Order Number 9334926

A comparison of grammaticality evaluation measurements: Testing native speakers of English and Korean

Masterson, Deborah Ann, Ph.D.

University of Hawaii, 1993

U·M·I
300 N. Zeeb Rd.
Ann Arbor, MI 48106

A COMPARISON OF GRAMMATICALITY EVALUATION MEASUREMENTS: TESTING NATIVE SPEAKERS OF ENGLISH AND KOREAN

A DISSERTATION SUBMITTED TO THE GRADUATE DIVISION OF THE UNIVERSITY OF HAWAI'I IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

IN

LINGUISTICS

AUGUST 1993

by

Deborah Ann Masterson

Dissertation Committee:

Robert Bley-Vroman, Chairperson
Roderick Jacobs
William O'Grady
Ho-min Sohn
Marshall Pihl

to Jay

ACKNOWLEDGEMENTS

Each of the members of my committee had valuable contributions to make to this dissertation. I thank Marshall Pihl for his cheerful demeanor, sharp questions, and his contribution to historical accuracy. Ricky Jacobs is appreciated for his sympathetic encouragement as well as for his syntactic acuity — in terms of both syntactic theory and pointing out to me my own twisted syntax. I credit Ho-min Sohn for patiently introducing me to Korean linguistics and for his suggestions regarding the Korean input. 교수님, 감사합니다! I thank William O'Grady for the large part he has played over the years in the development of my own understanding of syntax, from the basic knowledge I acquired in his introductory courses to the specific insights about the present study gleaned from our discussions of the details. Dr. O'Grady's comprehensive knowledge of and unflagging enthusiasm for his chosen field is an inspiration to all the students of our department. Above all, I gratefully acknowledge my chair, Robert Bley-Vroman, for his generosity, patient encouragement, and guidance in this project from its inception.

I received help from many quarters in completing this study. The data collection was supported in part by a grant to Craig Chaudron and Robert Bley-Vroman from the Dean's Research Fund of the University of Hawai'i College of Languages, Linguistics and Literature, Richard K. Seymour, Dean. Special thanks go to my brother-in-law, Choi Cheol-kyu, for arranging access to

the Korean participants. Many thanks for help with data collection go to Craig Chaudron, Soo ah Yuen, and Bill Johnston.

I must not fail to acknowledge here the many kinds of support I have received over the years from the diligent and brave Korean students in our department. 학생들, 감사합니다! I send my best aloha to the charter members of my support group, who were convinced I could do it, even when I had my doubts — to Julie Winter, who knew when it was time to work and when it was time to swim and to Kate Wolfe-Quintero, who took time she didn't have to help me define the scope of this study and navigate the analysis of the data.

Finally, to my husband, Jay Jeong-ryeol Kim — to even attempt to detail all the love and support he has provided would be to trivialize his contribution. He was with me in the beginning, and has seen me through to the sweet conclusion — my PhD., H.M., S.H.

ABSTRACT

This dissertation compares three instruments which have been used to obtain grammaticality evaluations in both first and second language research. In order to test the instruments, (1) a written questionnaire, (2) an elicited imitation task, and (3) a reaction time task, I assume the validity of a grammatical framework which accounts for a subset of modified nominal phrases in both English and Korean. These nominal phrases are analysed in Chapter 2 in terms of their grammaticality, referring to the DP analysis of Abney (1987) and the complementary DelP analysis of Yoon (1990), both formulated within the Government and Binding framework. Sentences used to exemplify the nominal phrases are also analysed in terms of their length in chunks, adopting a system developed by Abney (1991). The sentences serve as the input to an experiment designed to compare the three methods of obtaining grammaticality evaluations from 49 English and 44 Korean native-speaking participants.

Chapter 3 contains a discussion of the nature of metalinguistic awareness and its relation to performance, competence, and grammaticality evaluations. In Chapter 4, the overall research design and each of the three instruments being investigated are described. Chapter 5 contains the primary analyses of the data obtained. It is found that the written questionnaire is primarily, though not exclusively, sensitive to the grammaticality of the input sentences as opposed to the length in chunks. The elicited imitation instrument is sensitive to

grammaticality and length in chunks in more or less equal measure, and the reaction time instrument is sensitive to length in chunks alone. In Chapter 6, further analysis of the data reveals a challenge to the DelP analysis of Korean analysis.

In the final chapter, three main themes arising from the present investigation are identified and discussed. These are: (1) the importance of understanding what a specific grammaticality evaluation instrument measures, (2) the value of converging evidence, and (3) the need for independent data in validating linguistic theory.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ACC	accusative case marker (Korean)
AGR	agreement
AP	maximal projection of (A)djective
C	length in (C)hunks
CP	maximal projection of (C)omplementizer
DAT	dative case marker (Korean)
DP	maximal projection of (D)eterminer
DelP	maximal projection of (D)elimiter
ENGs	English-speaking participants
F&F	Freedman and Forster (1987)
G	(G)rammaticality
GB	Government and Binding framework (Chomsky, 1981)
GEN	genitive case marker (Korean)
IP	maximal projection of (I)nflection
KORs	Korean participants
L	(L)anguage
L1	first, or native language
L2	second language
LOC	locative case marker (Korean)
MOD–NP	a noun modified by a determiner and adjectives
NOM	nominative case marker (Korean)
NP	maximal projection of (N)oun
POS–NP	a noun modified by a genitive-marked possessive nominal, which in turn is modified by a genitive-marked nominal
PP	maximal projection of (P)reposition
PST	past tense marker (Korean)
SE	sentence end (Korean)

