions. The following quote is representative of the limited extent of Flynn's argument in support of this application (see also Flynn and Espinal 1985:101, Flynn 1987a:88-89):

> The basic, well-documented assumption underlying the experimental use of this test is that the active repetition of the stimulus sentence reflects input of the sentence to both the comprehension and the production systems of the subject, and that the grammatical structure of the stimulus sentence is relevant to this processing. (1986:137)

Although we are highly sympathetic to efforts to investigate elicited imitation and its relation to L2 knowledge (cf. Chaudron 1986), the current status of the measure is not as "well-documented" as Flynn suggests. While Flynn (1987a:89) cites her 1986 study as a "review" of this issue, where we find neither review nor adequate argument, she also cites Gallimore and Tharp's (1981) study of second dialect subjects as justification for her claim. However, Gallimore and Tharp's review of some L1 and L2 studies points out the considerable disagreement among researchers and findings as to the PRECISION of elicited imitation for specific grammar points. Gallimore and Tharp conclude on the basis of their own work that elicited imitation in a standardized test format can result in reliable and valid judgment of students' general grammatical abilities relative to one another. They avoid concluding that PARTICULAR points of grammar can be measured with any precision. The standardized battery that they propose would include a variety of grammatical structures.[10] In contrast, Flynn's study rests on the assumption of specific grammar point assessment by means of this one production measure and related error analyses.

3.2 TESTS OF THE HYPOTHESES IN PRODUCTION

Since Flynn hypothesized that L2 acquisition would be easier when L1/L2 match in head direction, she is interested in demonstrating that Spanish

[10] We are faced with the well-known dichotomy between a proficiency test and a diagnostic test. In order to justify a diagnostic capacity for elicited imitation, many more studies of outcomes correlated and compared with other measures of the same grammatical structures would be necessary (see Chaudron 1983 for discussion of this point).

learners are superior to the Japanese and Chinese learners. She also attempts to find different "patterns" of response to the tasks. Precisely what sort of evidence would constitute superiority or patterns was not, as we have seen, fully spelled out prior to the collection and analysis of the data. In the following discussion of Flynn's studies, we will first deal with the production data, and then with the comprehension results.

The elicited imitation test battery for production (27 sentences) included: 1) (Test 1) six sentences with postposed (head-initial) *when*-adverbial clauses, and six with preposed (head-final) *when*-adverbials.[11] Crossed with this factor, three of each six had pronoun anaphora in the subordinate clause, and three had null anaphora, as in the following sentences—*When he entered the office, the janitor questioned the man* (preposed, pronoun), *The professor answered the owner, when Ø preparing the lunch* (postposed, null); 2) (Test 2) three sentences with postposed *when* clauses, and three with preposed when clauses, all with full nouns (e.g. *When the man dropped the television, the woman hugged the child*—preposed); 3) (Test 3) three sentences with preposed *when* clauses but with pronouns in the final main clauses; and 4) six "juxtaposed" sentences, involving two main clauses and two types of *Redundancy,* three with subject repeated (*The man discussed the article; the man studied the notebook*), and three with verb and object repeated (*The mayor dropped the letter; the diplomat dropped the letter*).

The first principal result Flynn (1987a) brings to bear on the hypothesized superiority in the case of an L1/L2 match is the overall poorer performance of the Japanese subjects on production of these test sentences relative to the Spanish subjects.[12] Flynn sees the need to argue that the difference between Spanish and Japanese learners is not merely attributable to INTRINSIC

[11] Given the doubtful theoretical status of Flynn's attempt to unify the concepts of head direction and adverbial clause placement, we are disturbed by the use of "head-initial" and "head-final" in Flynn's experimental reports to refer to English sentences with adverbial clauses at the end and at the beginning of sentences, respectively. It makes it appear that these two options differ from each other in the same theoretical sense that head-initial and head-final language types differ. Henceforth, we will only refer to these sentences as postposed and preposed sentences, respectively.

[12] For this and further analyses, she performs separate analyses of variance for each subset of test sentences, despite the potential to compute multivariate analysis of variance on the complete test battery. The complexity of the number of independent factors and the unbalanced nature of the design weigh against such a single analysis, although the duplication of analyses on each set increases the likelihood of a Type I error, especially given that Flynn adopts a criterion level of significance of .05 throughout.

differences between the groups, but rather to their different abilities to PROCESS THE ADVERBIAL CLAUSES AND ANAPHORIC RELATIONS, since the placement test did not suffice to ensure equivalence in spoken proficiency:[13]

> To test whether Ss' performance in production of the experimental sentences was specifically due to the factors of head-direction and anaphora direction, it was necessary to assess possible general sentence processing differences between the two language groups. For this measure, an assessment of amount correct on imitation of simple juxtaposed sentences was used and was treated as a covariate in the statistical design. This covariate STATISTICALLY REMOVED DIFFERENCES DUE TO BASIC TWO-CLAUSE SENTENCE PROCESSING FROM THE TESTS OF EFFECTS OF THE MANIPULATED FACTORS . . .If differences in head-direction or anaphora direction across the factors are found, even after the statistical removal of possible processing differences in the basic set of sentences, then a significant amount of group differences in abilities must be accounted for by head-direction and anaphora direction. [Flynn 1987a:90; emphasis ours]

Flynn cites two statistics and design texts to explain why analysis of covariance (ANCOVA) is appropriate to adjust for initial differences in groups in an experiment.[14] While this rationale is plausible, and common in

[13] Nonetheless, the following quote from Flynn (1987a) suggests that she thinks these tests do establish comparability:

> Comparable ESL proficiency levels and a measure of baseline syntactic competence were established for the two groups of Ss. This ensured that any differences in acquisition between Spanish and Japanese Ss were due to principled structure-based differences between the two groups -- such as the match or mismatch of head-direction of the L1 and the L2 -- and not to spurious factors. These controls were also used to establish comparability between the two language groups. [p. 88, underlining in original]

We are not certain what is meant by the "also" in the final sentence—anything different from the first sentence?—and we will suggest in the following that in fact, the highly "spurious" cause of differences between groups on the production tests is the fundamental difference between the groups in speaking proficiency.

[14] Flynn also suggests (1987a:91) that the covariate is necessary to adjust for differences between and within the proficiency placement levels in her design, and to more precisely control "for any differences that might exist between the two groups in terms of, for example, listening comprehension abilities," because the tests were administered orally. This is a rather peculiar additional justification, since the placement tests supposedly were meant to differentiate precisely among proficiency levels, and to group as equivalent within levels, especially on listening proficiency. Other than as an adjustment for general speaking

educational research with intact groups, statisticians caution against abuse of such an application of ANCOVA. Moreover, the evidence from Flynn's production tests entirely disqualify the use of covariance as a potential error-reducing manipulation. Let us look at this important shortcoming in her analysis.

3.2.1 Test of sentences used as baseline imitation

Flynn reports analysis of variance results (1987a:120-124) for the juxtaposed (covariate) sentences in a 2 X 3 X 2 design (*Language* X *Level* X *Redundancy* type; see examples above).[15] Not surprisingly for anyone familiar with Hispanic and Japanese students, the Japanese subjects, although equated on the grammar and listening tests, were significantly poorer ($p < .0001$) in elicited imitation production on these juxtaposed sentences. The Spanish group performed on the whole almost TWICE as well as the Japanese group. Although there was a further significant difference among levels (the High group being significantly different from the Low), there was no *Level* by *Language* interaction, indicating that the Japanese were CONSISTENTLY significantly poorer in imitation ability.