SPEC	specifier
STM	short term memory
XP	maximal projection of an unspecified category
2ARG–NP	a nominal phrase containing an external argument plus one subcategorized internal argument
3ARG–NP	A nominal phrase containing an external argument plus two subcategorized internal arguments
θ-role	theta-role

CHAPTER 1: INTRODUCTION

Theories of grammar are typically constructed in a vacuum as far as independent evidence supporting the judgments of grammaticality by actual native users of a language is concerned, yet such evidence should be an essential pilar of any theory which claims even observational adequacy. Although this situation has been remedied to some extent in recent years, especially in the area of second language acquisition studies, independent supporting evidence of native language users' evaluations of grammaticality, including converging evidence from cross-validating measures remains a critical lack in much first language theory construction and validation.

Having said this, I embark on this study in the reverse fashion. In this study, in order to compare and evaluate different ways of measuring native language users' performance and interpreting the resulting data for the purpose of suggesting a picture of underlying competence, I am assuming as given the validity of the selected syntactic framework. This framework and the justification for its selection is presented in Chapter 2.

Chapter 3 is an overview of metalinguistic awareness. Metalinguistic awareness can be thought of as the ability to observe language as an object removed from its communicative function; this ability is present to varying degrees in all language users. Based upon observations of this ability in action, the above-mentioned picture of underlying competence might be formed. This picture

is useful in validating or challenging existing linguistic theory, but only to the extent to which the instrument by which grammaticality evaluations are collected can be trusted to provide a reasonably noise-free data. By "noise-free", I mean data that not only actually represents a true measurement of the object of interest — in this case, the grammaticality of the structures under investigation, but is also credibly free of effects from interfering and competing factors.

In Chapter 4, I give a detailed description of the overall research design with respect to the construction of the input, the participants, and the administration of the tasks. I also describe each of the three different instruments being investigated, the written questionnaire ("pencil-and-paper" or "P&P"), the reaction time task ("RT"), and the elicited imitation task ("EI"). Previous investigations and uses of these three instruments are summarized and the design of each instrument is detailed in turn, paying particular attention to design considerations unique to each instrument.

Chapters 5 and 6 contain the analyses of the data obtained via the three instruments. In Chapter 7, I summarize my findings and discuss three main themes regarding the relationship between linguistic theory and linguistic data collection which surface as a result of the investigation undertaken here.

CHAPTER 2: MODIFIED NOMINAL PHRASES IN ENGLISH AND KOREAN

2.1. General description of nominal phrases

The structure under investigation is modified nominal phrases, that is, phrases in which nouns combine with adjectives, genitives, complements, and determiners. English and Korean differ with respect to the types of nominal phrases permitted and, of particular interest here, with respect to the permissible positions of specifiers and modifiers within the phrases. The present study is concerned with three categories of modified nominal phrases, within which are found contrasts between the two languages. The sentences discussed here serve as the input to the grammaticality evaluation study undertaken (see Appendix A for the complete set in each language). The categories are outlined below.

2.1.1. MOD–NP category

The first category of nominal phrase under consideration is that in which a noun is modified by a determiner and adjectives (referred to hereafter as "MOD–NP"). This is exemplified[1] below in both English and Korean (Yale romanization of Korean is used throughout):

[1] Support for judgments of grammaticality appear in Section 2.2.

(1)	a.	the big blue book	a'.	ku khun phulun chayk *the big blue book*
	b.	* big the blue book	b'.	khun ku phulun chayk *big the blue book*
	c.	* big blue the book	c'.	khun phulun ku chayk *big blue the book*
	d.	* big blue book the	d'.	* khun phulun chayk ku *big blue book the*

In Korean, provided that the head noun[2] appears in phrase-final position, the position of the determiner *ku* is free with respect to the other elements of the phrase (1a'. –1c'.). In contrast, English permits determiners in the phase-initial position only (1a.; cf. *1b. and *1c.). Neither language permits phrase-final determiners (*1d. and *1d'.).

2.1.2. 2ARG–NP category

The second category of modified nominal phrase examined here consists of a nominal phrase containing an external argument plus one subcategorized internal argument, henceforth "2ARG–NP". Pre-nominal genitives are marked in English by attachment of the possessive morpheme *'s* to the noun; genitives in Korean are marked by the possessive morpheme *-uy*. Consider the sentences in (2) below; note that both the external argument, *Japan*, and the internal argument, *Seoul*, may appear as a genitive-marked nominal in both English (2c. and 2d.) and Korean (2 a'. and 2b'.). However, while the co-occurrence of both genitive-marked nominals is allowed in Korean (2 a'.)[3], such is not the case in

2 For the purposes of pre-theoretical discussion of these structures, the term "head noun" is not necessarily meant to indicate the syntactic head of the nominal phrase, rather it is used to refer to the conceptual or semantic head of the nominal phrase, i.e., the "thing" being referred to.

English (*2 a. and *2b.). In English, when one of the arguments of the head noun is genitive-marked, the other appears post-nominally as a complement of a preposition (2c. and 2d.). Alternatively, both may appear in post-nominal prepositional phrases (2e.). In contrast, all examples in Korean in which elements appear following the head noun are ungrammatical. (*2c'., *2d'., and *2e'.). Again, note that phrase-final determiners are ungrammatical in both languages (*2f., *2f'.). In English, a determiner is grammatical only in the phrase-initial position (2e.), and only in complementary distribution with a genitive-marked nominal argument (cf. 2c., 2d.).

(2) a. * Japan's Seoul's destruction
    a'. ilpon-uy sewul-uy phakoy
    Japan-GEN Seoul-GEN destruction
  b. * Japan's Seoul's the destruction
    b'. * ilpon-uy sewul-uy ku phakoy
    Japan-GEN Seoul-GEN the destruction
  c. Japan's destruction of Seoul
    c'. * ilpon-uy phakoy sewul-uy
    Japan-GEN destruction Seoul-GEN
  d. Seoul's destruction by Japan
    d'. * sewul-uy phakoy ilpon-uy
    Seoul-GEN destruction Japan-GEN
  e. the destruction of Seoul by Japan
    e'. * ku phakoy sewul-uy ilpon-uy
    the destruction Seoul-GEN Japan-GEN
  f. * destruction of Seoul by Japan the
  f'. * ilpon-uy sewul-uy phakoy ku
    Japan-GEN Seoul-GEN destruction the

---

3 2b' is ungrammatical on other grounds, as explained in Section 2.2.3.2.

2.1.3. 3ARG–NP category

A nominal phrase containing an external argument plus two subcategorized internal arguments is referred to as a "3ARG–NP". The situation here is very similar to the case of 2ARG–NPs. As seen above, only one of the arguments may appear as a genitive-marked nominal in English (3c. and 3d.); the remaining arguments appear post-nominally. Or, all three are permitted post-nominally as shown in 3e. All examples in Korean in which elements appear following the head noun are ungrammatical (*3c'., *3d'., *3e'.)[4]. As above, phrase-final determiners are ungrammatical in both languages (*3f., *3f'.). In English, a determiner is grammatical only in the phrase-initial position (3e.), and only in complementary distribution with a genitive-marked nominal argument (cf. 3c., 3d.).

(3) a. * Steven's to Jenny of flowers gift
  a'. Steven-uy Jenny-ey-key-uy kkoch-uy senmul
  Steven-GEN Jenny-DAT-GEN flowers-GEN gift
  b. * Steven's to Jenny of flowers the gift
  b'. * Steven-uy Jenny-ey-key-uy kkoch-uy ku senmul
  Steven-GEN Jenny-DAT-GEN flowers-GEN the gift
  c. Steven's gift of flowers to Jenny
  c'. * Steven-uy senmul kkoch-uy Jenny- ey-key-uy
  Steven-GEN gift flowers-GEN Jenny-DAT-GEN
  d. Jenny's gift of flowers from Steven
  d'. * Jenny-ey-key-uy senmul kkoch-uy Steven-uy
  Jenny-DAT-GEN gift flowers-GEN Steven-GEN
  e. the gift of flowers from Steven to Jenny
  e'. * ku senmul kkoch-uy Steven-uy Jenny-ey-key-uy
  the gift flowers-GEN Steven-GEN Jenny-DAT-GEN

---

4 3b' is ungrammatical on other grounds, as explained in Section 2.2.3.2.

f. * gift of flowers from Steven to Jenny the

f'. * Steven-uy Jenny-ey-key-uy kkoch-uy senmul ku
Steven-GEN Jenny-DAT-GEN flowers-GEN gift the

To summarize, modified nominal phrases in English and Korean are similar in some ways and contrast in others. First, with respect to permissible positions for determiners, English is superficially the more constrained system, allowing determiners in the phrase-initial position only. Korean allows determiners in a variety of positions within the MOD–NP structures, but, like English, does not allow them post-nominally, in the phrase-final position. This is presented diagrammatically in Figure 2–1. The numbers 1 through 4 indicate possible position for a determiner in a nominal phrase consisting of a head noun, two adjectives and a determiner, i.e., 1=phrase-initial, 2=pre-modifier (but not phrase-initial), 3=pre-noun, and 4=phrase-final. English and Korean agree that the phrase-initial position is an acceptable one for determiners and the phrase-final is not. They do not agree on the acceptability of phrase-internal determiners.

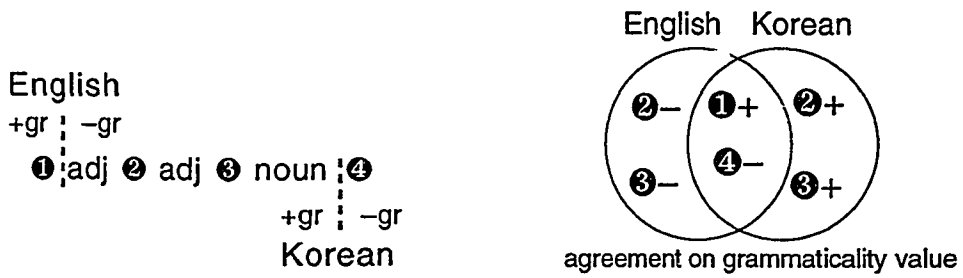


Figure 2–1: Possible positions for determiners within a MOD–NP structure

Second, in nominal phrases in which the head noun subcategorizes multiple arguments, both languages allow these arguments to appear as genitive-marked