We draw attention to this finding (which, it will be seen, resembles many language group comparison findings on the other test sentences throughout the study), because this significant difference in imitation ability between the two groups is in clear violation of a critical assumption of the use of analysis of covariance. The statistics texts that Flynn cites, and virtually any other on ANCOVA, make clear in various ways that a covariate may not be used to adjust for differences in a dependent variable, when the covariate interacts with a treatment variable or is in any way correlated with the independent factors (in this case, has a significant *phi* correlation with, the *Language* group division).[16] For example, Keppel (1973:479) states, "The analysis of

proficiency, one would not want to alter or eliminate intentional design-induced differences and equivalences.

[15] It would have been useful had Flynn (1987a) provided ANOVA or ANCOVA tables for her many tests; some of these are provided in Flynn (1986, 1987b); the information therein is often quite useful, especially given that she only reports selected significant findings in three- and four-factor analyses, which have many important testable main effects and interactions.

[16] Numerous treatments besides those listed here attest to this assumption and the consequences of its violation (Kirk 1968:455- 458, Wildt and Ahtola 1978:16-17, Huitema 1980:98-122), although many commentators, in assuming randomized assignment to experimental treatment groups, do not comment directly on the sort of obvious violation evidenced in Flynn's study.

covariance consists of a statistical adjustment...for CHANCE differences for the treatment groups." (The emphasis is ours—i.e. not SIGNIFICANT differences.) Or Pedhazur (1982) comments on "extrapolation errors" in the use of ANCOVA:

> When there are considerable differences on the covariate between, for example, two groups so that there is little, or no, overlap between their distributions [i.e. they are significantly different], the process of arriving at adjusted means involves two extrapolations....[I]t would be more appropriate to speak of 'fictitious means' (p. 522) (bracketed comments ours)

An excellent illustration of the bizarre consequences of this type of extrapolation error is that Flynn's ANCOVA derives "adjusted" means of .37 and .33, where the actual means and variances were ZERO, for the Low Japanese subjects on two test measures (1987a Table VII.4, p. 127 for pronominal and null anaphora sentences, and Table VII.5, p. 135 for no anaphora sentences).[17] Because Flynn's ANCOVA results are questionable,

[17] It must be pointed out further that Flynn has not provided the important information that is needed in order to justify use of ANCOVA anyway. Several other assumptions or limitations of its use are (see relevant references above): 1) that the covariate(s) and dependent measure(s) are correlated (we actually assume this to be likely—in fact, that the subjects' relative performance on elicited imitation of the various sentences will be quite similar regardless of the sentences involved—but Flynn does not report the correlation); 2) that the variables have normal distributions; 3) that there is homogeneity of variance of the dependent variable measures across groups; (2) and 3) are also assumptions of analysis of variance); and 4) that there is homogeneity of within-group regression on the covariate. With each of these limitations, we lack the important test results—it appears doubtful, in fact, that points 2) through 4) are met, which calls both the ANCOVA and any ANOVA results into question. For example, the radically different (sometimes null) standard deviations reported across groups on most of Flynn's measures suggest non-homogeneity of variance.

While we are considering potential sources of error, the statistical analysis throughout Flynn's research on anaphora is jeopardized by there being unequal sample sizes across group levels (Spanish samples were 16, 21, and 14, Japanese were 7, 25, and 21, and Chinese were 11, 20, and 29, for Low, Mid, and High levels, respectively). For both ANCOVA and ANOVA, this is an important matter, more important than non-homogeneity of variance among groups. Unbalanced designs in which there is ALSO non-homogeneity have a high risk of being either too liberal or too conservative in their estimation of significant differences, depending on the relationships between group sizes and variances (see the above references as well as Elashoff 1969, Glass, Peckham and Sanders 1972, Hsu and Sebatane 1979, Hsu, Abunnaja, Zikri and Bugbee 1984, among many others). On the other hand, given that Flynn in fact finds very few significant differences in numerous tests of differences, even with her choice of a liberal level of significance, we might judge the overall gravity of a Type I error to be a relatively minor risk.

we will henceforth base our discussion as much as possible on the unadjusted means reported in Flynn (1984, 1986, 1987a) and Flynn and Espinal (1985).[18]

Despite the inappropriate use of the juxtaposed sentences, subjects' results on them are in fact quite interesting, for they suggest a different account of the ability to perform on the imitation tasks. Flynn reports (1987a:124) that *Redundancy* was a significant factor, with repetition of the verb-object redundancy sentences superior to the subject-redundancy, and this did not interact with the *Language* group factor. Flynn's preferred explanation for this is curious, primarily because she invokes the vague construct of "impose structure" in order to account for a result that on the surface has a straightforward explanation:

> . . .even in sentences where there is no overt embedding and no syntactic connector, Ss may still impose structure on their interpretation of these sentences. . . [B]oth groups of language learners approach the L2 language acquisition process in a structure-sensitive way. (p. 124)

Because the means by *Language* group and *Level* on the verb-object redundancy sentences were almost consistently at least TWICE as high as the subject-redundancy sentences (excepting where there were evident ceiling effects for the Spanish subjects), we judge their success to be based on the twice as great ease of imitating a stimulus in which TWO out of three constituents are repeated instead of ONE out of three. This behavior is consistent with the view that the subjects are to some extent simply repeating strings of words verbatim. Flynn in fact acknowledges this possibility, but only in a footnote: "Alternatively, these results MIGHT suggest..." (1987a:175, our emphasis). Later, we will have further occasions to propose simpler, processing explanations for these subjects' performance instead of the somewhat forced accounts provided by Flynn.

The fact that there are initially significant differences between the groups in speaking proficiency rules out any interpretation of the superiority of the

[18] However, for want of details on the significance of main and interaction effects on analysis of variance, we will assume that all effects reported for analyses of covariance are likely analogous to those for ANOVA. Comparison of unadjusted and adjusted means following her use of ANCOVA throughout her analyses on production tests (1987a:Chapter 7) shows that little is altered of the fundamental differences between the Japanese and Spanish groups in overall success on the tasks (or for that matter, little of the differences between proficiency levels within language groups). We will furthermore make the standard assumption that any effects not reported as significant are therefore non-significant.

Spanish group as being attributable to their greater ease in processing only the experimental sentences. On the whole, we will not question the LOCATION of effects, merely the INTERPRETATION of them as having to do SOLELY with the subjects' success or failure in imposing the targeted grammatical structures (i.e. subordination direction, anaphora type and direction) on the items.

3.2.2 Test of sentences involving *Directionality* and no anaphora

One set of test sentences (Flynn's "Test 2"—see the description above) was examining the directionality effect with adjunct adverbials alone. Flynn tested this in a *Directionality* (pre- and postposed) X *Language* (Japanese and Spanish) X *Level* (Low, Mid, and High) design. According to Flynn's specific hypotheses for this set, there should be 1) a significant difference on the *Language* group factor (due to the greater difficulty in processing these complex sentences by Japanese—and we suppose Chinese—learners), and 2) differences in "production abilities between pre-posed and post-posed complex sentence formation" (1987a:96). It is unclear exactly what the nature of the latter differences should be.

With respect to the first hypothesis, a main effect was found for the *Language* factor, where the Spanish subjects performed better than the Japanese (Flynn 1987a:136; we take these ANCOVA results to be equivalent to what ANOVA would result in). We find this result unsurprising, for as we argued in the case of the juxtaposed sentences, the Japanese are at a lower level of oral proficiency. As to the second hypothesis specifically regarding directionality, there were no significant results for the relevant main effects or interactions.[19]

Instead of reacting to the null results for *Directionality*, Flynn highlights the *Language* effect. Her summary (1987a:136) of the result is technically correct, while sidestepping the pattern hypothesis: "the two language groups differ significantly in their ability to produce complex sentences even when no anaphora or redundancy is involved." Nonetheless, Flynn and Espinal (1985)

[19] The overall UNADJUSTED means on the Test 2 sentences were 1.37 (Spanish) and .30 (Japanese), out of a possible score of 3. (The Japanese mean was reported as .37 in Flynn (1984:81), but we take the .30 to be the correct figure, as it has appeared several times in more recent publications and conference handouts.) These low scores and the large standard deviation (.42 for Japanese) suggest that the test battery was too difficult for the Japanese learners, probably with a floor effect. The low overall mean would have been still lower, were it not for the advanced Japanese subjects, who managed to achieve a mean score over 1 correct (1.05) on the postposed adverbial clause set. The other Japanese cell means were .38 and lower. This is probably one reason that there was no main effect for *Level*.

use the between-group differences in imitation as support for the hypothesis that the L1/L2 mismatch in head direction for the Japanese learners makes L2 acquisition more difficult, and they add a further claim as to the learners' internal psychological processes:

> [R]esults on imitation of the [Test 1 and Test 2] sentences ... indicated that the Japanese did not simply fare worse than the Spanish speakers because of the mismatch in head-direction between the L1 and L2. Rather results suggested that the Japanese learners were attempting to organize the L2 grammar around the head-initial configuration in English. (p. 96)

The source of Flynn's claim that the Japanese learners are adopting the L2 head direction parameter derives from one significant finding that supposedly supports a directionality effect. Flynn (1987a:136) reports a significant ($p = .03$) simple effects test (what she mistakenly calls a "main factor") of the difference in *Directionality* within only the Japanese High level subjects, with the postposed clause sentences (mean = 1.05) favored over the preposed clause ones (mean = .38). While such a finding appears to be an encouraging confirmation of the second specific hypothesis, it must be emphasized that Flynn's main effects and interaction results were not significant, and that, given the number of statistical tests she has conducted, such a finding of simple effects between two cells is of questionable reliability. It is not legitimate to perform such post hoc tests when there is no significant main or interaction effect involving *Directionality* and *Language Group* or *Level*.[20]

Flynn and Espinal (1985) repeat this error in their analysis of the Chinese group. No significant main effect for *Directionality* was reported for the Chinese group, nor a *Directionality* by *Level* interaction, yet the authors report a significant effect within the High level group, who also favored the postposed clauses over the preposed ones (means = 1.83 versus 1.21). They proceed to justify this result too with a parameter-setting interpretation. However, as we will see in the next section, there was no evidence of such a *Directionality* effect for either the Japanese or the Chinese learners on the Test 1 sentences (involving anaphora) at any level. We must question how Flynn can place so much stock on two post hoc intragroup and intralevel comparisons, when so many other results reveal no effects. The degree of

[20] This approach is even less legitimate when a liberal experiment-wide criterion of significance is maintained, and because it appears Flynn has used only the within-level error term in her test of simple effects.

selective treatment of results illustrated here seems unwarranted, especially when, as we will argue, there are alternative accounts of the overall results that involve less complex assumptions about the subjects' internalized grammars.

3.2.3 Test of sentences involving *Directionality* with anaphora

The Test 1 sentences involved the same *Directionality* X *Language* X *Level* factors as before, but the additional factor of *Anaphora* type was included (null or pronoun), occurring only in the subordinate adverbial clauses. The hypotheses remain that Japanese (or Chinese[21]) subjects should have greater difficulty in imitating these sentences, and that different patterns in imitation should result for these groups compared to the Spanish subjects. However, in the specific predictions for these sentences, Flynn (1987a) also hypothesized 1) that null and pronoun anaphora should be imitated similarly ("Spanish and Japanese would generalize over both pronoun and null anaphora," p. 95), but 2) that there would be an interaction of *Language* X *Level* X *Directionality* X *Anaphora* type. Flynn states specifically that:

> The Spanish *Ss* should demonstrate a significant preference for forward pronoun anaphora. We would expect the strongest effects of this interaction to be evidenced at the *Mid* or *High* levels for these speakers. This pattern would resemble L1 learners' acquisition of English (1987a:94)

Note at the outset that these two predictions risk being in conflict with one another. Any interaction involving *Anaphora* type could mean that hypothesis (1) was disconfirmed.

Furthermore, it is difficult to see how these specific predictions are related to Flynn's general hypotheses. She provides no rationale for the prediction that null anaphors and pronouns will be imitated similarly.[22] Likewise, the prediction of a four-way interaction does not derive in any obvious way from her exposition of parameter theory. She does not clarify the question of "early" and "later stages" of L1 acquisition sufficiently to lead to any such specific prediction of cross-linguistic differences in L2 acquisition by proficiency

[21] We must assume that predictions which Flynn (1987a) makes for Japanese subjects will hold for Chinese subjects as well.

[22] Indeed, she points out elsewhere (correctly) that null anaphors like those in the test sentences are accounted for by "a separate module of UG" from that which accounts for pronouns (1987a: 45).

level. We do not understand how she was able to foresee, prior to data collection and analysis, that the effect in question would only show up among the Spanish subjects, and then only at the Mid and High Michigan Placement Test levels, and there just for pronoun anaphora.

Flynn (1987a:128-133) reports that the four-way ANCOVA revealed significant main effects for *Language* group (Spanish again not surprisingly superior to Japanese), *Level* (each group was significantly poorer than the next more advanced one), and *Anaphora* (pronoun sentences were superior to null anaphors). These effects are at a high enough level of significance ($p < .0001$) that we trust the ANOVA results would be similar. As with the previous Test 2 sentences, there was again no effect for *Directionality*, nor any interaction,[23] except for the predicted four-way interaction of all factors (reported $p = .022$), and a two-way interaction of *Language* group by *Level* ($p = .04$).[24]

The main effect for *Anaphora* type has little bearing on the issues in question, although it does contradict Flynn's first specific prediction that the two types would elicit similar effects. We suggest that the sort of participial construction involved in the null anaphora sentences may have been unfamiliar to both the Spanish and Japanese subjects.[25] The effect is not

[23] Nor were Test 1 *Directionality* results significant in the Chinese study of Flynn and Espinal (1985), although Flynn does not report results of null pronoun sentences in this case.

[24] In these last two results, we question whether legitimate ANOVA results would obtain significance (this analysis, like the previous one, involved several of the lower level Japanese cells having null or extremely low means and variances, where analysis of covariance has evidently extrapolated the sort of "fictitious" means referred to earlier). Moreover, even if it were a true interaction, the *Language* X *Level* one is the least interesting result, since it is surely due to a floor effect for the two lowest Japanese groups. Furthermore, as is the case in many of Flynn's results, the means reported hide the fact that VERY FEW sentences are imitated correctly. For a group of 7 subjects (the Low level Japanese), each repeating 3 sentences of a given type, a mean of .14 out of 3 translates to ONE SENTENCE CORRECT out of all 21 sentences. According to Flynn's Table VII.4 (1987a:127), the Low level group had means of .14, .00, .00, and .14 on the preposed and postposed, null and pronoun anaphora items. Thus, two out of a total of 84 sentences, where at most TWO individuals account for the correct production. Similarly, for the Mid level Japanese group (n= 25), a mean of .04 indicates one sentence correct; the corresponding means for this group are .04, .08, .36, and .44, which translates into 1, 2, 9, and 11 sentences correctly repeated, out of 300 total. It is imaginable that in a group of 25 subjects, it would only take four or five rather proficient speakers to produce ALL the correct items. This is the likely source of the large standard deviations for most of these lower proficiency subjects on the test items (for the Low Spanish and especially the Low and Mid Chinese subjects reported in Flynn and Espinal 1985, as well), and the source of our concern that the data are highly skewed in many cases.

apparent at the Low levels, however (cf. footnote 24). In further analyses, we will see that Flynn fails to take the significant *Anaphora* type difference into account.

Since there was a four-way interaction, it is important to examine the potential source of it before any further analysis is conducted. As Flynn does not adequately explore this, we have plotted all the cell means for both Japanese and Spanish groups on the entire set of Test 1 sentences in Figure 1a. A plot of the Test 2 sentences is presented in Figure 1b for contrast.