nominals. Korean allows all arguments to be so marked, in English, only the phrase-initial argument is marked as genitive; the remaining arguments appear in post-nominal prepositional phrases (PPs), or, alternatively, all arguments may appear post-nominally. In both languages only pre-nominal arguments are genitive-marked with the possessive morpheme. With respect to allowable positions for arguments of the head noun, Korean appears as the more constrained system, allowing arguments to appear only pre-nominally, i.e., in all nominal phrase structures, the head noun *must* be final. These observations are presented diagrammatically in Figure 2–2. The numbers 1 through 4 indicate possible positions for the head noun in a nominal phrase consisting of a head noun and three subcategorized arguments. For ease of exposition, possible positions for determiners are not shown here. Possible positions for the head noun are: 1=phrase-initial, 2=following the first argument, 3=following the second argument, 4=phrase-final (following the third argument). Neither language allows a head noun in the number 3 position. In both languages pre-nominal arguments are genitive-marked.

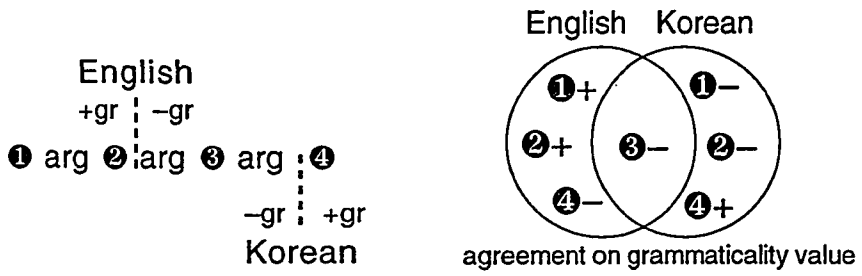


Figure 2–2: Possible positions for head noun within a nominal phrase

In the following section I review analyses of these structures formulated within the Government and Binding framework which account for the above-described phenomena.

2.2. Review of the literature: Analyses of nominal phrases

In this section, before reviewing the analyses adopted, I discuss first the concept of functional versus lexical categories with respect to nominal phrases in both English and Korean. This distinction between category types is widely accepted within the Government and Binding (GB) framework (Chomsky, 1981). As for the analysis of English, I rely heavily on Abney's (1987) "DP analysis" of nominal phrases; for Korean, I refer primarily to the "DelP analysis" (Yoon, 1990), a further refinement of Abney which accommodates nominal phrase structures in Korean, including the phenomenon of "scrambling". In addition, I present an alternative view of Korean nominal phrases, based on the work of Fukui (1986). In the discussions that follow, I seek to avoid confusion that can come from the use of the term "NP" in expositions of the DP analysis of the nominal phrase. Following Abney, I never use "NP" as an abbreviation for "noun phrase"; I refrain from using the term "NP" to refer to anything except the maximal projection of N. The structures under investigation are referred to in general as "nominal phrases".

2.2.1. Lexical versus functional categories

A distinction important to the discussion here is that of *functional* versus *lexical* categories. In 1824, Japanese linguist Akira Suzuki distinguished elegantly between these two types of elements by describing lexical or "thematic" elements, *zi*, as "denot[ing] something" and functional elements, *si*, as "denot[ing] nothing; [they] only attach 'voice of heart'" (quoted in Makino, 1968, p. 12). Lexical categories are described by the features +/–N and +/– V. These categories are: N(oun), V(erb), A(djective), and P(reposition) and encompass the vast majority of the lexicon of a language. Lexical heads are open-class items with phonetic content at S-structure that can optionally take arguments. In contrast, functional categories, C(omplementizer), I(nflection), and D(eterminer), constitute small closed classes. Members of functional categories are generally phonologically and morphologically dependent. They are generally stressless, often clitics or affixes, and sometimes even phonologically null. According to Abney (1987, p. 65), "Functional elements lack... 'descriptive content'. Their semantic contribution is second order, regulating or contributing to the interpretation of their complement. They mark grammatical or relational features, rather than picking out a class of objects."

Functional categories obligatorily select only one and only one complement[5] which is not plausibly either an argument or an adjunct of the functional element and which is inseparable from it. Lexical elements, on the other hand, may select more than one argument as complement. Unlike functional heads, a lexical head is a conceptually distinct object from the complement that it selects.

5 According to Abney's (1987) analysis, although D generally requires a complement (analogous to *if* + S, *of* + NP, *and* + various), there are some that don't, e.g., *That is on the table.* Still, the Abney analysis has certain advantages over previous analyses that required an empty noun phrase head or N-bar deletion. These cases are not discussed here.

2.2.2. DP analysis of the English nominal phrase: Abney (1987)

Abney was not the first to propose the functional category D(eterminer) as the head of the nominal phrase[6], yet the work of Abney constitutes the most thorough and comprehensive examination of this proposal to date. In his 1987 dissertation, Abney seeks to simplify and unify existing syntactic theory by analysing nominal phrases using the same structures and principles as those currently used to analyse sentences and verb phrases. By analogy to C(omplementizer), the functional head of the sentence, and I(nfl), the functional head of the verb phrase, Abney posits "D" as the functional head of the nominal phrase. Thus, under his analysis, the nominal phrase is now referred to as a "determiner phrase", or "DP", just as the sentence and verb phrase are represented in current theory as an "CP" and "IP" respectively. I refer to this analysis hereafter as the *DP analysis*.