Flynn tries to determine the source of the four-way interaction by making post hoc comparisons WITHIN the Spanish language group on the PRONOUN sentences, where she reports a significant *Directionality* by *Level* effect: "sentences with forward pronouns [i.e. postposed pronoun anaphora] are significantly easier to imitate for the Spanish Ss at the Mid level than are sentences with backward pronouns." (1987a:130)[26] While Figure 1a does in fact suggest that this could be the source of the four-way interaction, it also looks as if the High Spanish group may have reached a ceiling in imitation of pronoun anaphora sentences, thus contributing to a *Directionality* by *Level* effect within the Spanish/pronoun data.[27] Thus, the apparent difference in *Directionality* production of pronoun sentences may be an artifact of the relative difficulty of only a few test items. We see no other major differences or crossing of pattern. Since the general trend in the results appears to be towards no differences on *Directionality* for any of the subgroups, the rather large difference for pronouns at the Spanish Mid level looks interesting. But why would this effect be so limited?

3.2.4 Interpretation of results from Tests 1 and 2

There are three results from the preceding in which subjects favored

[25] Flynn tests both groups separately to demonstrate this, although there was no *Language* by *Anaphora* interaction to justify doing so. She offers no explanation for the result, however.

[26] This is again a case where, without other significant main or interaction effects (e.g. a *Language* by *Anaphora* effect, or a *Directionality* by *Language* by *Anaphora* effect), the finding of simple effects within one level of a language group on one type of anaphora may not have correctly localized the source. We would like to see the complete set of contrasts and error terms used to determine the significant differences.

[27] For the High group (n=14), the non-significant difference is four items less correct on the postposed sentences (30 sentences versus 34 for the preposed, out of 42 (14 X 3) total items for each type). For the Spanish Mid group (n = 21), the difference in total sentences correct on pronoun sentences is 16 (47 for the postposed sentences, and 31 for the preposed, out of 63 of each type).

forward directionality: Japanese and Chinese advanced learners on Test 2 non-anaphoric subordination, and Spanish intermediate learners on Test 1 pronoun anaphora sentences. Despite our reservations concerning the appropriateness of Flynn's methods of analysis, we will attempt to interpret these findings, but first we must see what interpretation Flynn offers. The key difference between language groups is NOT in the direction of the effect, but in the level where the effect appears, which of course appears to support the notion of Spanish learners acquiring the L2 direction parameter EARLIER than those groups whose L1s do not match the parameter. This is indeed Flynn's preferred interpretation. However, we find that her representation of the results greatly obfuscates the fact that the findings are really quite limited.

Especially with regard to the Spanish results, Flynn pays no attention to the overall finding of NO effect for *Directionality*, but selects instead the single supposed difference at the MID level on PRONOUN anaphora to justify her claim that the match in L1-to-L2 head direction leads to "sensitivity" to the parameter in question (see e.g. 1984:82). The phrasing of this claim varies from a technically correct description (1987a:136, para. 2), to the following unwarranted logical connection in a summary section:

> Test 1 demonstrated that for the Spanish Ss -- but not for the Japanese Ss -- Directionality of HD/AD [head direction/anaphora direction] combined does have a significant intra-language group effect, as seen in the significant interaction in Language Level. (1987a:140)

In fact, of course, no such demonstration occurred, especially if we consider the null effect for the other levels and for the null anaphora condition. Flynn also completely glosses over the *Language* group, *Level*, or Test sentence type (1 versus 2) limitations on this finding, as seen in the following quotes:[28]

> . . .[sentences] in which the antecedent preceded the pronoun . . .were also significantly easier for the Spanish Ss to imitate (Flynn and Espinal 1985:96, see also p. 103);

---

[28] Of course Flynn identifies the *Level-*, *Language-*, and sentence type-specific location of this effect at other points in some of these articles, but does not draw attention to these when reaching her conclusions. Flynn (1986), discussing only the Spanish data, does not even mention the lack of such a result in the Test 2 (null pronoun) and the Test 1 (no anaphora) sentences. This allows her to attribute greater implications, for the purpose of that article, to the supposed contrast between the lack of a directionality effect in the act-out comprehension sentences and the presence of one in the single case of pronoun anaphora sentences with Mid level subjects.

> ANOVA on amount correct in imitation indicate significant effects of directionality (1986:146);
>
> Patterns of acquisition of backward and forward pronoun anaphora in pre- and post-posed subordinate adverbial *when* clauses as measured by an elicited imitation (production) and an act-out (comprehension) task differed significantly. (1986:154)

Flynn brings a variety of other data analysis to bear on these issues, including error analyses of errors in production of each type of sentence (Flynn 1987a:141-157 for Spanish and Japanese, and selected results for Chinese in Flynn and Espinal 1985:104-106); comparison with production on the Test 3 sentences with preposed subordinate clauses but forward anaphora (1987a:137-140); and discussion of the comprehension act-out test items (1987a:157-172, also 1986:146-153). For lack of space, we will only deal selectively with some of these analyses. We find Flynn's arguments on the whole to be no more convincing than heretofore.

3.2.5 Alternative interpretations

Before we deal with these additional points, however, we would like to suggest what we believe to be a more parsimonious, less presumptive interpretation of the "pattern" of results evident in Figure 1 and, taking them cautiously, the significant comparisons within language groups and levels. Our interpretation involves three basic principles:

1) Retaining a subordinate clause in memory before processing a main clause puts added burden on recall memory (a principle suggested for native speakers in a variety of studies, e.g. Jarvella and Herman 1972, Townsend and Bever 1977, 1978).
2) As we have suggested earlier, there are universal tendencies favoring forward anaphora (Carden 1982, 1986). Reinhart puts it thus: "forward anaphora is the easiest form of anaphora to process while backward anaphora requires holding the pronoun in memory and going back to it." (1986:140)
3) Once learners attain a certain threshold in recognition of L2 clause structure, they will exhibit the principle 1 preference in production for the main-subordinate order.

Note that none of these principles invokes contrasts across L1s, not because we believe that contrastive specifications never have bearing on L2 acquisition, but because the contrastive use of Flynn's parameter-setting

model is insufficiently motivated in this on-line text processing study. In such a study, one would prefer an explanation based on processing strategies, especially when linguistic theory does not provide a precise linking of the phenomena involved.

Aside from possible idiosyncratic problems with sets of sentences and with certain subgroups in Flynn's study, we view the pattern of results seen in Figure 1[29] as an overall trend toward no effect for *Directionality* in either subordinate adverbials or anaphoric relationships.Conforming to Gallimore and Tharp's (1981) conclusions regarding elicited imitation, the learners are generally able to repeat sentences verbatim (incrementally as oral proficiency develops) regardless of the syntactic structures of the sentences.

However, once a given group of learners has attained a specific level of oral proficiency, they will tend to begin to process the presented sentences for MEANING, and thus be more subject to the processing constraints seen in psycholinguistic research on native speakers. The Mid level Spanish learners may have reached such a level, while only the High groups of Chinese and Japanese learners have done so. As we have seen, Flynn's grouping of subjects on the basis of a grammar and listening test did not equate the subjects on oral proficiency, nor does analysis of covariance correct for this mismatch. Thus, in comparison to Spanish learners, what is "late" is merely the development of Japanese or Chinese learners' oral proficiency relative to their grammatical development.

Moreover, we observe in Figure 1b a somewhat aberrant result for the Test 2 sentences (without anaphora), which we suspect may underlie the lack of differentiation for Spanish subjects as compared with the postposed favoring by Japanese and Chinese subjects. Note that the advanced Spanish learners appear to reach a ceiling on these sentences, thereby not allowing differentiation. The advanced Chinese subjects surprisingly attain an equivalently high performance, although only on the postposed subordinate clauses. The advanced Japanese subjects, performing considerably lower on these, seem radically affected by the preposed/postposed difference. Flynn does not call attention to these cross-group differences, and we do not see a likely explanation, other than that a threshold effect is being specially tapped

[29] The Chinese results, except for the advanced level that appears superior to the advanced Japanese in imitation ability, closely overlap the Japanese ones on anaphoric sentences, and show superiority on the non-anpahoric ones at both Mid and High levels. These results are consistent with our suggestion.

by this particular set of sentences. We can understand why there may be a lower ceiling on performance with these sentences (i.e. they are harder), as seen in the Spanish curve, for they contain the least redundancy of any of the sentence types (they are of the type: NP1 - VP1 - NP2 - when - NP3 - VP2 - NP4).

In the results with Test 1 and 2 sentences seen in Figure 1, as well as in the case of the subjects' imitation of juxtaposed sentences discussed earlier, we see that basic language processing principles provide a more straightforward account of the results. We will suggest a similar simpler alternative in our interpretation of Flynn's reported error analyses, and of the comprehension act-out sentences.

3.2.6 Analysis of error types

It should be noted from the start that Flynn's analysis of subjects' errors on the imitation tests (Flynn and Espinal 1985:103-106, Flynn 1986:146 and Table 2, 1987a:141-157) are presented largely as percentages of total errors for a given test and sentence type, and that compilations of these figures for different error types DO NOT ADD UP TO 100%.[30] That is, in the complete listing of these error analyses for Japanese and Spanish subjects, by Test, *Level*, and *Directionality* (Flynn 1987a), the SUM of percentage values for "lexical" errors (Table VII.7, p. 142), "one-clause repetition" errors (Table VII.9, p. 144), "conversion to coordination" errors (Table VII.12, p. 149) and "anaphora" errors (Table VII.14, p. 152) ranges between 65% and 110% for Test 1, and between 57% and 100% for Test 2. Flynn does not provide an adequate enough illustration of these error types or their calculation for the reader to determine where there might be overlap among them (thus leading to sums over 100%), or what other sorts of errors would make up the difference for those which fall well short of 100%. There does not appear to be any systematicity to the sum of percentages for any particular language or level group. Had frequencies of errors been provided, the reader could have computed the proportions.

Due to the difficulty of interpreting the reported percentages, we will not

[30] Recall that Flynn scored items as either correct, or in error, resulting in the total possible of 3 for each type of sentence in production (2 in comprehension). It is unfortunate she does not present the actual FREQUENCIES of errors in discussing the error analyses, for percentage analyses in the widely different types of sentences tends to obscure the relationships between errors. As will be seen, with percentages based on low frequencies, the differences reported appear to be based on only one or two errors.

dwell on each comparison Flynn (1987a) makes, only pointing out that numerous quantitatively based comparisons ("more," "greater," "significantly," etc.) are made (pp. 141-157), with neither statistical tests nor clarity about the quantities being compared (e.g. percentages are reported as "means").[31] We will illustrate especially by questioning the significance of conversion to coordination errors and anaphora errors.

Conversion to Coordination Errors

Flynn counts as "conversion to coordination" any repetition of the (*when*) subordinate-main clause stimulus items as coordinated "and" clauses. Such errors on Test 1 (3% for Spanish, 9% for Japanese) account for the smallest percentage of the total number of errors among the four types of errors. Yet, because Japanese learners had a higher proportion of this sort of error (19%) than Spanish learners (3%) when Tests 1, 2 and 3[32] were combined, Flynn attributes this greater relative proportion of Japanese error to their "difficulty maintaining a head-complement [sic] relation in the L2" (1987a:148). Moreover, she uses the comparison between the Japanese proportion on Test 2 sentences (with no anaphora - 33%) and Test 1 sentences (9%) to argue that a conversion to coordination error will only occur when learners can "maintain the requisite two-clause structure." The implication is that they can better process a two-clause structure with the Test 2 full sentences than with the anaphoric Test 1 sentences. Finally, Flynn claims that:

> . . .both groups of Ss differentiated the stimulus sentences structurally. There is a tendency for both groups to convert post-posed sentences rather than pre-posed sentences to coordinate structures (p. 150).

For all of these arguments, however, we suggest that again, a more parsimonious interpretation is derived from simple sentence processing constraints, based on the saliency of the sequence of words presented and the Japanese learners' low level of speaking proficiency. Their level of proficiency alone inhibits their production of complex sentences. The observation that maintenance of a two-clause structure is a prerequisite for the production of coordination errors seems indisputable, but it is not logically connected to the

[31] One exception is a reported analysis in Flynn (1986:146, and Table 2). Figure 2 (p. 148) related to these data is evidently erroneously drawn, for the values do not correspond to those in Table 2.

[32] We have not discussed Test 3 items for lack of space. Results for these are not critical to our position.

head direction parameter advocated by Flynn. The reason that Spanish learners show no differences in this regard is that they already have attained that level of speaking proficiency.

The most difficult argument to accept, however, is that the "tendency" for conversion to coordination of postposed sentences rather than preposed sentences indicates any sort of "structural" differentiation. A brief look at the stimulus sentences reveals that all postposed sentences have the *when* embedded in the sentences, while the preposed subordinate structures all BEGIN WITH *when*. Memory constraints in imitation would be sufficient to result in the slight differences[33] evident in Flynn's results, where even PERCEPTION and production of *and* instead of *when* could more likely occur in mid-sentence than sentence-initially.[34]

Anaphora Errors

Flynn divides "anaphora errors" into two types: "blocking" and "modification". In blocking errors, the anaphor is repeated as a full noun phrase, or vice versa. Modification errors involve a variety of types, including failing to repeat the pronoun and using a pronoun when the stimulus contained a null anaphor. In the next section, we point out that these "errors" do not necessarily indicate the failure of the subjects to assign anaphoric relations, nor for that matter is successful repetition a guarantee of correct anaphoric interpretation.

The results show Spanish subjects producing a greater relative proportion of anaphora "errors" than the Japanese (which she again attributes to the greater ability of Spanish subjects to maintain a two-clause structure). But since BOTH Japanese and Spanish subjects produce more anaphora "errors" on

[33] There is only a statistically untested, and not entirely evident numerical difference in the values in Table VII.12, p. 149.

[34] Flynn's (1987b) response, in a footnote, to this point made by a journal reviewer is that such an alternative explanation would lead to other predictions about her data. She claims that it would predict a consistently greater rate of conversion to coordination errors by the subjects on postposed clauses. Her data in Table VII.12 (1987a:149) reveal only two cases (out of 12 possible with Test 1 sentences) of inconsistency—with Low Japanese learners producing more such errors (10% versus 5%) with preposed clauses in the Test 1 pronoun sentences, and Mid Spanish learners producing more (5% versus 0%) in Test 1 null anaphora sentences. Yet, Flynn's position can offer no explanation for such INCONSISTENCY either, so that without specifics as to frequencies or the combinatory nature of these errors, and given the quite low rates and differentials of conversion errors, we find her defense rather weak.

preposed sentences than on postposed sentences, Flynn attributes this L2 consistent *Directionality* effect to acquisitional factors. She states, "This indicates that both groups are attentive to differences in head direction in the L2 and have more difficulty with the head-final sentence structures in L2." (p. 155) Since the two language groups behave similarly, however, this is again no argument specifically supporting a parameter-setting model over any other explanation, for example, one based on a universal preference for forward anaphora.[35]

3.3. TESTS OF THE HYPOTHESES IN COMPREHENSION

3.3.1 The need for a comprehension task

Flynn's model predicts that L2 learners will differ in their acquisition of anaphoric relationships depending on the primary head direction of the native language. In fact, Flynn's test of repetition accuracy does not itself provide evidence of anaphoric relationships. That is, when a learner attempts to repeat *When he delivered the message, the man questioned the lawyer*, we in fact have no way of knowing whether the learner construes *he* as coreferential with some NP in the sentence, or with some NP outside of the sentence, or even whether the learner conceives of *he* as anaphoric. The repetition results are irrelevant to Flynn's hypothesis unless the learner interprets *he* in the subordinate clause as coreferential to one of the two NPs of the main clause.[36]

The pattern of production errors might have provided evidence for coreference assignment. However, Flynn's reports of the error patterns do not

[35] It must be pointed out that one type of blocking error (Type II) has nothing whatsoever to do with anaphora direction per se. Moreover, Flynn conflates the results for Test 1 null and pronoun anaphora errors in her discussion, when, it will be recalled, there was found to be an overall SIGNIFICANT difference in rate of error between these two sentence types in her earlier ANOVA analysis. Any conflation of such significantly different measures is entirely illegitimate.