There are many advantages to the DP analysis of nominal phrases. In designating D as the functional head of the nominal phrase, Abney relates the elements appearing under D to other functional elements, specifically complementizers. As with other functional heads, D can also be the site for AGR(eement). The SPEC(ifier position) of functional categories, including D, unlike the lexical categories, provides a position for overt case-marked Subjects as well as serving as the landing site for various types of movement, thus accounting both for various agreement phenomena and also providing "room" for the full range of specifiers found in the nominal phrase. A major contribution of this analysis is to preserve the restrictive characterization of X-bar theory

6 Previous Det-as-head proposals include those of Montague (1974), Hale (1982), Higginbotham & May (1980), Barwise & Cooper (1981), Brame (1982), Szabolcsi (1981, 1984), Hudson (1984), Kornfilt (1984, 1991), Horrocks & Stavrou (1985), Reuland (1985), Fukui (1986), Fukui & Speas (1986), Hellan (1986), Kuroda (1988), and Bowers (1987).

which has motivated the CP/IP analysis of the sentence while accounting for many nominal phrase phenomena without having to resort to various nominal phrase-specific constraints and mechanisms. In the following sections, I rely on Abney's DP analysis in making predictions of grammaticality for each of the sentence example types in the three modification categories under examination.

2.2.2.1. DP analysis: MOD–NP

As described by Abney's DP analysis, the structure shown in Figure 2–3 accounts for the judgments on sentences in Section 2.1.1. above. D is the syntactic head of the phrase; NP is its obligatory complement.

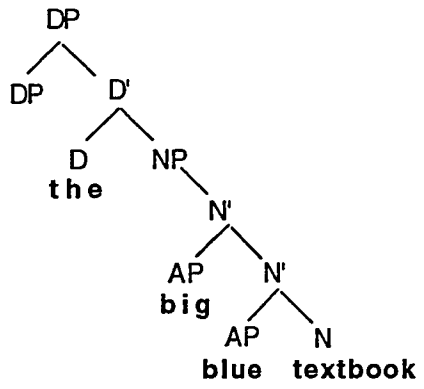


Figure 2–3: Analysis of NP–MOD (1a.) (after Abney, 1987)

The figure describes the only grammatical example in the set, 1a.[7] Examples *1b. and *1c. in which the determiner intervenes between the adjectives or between the adjectives and the noun is disallowed because in a head-initial language such as English, there is no other possible position for the determiner, *the,* apart from at

[7] Abney actually analyses adjective phrases as "Deg(ree)P(hrase)s" in order to capture the full range of English specifier structures; however, for simplicity's sake, I refer to the maximal projection of adjectives as "AP", which is adequate for my purposes here.

the front of the nominal phrase; unlike Korean, in English there is no landing site available for scrambled elements. The same is true for example *1d. in which the determiner appears phrase-final.

2.2.2.2. DP analysis: 2ARG–NP

Figure 2–4 shows one possible structure which describes a nominal phrase in which the semantic head noun subcategorizes one internal and one external argument. The sentences in Section 2.1.2. above exemplify these structures.

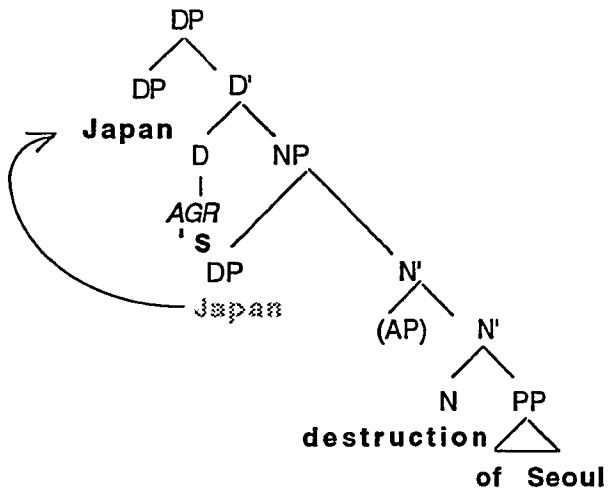


Figure 2–4: Analysis of 2ARG–MOD (2c.) (after Abney, 1987)

The figure describes the grammatical example 2c., *Japan's destruction of Seoul.* The internal argument, *Seoul,* is assigned the θ-role of Patient by the semantic head *destruction* and appears as a PP complement; partitive case (Abney's term) is assigned by the head of the PP, *of. Japan,* as the external argument, or subject, of *destruction* is base-generated in the SPEC position of NP, analogous to the generation of the subject of a sentence in the SPEC position of IP. However,

although the noun *destruction* assigns the θ-role of Agent to the DP (in earlier versions, NP) *Japan*, it cannot assign case. The DP, *Japan*, must move to a case-assigned position, namely, the SPEC of DP. Genitive case is assigned by AGR in D. Genitive case marking is realized as *'s* on *Japan's*. The case of sentence 2d., *Seoul's destruction by Japan*, is similar. In this case of a passive construction, *Seoul* is the external argument, or subject, of the semantic head *destruction*; the Agent, *Japan*, appears in the *by*-phrase complement position. As above, the external argument must move to the SPEC position of DP to receive genitive case, which is assigned by AGR in D. In the third grammatical case, *the destruction of Seoul by Japan*, D is already occupied by the determiner *the*, thus, both arguments appear as PPs.