[36] Flynn suggests (1987a: 48) that the repetition task tests 'construal', rather than 'coreference'. Basically, construal involves determination of whether the structural conditions for a pronoun-antecedent relationship exist, while coreference assignment involves the (often pragmatically determined) decision of whether the elements are in fact coreferential. While the construal/coreference distinction is important to the theory of anaphora (see especially Hust and Brame 1976), we fail to see just how Flynn's studies separate the questions. Does she intend to claim that the subjects in the repetition task are making construals (in some sense) while not assigning coreference? How does she know? Flynn (1987a:47) devotes only three sentences to the construal/coreference distinction.

provide unequivocal evidence that coreference assignments are consistently being made or attempted. One-clause repetitions, the largest proportion of errors (over half the Japanese errors, over 80% of the Low Japanese level subjects' errors), provide no information about understood pronoun-antecedent relationships. Even when enough of the sentence is preserved to make the remaining error types informative, the picture is far from clear. Some "blocking errors" (Flynn 1987a:Appendix J), in which the learner replaced the pronoun in the subordinate clause with a full NP identical to one of the NPs in the main clause, show that the intended coreferentiality link was probably made (repeating *After he came in, John saw the girl* as *After John came in, John saw the girl*)[37]. But in other blocking errors, the pronoun is replaced by an NP from outside the sentence (repeating *When he entered the office, the janitor questioned the man* as *When the doctor entered the office, the janitor questioned the man*). This kind of error suggests that at least in some cases, learners are not interpreting intrasentential coreference. In a third kind of error, the learner switches pronoun and antecedent (repeating *After he came in, John saw the girl* as *After John came in, he saw the girl,* for example). In such examples, anaphoric relations are almost certainly interpreted as intended. Since Flynn unfortunately groups these error types with others in her analysis, it is impossible to draw any conclusions with respect to coreference.

For these reasons, and because the effects shown on the imitation tests are so limited, Flynn's comprehension test results are necessary to determine whether any effect exists at all. We will see, however, that they appear to conflict with the imitation test results.

3.3.2 Test of act-out sentences

Comprehension of pronoun anaphora was tested using an act-out task. Each

[37] The classification of such anaphora successes as "anaphora errors" derives from Lust. On this classification, Lasnik and Crain (1985) have commented:

> Imagine an experiment in which children were asked to imitate French sentences, and we found that some of them made the mistake of translating them into corresponding English sentences. We would surely conclude that children who made this error were in command of the rules of French. The repetition 'errors' in Lust's [1981] study have precisely the same character as those in our *gedanken* experiment. Thus, the conclusion must be the same; children who translated (30) [*Because she was tired, Mommy was sleeping.*] as (30') [*Because Mommy was tired, she was sleeping.*] must have access to rules allowing backwards anaphora. (pp. 149-150)

subject was asked to act out the meanings of orally presented test sentences using a set of colored plastic geometric shapes. The sentences described movements of the shapes with respect to each other, while manipulating *Directionality* and *Anaphora* type, giving four sentence types, as in imitation Test 1. The anaphor could be either a pronoun (*when it moved up and down*) or an empty subject (*when moving up and down*). The subordinate clauses could either be in sentence-initial position or in sentence-final position (*When it moved up and down the blue triangle touched the red square* or *The blue triangle touched the red square when it moved up and down*).

Flynn also introduced a new factor in the comprehension test, not present in the production test. The test sentences were sometimes presented with a "pragmatic lead"—a preceding sentence indicating what the test sentence is to be "about".[38] For example: *I'm going to tell you a sentence about a red square. When it moved up and down the blue triangle touched the red square.* In every case, the pragmatic lead (*PL*) established the direct object of the test sentence (the *red square* in the example) as what the sentence was about. Each of the four structural possibilities was presented with and without a pragmatic lead, and two sentences of each type were used, making a total of 16 stimulus sentences (and a possible score of 2 on each combination of variables).

---

[38] In the Lust studies on which Flynn's are modelled, the pragmatic lead variable is typically included in both the repetition and the act-out tasks. Flynn does not say why she decided not to include it in both. Moreover, the pragmatic lead variable has no relationship to Flynn's hypotheses, which deal solely with the claimed relationship between head direction and anaphora interpretation. Thus, if it should turn out that the pragmatic lead variable did matter, it would be difficult to interpret the result. Flynn apparently included this variable to explore properties of the test itself: in order to "evaluate whether the use of this biasing context would affect the subjects' judgments of coreference between the pronoun and the antecedent in these complex sentences. If so, this finding would help us determine what aspect of language knowledge comprehension evaluates most directly." (1986:143)

An important problem is that the pragmatic lead condition obscures the difference between forward and backward anaphora which is so crucial to Flynn's hypotheses. In an example like *I'm going to tell you a sentence about a red square. When it moved up and down the blue triangle touched the red square,* the pronoun *it* may well be taken to refer to the red square. But to which *red square*? Is it referring to *red square* in the following main clause or is it referring to *red square* in the preceding sentence (with the second *red square* referring directly back to the earlier *red square* by the device of repetition)? Or is it simultaneously referring in both directions? Flynn does not discuss this important problem.

3.3.3 Discussion of comprehension results for *Directionality*

Flynn (1987a) has reported comprehension test results only for the Spanish and Japanese subjects of her dissertation. Comprehension results for the Chinese were not reported in Flynn and Espinal (1985), and Flynn (1987b) presents only portions of the Spanish and Japanese PRODUCTION data.[39]

Before considering the results for *Directionality*, let us ask what Flynn would have expected to find. She makes no specific predictions in advance (1987a:98-99). We can, however, reconstruct what the predictions ought to be from her analysis of the repetition test results: Flynn interpreted the lack of a forward preference among the Japanese, in contrast to a preference in the Mid-level Spanish speakers (albeit on pronoun anaphora only), as proof that the Japanese are having trouble working out the head direction of English and its consequences for anaphora. One would thus expect a similar prediction for comprehension.

The results of the *-PL* comprehension condition showed *Directionality* to be not significant (1987a:160), although it interacted with *Language*. In the *+PL* condition, there was a significant effect for *Directionality*, favoring the forward anaphora sentences (1987a:164-165). Further analyses revealed, however, that in both experimental conditions, it appears to be the Japanese group which evidences a significant forward preference.[40] That is, in the *-PL* condition, the *Language* by *Directionality* interaction was caused by the Japanese forward preference, with no preference on the part of the Spanish; in the *+PL* condition, Flynn's post hoc within-language comparisons (although not exactly justified here, as the results showed no interaction with *Language*) found that the Japanese forward preference was significant, while the Spanish one was not. Far from buttressing the weak and unclear results of the

[39] The null anaphor results were not included in Flynn (1986), but they are reported in Flynn (1987a:159 ff.). On the combined null and anaphora sentences, the Japanese subjects, as we have come to expect, did not do as well as the Spanish subjects. Without a pragmatic lead, the Spanish overall correctness score was 1.22 as compared to Japanese .87. With a pragmatic lead, the figures are Spanish .98, Japanese .49. Flynn (1987a) reports both these results to be significantly different, but ventures no explanation. There was no *Anaphora* effect, nor interactions with *Anaphora*.

[40] The Japanese results are: without pragmatic lead, forward 1.01, backward .69; with pragmatic lead, forward .64, backward .36. This forward preference holds for both null and pronoun anaphora, and at all levels. The Spanish results with pragmatic lead reveal a slight favoring for forward directionality. The scores (out of 2 possible for each condition) for the cases without pragmatic lead are reported by Flynn (1987a:160); the figures for the cases with pragmatic lead were calculated by us based on Flynn's table VII.18, p. 163.

production tests, the comprehension test results seem actually to undermine them.

Nowhere does Flynn directly address the apparently contradictory findings of the comprehension and production tasks, or how her general hypotheses could possibly have predicted the lack of a forward preference for Spanish speakers and a substantial forward preference for Japanese speakers. In the published journal articles, either the production data are presented alone (as in Flynn and Espinal 1985, Flynn 1987b) or when both production and comprehension are reported (as in Flynn 1986), then only the Spanish data are presented. In Flynn's dissertation (published as Flynn 1987a), the problematic Japanese results are dealt with in the comment, "there is also a significant overall effect of directionality for the Japanese." (p. 174)

In the context of discussing a limited subpart of the Spanish results (only the pronoun anaphora sentences), while ignoring the Japanese results entirely, Flynn (1986) did propose that act-out tasks are somehow less satisfactory as a test of structural aspects of language knowledge than imitation tasks. She argued as follows:

> From these results, we can conclude that while production (elicited imitation) and comprehension (act-out) both elicit data that can be evaluated for evidence of linguistic competence, the degree to which each accesses this knowledge differs significantly. Specifically, the lack of a significant directionality effect [for the Spanish subjects] as well as the enhanced performance in comprehension suggest that in act-out (comprehension) tasks a subject need not tap into structure as directly as in production. (1986:154)

With a similar argument Flynn might propose that the Japanese comprehension results should not be taken seriously either. However, the evidence which she cites to demonstrate this limitation of act-out is nothing more than the lack of the directionality effect.[41] We have seen that only the

[41] Flynn does also mention "enhanced performance" on the comprehension test. We fail to see how a comparison between the act-out and imitation scores can be fairly made—the tasks are very different, the scoring systems use different criteria, even the number possible correct is different, so the scores are basically incommensurable. The difference in the comprehension test between the cases with and without pragmatic lead might be taken to indicate a special context-sensitivity for act-out. But since Flynn's studies provide no information about pragmatic lead in imitation, it is hard to conclude much. No doubt, different tests measure different aspects of linguistic knowledge, but we doubt that it is simply that comprehension tests are measures of linguistic knowledge to a lesser degree. If anything, one might well hold

Spanish results show this lack. On the basis of the Japanese data, she should conclude that it is act-out, rather than imitation, which "taps into structure most directly". As far as we can see, the only reason to discount the problematic comprehension results is that subjects do not behave as Flynn believes they should.

3.3.4 Alternative interpretation

We will only briefly propose an alternative account of the comprehension test results, which we believe again is more parsimonius and reasonable, and which is consistent with our position earlier with regard to the imitation test results. The evidence tends to support a forward directionality preference, independent of L1, which will show up if the learners are at an appropriate level of development and the test is sensitive to that effect. If the learners are at too low a level of development, no effect shows up because of the learners' inability to process the stimuli to the requisite depth of syntactic analysis, and because the tests themselves are too difficult to detect the effect.[42]

Accordingly, we suggest that Flynn's comprehension task lends itself to a fairly direct interpretation by the typical second language learner, so that after a certain level of proficiency is reached, manipulation of *Directionality* will not elicit effects. However, learners who are at a particular lower stage in their development will evidence the forward preference in sentence processing that is evoked by the specific set of stimulus sentences used. Another sort of task or stimuli might reveal the forward preference even with more advanced learners. The difference between the production and comprehension data in the two language group's performance suggests that the comprehension task is generally easier for these subjects' degree of syntactic develoment. Thus, the Spanish learners are too advanced to evidence an effect for forward directionality on the comprehension task, while all the Japanese learners fall into the range of proficiency on which the task succeeds in eliciting the forward preference. Note that the lower level of performance on the $+PL$

(as psycholinguists generally have held) that act-out is a particularly appropriate test when anaphoric relations are at issue.

[42] The astute reader will of course have noticed that, as in our interpretation of the production results, we are, like Flynn, making use of the notion of "tapping into [learners'] structure". We emphasize again that we have no objections to such a notion. We are only questioning the validity of the rather powerful theoretical model which Flynn attempts to invoke, but which has not served her well in reconciling her experimental production and comprehension results.

condition (for both groups) led to a tendency for the Spanish subjects to demonstrate the forward preference as well.[43]

4. CONCLUSION

We emphasize, in conclusion, that we find valuable the sort of cross-linguistic comparisons in second language acquisition attempted by Flynn. More experimental research is needed, testing a variety of linguistically and psychologically motivated hypotheses, in order to expand our knowledge of the internalized grammars and language processing capabilities of L2 learners, as well as to evaluate more precisely the use in L2 studies of different elicitation techniques and methodological paradigms. Methods developed for first language learners may prove unsatisfactory in L2 contexts, either because of the learners' difficulty and unfamiliarity with the testing procedures, or because of unnoticed assumptions about the learners' abilities with phonology, vocabulary, and syntax, or differential productive and receptive proficiency.

Certain of these and other methodological problems, along with numerous analytical errors, have considerably reduced the power of Flynn's study to support her intended goal, namely, to develop a predictive parameter-setting model of second language acquisition. Flynn's clearest finding is that at a given intensive course placement level, Spanish learners outperform Japanese and Chinese learners in tasks of oral comprehension and production. We consider this entirely unsurprising. In addition, in some cases there were signs of a preference for postposed adverbial clauses and for forward anaphora. This tendency seems to us to hold independently of native language and is quite consistent with pragmatic, processing accounts of linguistic performance.

[43] The lower scores on *+PL* may have been the result of the pragmatic oddity of the lead stimulus. The lead sets up a topic, but the next sentence then places that topic in main clause object position (a prototypical position for non-topics), while it places a different, newly introduced entity in main clause subject position (the prototypical place for known, topic elements). This is not the usual way to encode discourse structure. In fact, the pragmatic lead might be called the pragmatic confuser. Adding a subordinate clause with a pronoun subject makes things even worse. A natural tendency would be to interpret the pronoun as coreferential with the main clause subject. But, alas, the pragmatic confuser has just introduced a different salient discourse topic, and thus, a competing tendency is generated to identify the pronoun with that topic. In addition, the pragmatic lead is given a distinctly odd formulation—odd at least to our ears. A native speaker would not normally use the collocation *tell a sentence*: one tells stories, not sentences. No wonder comprehension is difficult!

Only through selective focus is Flynn able to argue a connection between her basic hypotheses and the findings.

Flynn's goal is further undermined by the inadequate development of a model of parameter setting. Perhaps due to her initial dependency on the Principal Branching Direction model, which she abandoned as a rationale in favor of the Universal Grammar-like head direction model, Flynn is left with an incoherent and often vague set of hypotheses and predictions with regard to the different languages and L2 learners in her study.

We fully accept the notion that L2 learners need to develop grammatical systems in order to become proficient in sentence comprehension and production. We should investigate these grammars through careful, reliable elicitation in experimental conditions. What we have claimed in the current paper is that Flynn's model is ill-defined, her methodology problematic, and her interpretation too forced. Thus, her study cannot reveal the precise constraints underlying the grammatical system that these learners have acquired.