There can only be a single head, D, of a nominal phrase. AGR may occupy this position and assign genitive case to an external argument which has moved into the SPEC of DP. The examples in *2a. and *2b show two genitive-marked arguments and thus violate the structure. *2b. is also ungrammatical on the grounds that D, as head of the nominal phrase can accommodate only one determiner-like element, either a genitive-case-assigner (AGR) or a determiner, such as *the*. Example *2f. is ungrammatical not only on the same grounds as *2a., but also because of the existence of a phrase-final determiner.

2.2.2.3. DP analysis: 3ARG–NP

The case for the 3ARG–NP structures is slightly different from that of the 2ARG–NP structures. In the ungrammatical 2ARG–NP examples, additional preceding arguments are genitive-marked by *'s* thus constituting an illegal multiple-headed construction. In the 3ARG–NP examples *3a. and *3b, and *3f., additional arguments in the form of PPs that precede the semantic head, *gift*,

also constitute structural violations. In English, a head-initial language, complements of a noun within an NP must follow the head N; the ungrammaticality of a PP complement preceding the noun this accounts for *3a. and *3b, and *3f.

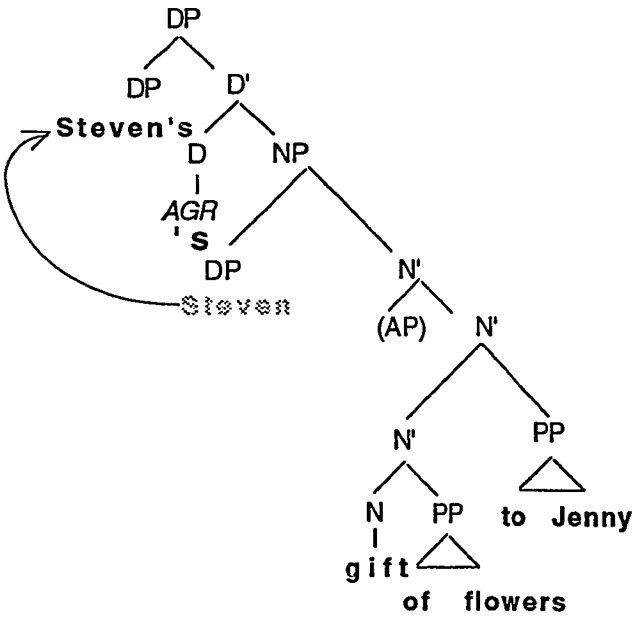


Figure 2–5: Analysis of 3ARG–MOD (3c.)

In the grammatical examples 3c. (Figure 2–5) through 3e. a single determiner-like element occupies the head of the nominal phase, whether it be the determiner, *the*, or a nominal argument which then raises to the SPEC of DP to receive genitive case; arguments occurring as PP complements follow. Example *3f. is ungrammatical not only on the same grounds as *3a. and *3b., but also because of the phrase-final determiner.

2.2.3. DelP analysis of the Korean nominal phrase: Yoon (1990)

In his 1990 dissertation, Yoon takes as the starting point for his analysis of Korean clausal and nominal structures Abney's proposal that in English (and perhaps universally) these are headed by functional categories. Yoon's work is based on earlier partial treatments, i.e., topic vs. subject prominence (Li and Thompson , 1976), configurational vs. non-configurational (Hale, 1982, 1985), and parameterization according to absence/presence of functional categories (Fukui, 1986; Fukui & Speas, 1986; Ahn, 1988; Ahn & Yoon, 1989). Yoon claims that the major syntactic building blocks of Korean can likewise be represented as maximally parallel to each other. This has the advantage in Korean, as in English, of eliminating the need for various nominal phrase-specific mechanisms. Additionally, Yoon endeavors, with one unified theory, to account for scrambling in both nominal phrases and clauses.

Yoon proposes, as does Abney in English, that in Korean, external arguments are assigned case in the specifier position of a functional category which contains AGR(reement): i.e., SPEC of AGR in clauses and SPEC of D in nominal phrases. The functional categories containing AGR are in turn dominated by another functional category, e.g., IP in case of Agr — the same as in English. In order to account for scrambling in nominal phrases, Yoon posits an additional functional category, Del(imiter)[8] which takes a DP as its obligatory complement. Specifier positions of I(nfl) and Del are reserved as landing sites for scrambled elements in clauses and nominal phrases respectively. The concern here is only with nominal phrases — it will be shown that the specifier position

8 Yoon posits this position to accommodate delimiters such as *man* 'only, *cocha* 'even', *mace* 'even, including', *kkaci* 'up to', etc. See discussion in Yoon 1990, also cf. Ahn (1988, 1989) and Yang (1972).

of Del provides a position for both post-adjective determiners and additional genitive-marked arguments in Korean.

2.2.3.1. DelP analysis: MOD–NP

Yoon (1992, p. 454, fn. 23), in a footnote, claims "in Korean there are no determiners", he refers to words such as *ku, i,* and *ce* as "demonstratives". Nevertheless, to simplify comparisons of English and Korean, I use "determiner" in the following discussions to refer to the class of words such as *a, the,* and *this* in English and words such as *ku, i,* and *ce* in Korean. These words have the following characteristics in common: (1) they uniquely identify a noun, (2) they belong to a small closed set of lexical items, and (3) they are limited to one occurrence per nominal phrase. The position of the determiner under DP in Korean differs from the position in English. In the English DP structure, as shown in Figure 2–3 above, the determiner *the* appears as the head, under D. In Korean, according to Yoon, a determiner such as *ku* is base-generated in the SPEC position of DP. This is necessary not only to account for various case-marking and agreement phenomena, but also to maintain the noun in the phrase-final (head) position[9]. Figure 2–6 illustrates the structures for the examples 1a'. through 1 c'.

Adjectives (*khun* 'big' and *phulun* 'blue') are identified in the above structures as A(djective) P(hrase)s, following Yoon (1990, p. 299). However, as non-attributive, or predicative, adjectives, these may also be analysed as null subject relative clauses. According to Yoon (1990, p. 327), "...only non-attributive adjectives are compatible with the tense morpheme and can be used as

---

9 Actually Del, as head of the nominal phrase occupies the phrase-final position. Del is realized as a delimiter, (see footnote 8). However, here we are not concerned with delimiter-marked nominal phrases.

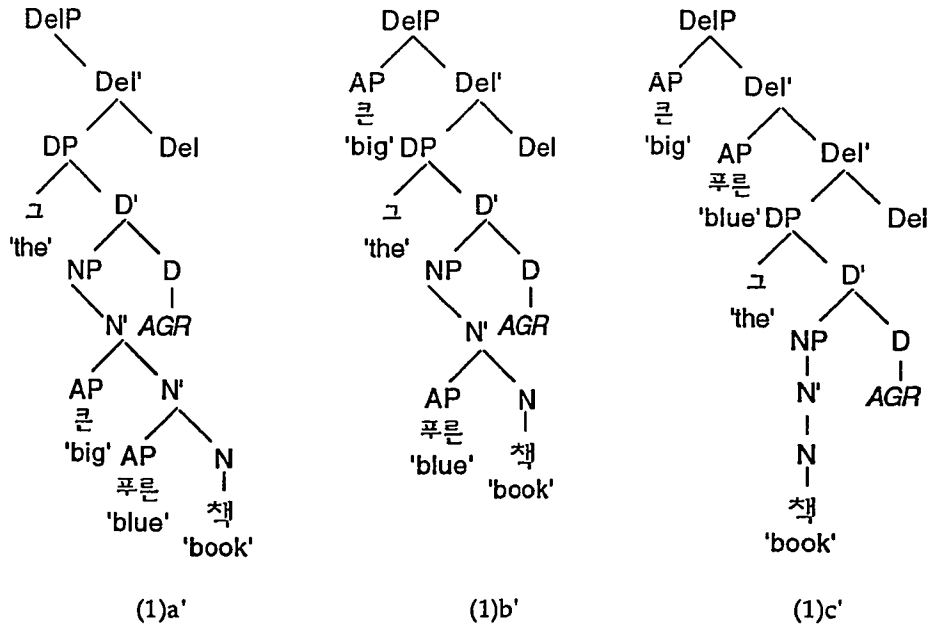


Figure 2–6: Analysis of NP–MOD (1a'. through 1 c'.) (after Yoon, 1990)

predicates. We propose that the pre-modifying phrases which precede a demonstrative [determiner] or a specifier DelP are relative clauses....non-attributive adjectives can be analysed either as APs or relative clauses depending on the occurrence of the tense morpheme." Yoon does not discuss the case of two pre-modifying phrases preceding the determiner; I assume that this can occur through Chomsky-adjunction of Del' as shown in Figure 2–6, 1c'. above.*1d'. is ungrammatical on the grounds of the illegal phrase-final determiner.

2.2.3.2. DelP analysis: 2ARG–NP

Yoon's analysis of the structure referred to here as 2ARG–NP in general follows that of Abney on the same structure. Figure 2–7 describes the grammatical example 2a'., repeated here:

(2) a'. ilpon-uy sewul-uy phakoy
Japan-GEN Seoul-GEN destruction

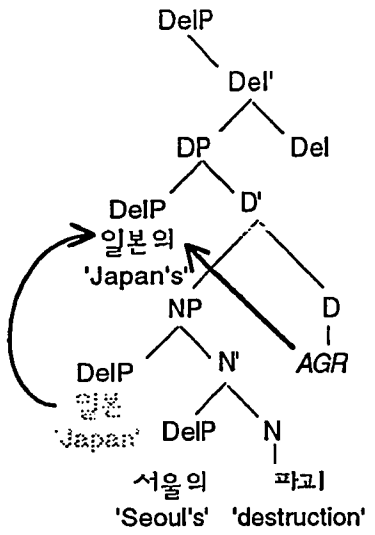


Figure 2–7: Analysis of 2ARG–MOD (2a'.) (after Yoon, 1990)

The internal argument, *sewul* 'Seoul', which appears as a complement of the head, *phakoy* 'destruction', is assigned the θ-role of Patient by the head, which also internally assigns genitive case. The external argument or subject of *phakoy* is *ilpon* 'Japan'. The external argument is base-generated in the SPEC position of NP, analogous to the generation of the subject of a sentence in the SPEC position of IP, as both Abney and Yoon point out. Although the head, *phakoy*, assigns the θ-role of Agent to the DP, *ilpon*, it cannot assign case; *ilpon* must move to a case-assigned position, namely, the SPEC of DP. Genitive case is assigned by AGR in D. Genitive case marking is realized as *–uy* on *ilpon-uy*, 'Japan's'.