Received December 20, 1987.
Authors' address for correspondence:

Robert Bley-Vroman
Craig Chaudron
Department of English as a Second Language
University of Hawai'i at Manoa
1890 East-West Road
Honolulu, HI 96822

REFERENCES

Bley-Vroman, Robert. In press. The logical problem of foreign language learning. *Linguistic Analysis.*

Carden, Guy. 1982. Backwards anaphora in discourse context. *Journal of Linguistics* 18.361-387.

_____. 1986. Blocked forwards coreference: theoretical implications of the acquisition data. In B. Lust (ed.) *Studies in the acquisition of anaphora: Volume !: defining the constraints,* 319-358. Dordrecht: D. Reidel.

Chaudron, Craig. 1983. Research on metalinguistic judgments: a review of theory, methods, and results. *Language Learning* 33:3.343-377.

_____ . 1986. Intake: on models and methods for discovering learners' processing of input. *Studies in Second Language Acquisition* 7:1.1-14.

Chomsky, Noam. 1981. *Lectures on government and binding.* Dordrecht: Foris.

_____ . 1982. *Some concepts and consequences of the theory of government and binding.* (Linguistic Inquiry Monograph 6). Cambridge. Mass.: MIT.

_____ . 1985. *Knowledge of language: its nature, origins, and use.* New York: Praeger.

Clahsen, Harald and Pieter Muysken. 1986. The availability of universal grammar to child and adult learners. *Second Language Research* 2:2.93-119.

Elashoff, Janet D. 1969. Analysis of covariance: a delicate instrument. *American Educational Research Journal* 6.383-401.

Flynn, Suzanne. 1983. *A study of the effects of principal branching direction in second language acquisition: the generalization of a paramater of universal grammar from first to second language acquisition.* Cornell University dissertation.

_____ . 1984. A universal in L2 acquisition based on a PBD typology. In F. R. Eckman, L. H. Bell, and D. Nelson (eds.) *Universals of second language acquisition,* 75-87. Rowley, Mass.: Newbury House.

_____ . 1986. Production vs. comprehension: differences in underlying competences. *Studies in Second Language Acquisition* 8:2.135-164.

_____ . 1987a. *A parameter-setting model of L2 acquisition: experimental studies in anaphora.* Dordrecht: D. Reidel.

_____ . 1987b. Contrast and construction in a parameter-setting model of second language acquisition. *Language Learning* 37:1.19-62.

Flynn, Suzanne, and I. Espinal. 1985. Head-initial/head-final parameter in adult Chinese L2 acquisition of English. *Second Language Research* 1:1.93-117.

Gallimore, Ronald, and Roland G. Tharp. 1981. The interpretation of elicited sentence imitation in a standardized context. *Language Learning* 31:2.369-392.

Glass, Gene V., Percy D. Peckham, and James R. Sanders. 1972. Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. *Review of Educational Research* 42:237-288.

Hilles, Sharon. 1986. Interlanguage and the pro-drop parameter. *Second Language Research* 2:1.33-52.

Hsu, Tse-chi, and E. Molapi Sebatane. 1979. Unequal covariate group means and the analysis of covariance. *Journal of Experimental Education* 47.222-229.

Hsu, Tse-chi, Salaheddin S. Abunnaja, Lawrence Zikri, and Alan C. Bugbee, Jr. 1984. The robustness of the analysis of covariance to the violation of various assumptions. Working Paper #40, Department of Educational Research, University of Pittsburgh, Pittsburgh.

Huang, Cheng-Teh James. 1982. *Logical relations in Chinese and the theory of grammar.* MIT dissertation.

Huitema, B. E. 1980. *The analysis of covariance and alternatives.* New York: Wiley and Sons.

Hust, Joel and Michael K. Brame. 1976. Jackendoff on interpretive semantics. *Linguistic Analysis* 2.243-277.

Jarvella, Robert J., and Steven J. Herman. 1972. Clause structure of sentences and speech processing. *Perception and Psychophysics* 11.381-384.

Keppel, Geoffrey. 1973. *Design and analysis: a researcher's handbook.* Englewood Cliffs, New Jersey: Prentice-Hall.

Koopman, Hilda. 1985. *The syntax of verbs: from verb-movement rules in Kru languages to universal grammar.* Dordrecht: Foris.

Kirk, Roger E. 1968. *Experimental design: procedures for the behavioral sciences.* Belmont, California: Brooks/Cole.

Langacker, Ronald. 1969. Pronominalization and the chain of command. In David Reibel and Sanford Schane (eds.) *Modern studies in English*, 160-186. Englewood Cliffs, N.J.: Prentice Hall.

Lasnik, Howard, and Stephen Crain. 1985. Review article: on the acquisition of pronominal reference. Review of *Pronominal reference: child language and the theory of grammar*, by L. Solan. *Lingua* 65.135-154.

Liceras, Juana M. 1985. The value of clitics in non-native Spanish. *Second Language Research* 1:2.151-168.

Long, Michael H. 1987. Maturational constraints on language development. Plenary address at the Seventh Second Language Research Forum, University of Southern California, February 20, 1987.

Lust, Barbara. 1981. Constraints on anaphora in child language: a prediction for a universal. In S. L. Tavalokian (ed.) *Language acquisition and linguistic theory*, 74-96. Cambridge, Massachusetts: MIT.

_____. 1986 (ed.) *Studies in the acquisition of anaphora: Volume I: defining the constraints*. Dordrecht: D. Reidel.

Lust, Barbara, and Tatsuko Kaneda Wakayama. 1979. The structure of coordination in children's first language acquisition of Japanese. In F. R. Eckman and A. J. Hastings (eds.) *Studies in first and second language acquisition*, 134-152. Rowley, Massachusetts: Newbury House.

Lust, Barbara, and L. Mangione. 1983. The principal branching direction parameter constraint in first language acquisition of anaphora. In P. Sells and C. Jones (eds.) *NELS 13* , 145-160. Amherst, Massachusetts: University of Massachusetts.

Lust, Barbara, and Y. C. Chien. 1984. The structure of coordination in first language acquisition of Mandarin Chinese: evidence for a universal. *Cognition* 17:1.49-83.

O'Grady, William, Yoshiko Suzuki-wei, and Sook Whan Cho. 1986. Directionality preferences in the interpretation of anaphora: data from Korean and Japanese. *Journal of Child Language*, 13.409-420.

Pedhazur, Elazar J. 1982. *Multiple regression in behavioral research: explanation and prediction, 2nd edition*. New York: Holt, Rinehart and Winston.

Reinhart, Tanya. 1976. *The syntactic domain of anaphora*. MIT dissertation.

_____. 1983. *Anaphora and semantic interpretation*. London: Croom Helm.

_____. 1986. Center and periphery in the grammar of anaphora. In Barbara Lust (ed.) *Studies in the acquisition of anaphora: Volume I: defining the constraints.* Dordrecht: D. Reidel.

Roeper, Thomas and Edwin Williams. 1987 (eds.) *Parameter setting.* Dordrecht: D. Reidel.

Townsend, David J., and Thomas G. Bever. 1977. Main and subordinate clauses: a study in figure and ground. Bloomington, Indiana: Indiana University Linguistics Club.

_____ . 1978. Interclause relations and clausal processing. *Journal of Verbal Learning and Verbal Behavior* 17:509-521.

Travis, L. 1984. *Parameters and effects of word order variation.* MIT dissertation.

White, Lydia. 1985a. The "pro-drop" parameter in adult second language acquisition. *Language Learning* 35:1.47-61.

_____. 1985b. The acquisition of parameterized grammars. *Second Language Research* 1:1.1-17.

Wildt, Albert R., and Olli T. Ahtola. 1978. *Analysis of covariance.* Beverly Hills, California: Sage Publications.