The example in *2b'. is repeated here:

(2) b'. *ilpon-uy sewul-uy ku phakoy
Japan-GEN Seoul-GEN the destruction

This example is not grammatical according to Yoon for the following reason: the external argument, *ilpon*, base-generated in the SPEC of NP, must move to the SPEC of DP to receive case. However, SPEC of DP is also the position in which the determiner, *ku*, 'the', must be generated. These two elements, i.e., a genitive-marked DP external argument of the head and a determiner, cannot co-occur.[10] Examples *2c'. through *2f'. are all ungrammatical on the same grounds: as shown in Figure 2–7, the structure does not provide post-head positions for arguments of the head.

2.2.3.3. DelP analysis: 3ARG–NP

The analysis of the examples in 3a'. through *3f'. for the most part parallels that of examples 2a'. through *2f'. The 3ARG–NP sentences include an additional subcategorized internal argument. In addition to assigning the Agent role to *Steven*, the noun *senmul* 'gift' assigns the θ-roles of Theme and Recipient to *kkoch* 'flower(s)' and *Jenny*, respectively. The Recipient argument is marked morphologically for both dative and genitive case. Sentences such as 3a'., repeated here, are not specifically addressed in Yoon's analysis.

(3) a'. Steven-uy Jenny-ey-key-uy kkoch-uy senmul
Steven-GEN Jenny-DAT-GEN flowers-GEN gift

10 In an informal survey, I found that the examples in 2b' and 3b' are not considered ungrammatical by a number of native speakers of Korean. This issue is taken up again in Section 2.2.4.

Drawing on Yoon's work, I project the structure shown in Figure 2–8 to accommodate 3ARG–NP sentences:

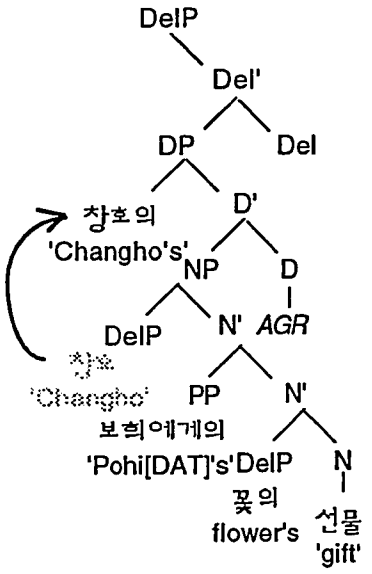


Figure 2–8: Analysis of 3ARG–MOD (3a'.)

2.2.4. An alternative analysis for Korean nominal phrases: Fukui (1986)

It is important to note that there is an alternative view of Korean nominal phrases which can be extracted by analogy to the analysis of Japanese phrase structure by Fukui (1986; also Fukui and Speas, 1986), supported (with reservations) by Radford (1990). Fukui divides languages into two types: the first type is represented by those that have both functional categories and lexical categories, like English, and the second, by those that lack functional categories, like Japanese, and by extension, Korean. Fukui's claim is that the presence or absence of functional categories in a language can account for many differences between two typologically distinct languages like English and Korean at all levels of syntax.

Under this analysis, a crucial way in which lexical categories differ fundamentally from functional categories is in that only the former are iterable, that is, only lexical categories allow "free recursion" of specifiers at the X' level as long as all the specifiers are licensed and can be interpreted at LF. The head of a functional category, on the other hand, selects a unique specifier which "closes off" any further projection at either the X' or XP level (DP, IP, and CP). In the case of the nominal phrase in either language, the iteration of specifiers at the N' level provides places for multiple modifiers of the head noun; these appear under N' as shown in Figure 2–9 (a. and b.).

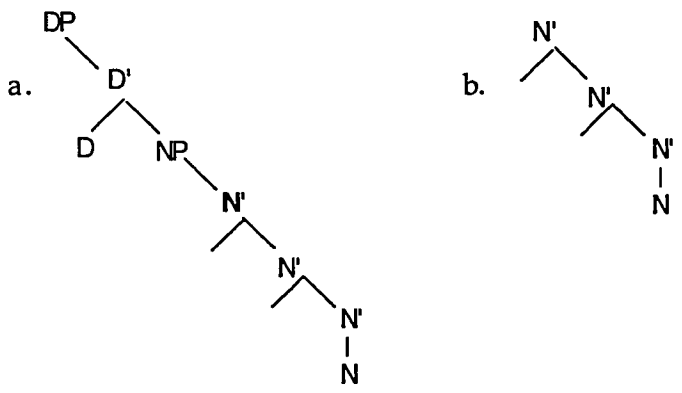


Figure 2–9: Fukui's (1986) analysis of a. English nominal phrases and b. Japanese nominal phrases

Similar to Abney's analysis, in Fukui's analysis, English determiners, as functional categories, act as the head of the nominal phrase; the noun is the obligatory complement of D(eterminer). This has the effect of placing the head of the nominal phrase on the left as is to be expected in a right-branching language (Figure 2–9a.) and also limiting the occurrence of specifiers to one per nominal phrase. In Japanese, a left-branching language which lacks functional categories,

the head of the nominal phrase, the lexical category N, appears in the rightmost position (Figure 2–9b.). Determiners in Japanese distribute with other modifiers, thus they are of the type of adjunct that freely iterates. Presumably an additional, possibly semantic, constraint is required to limit the number of determiners per nominal phase to one. Thus, Fukui claims that as neither genitives nor determiners close off the projection of N (p. 236); they are not heads. Thus, in a language like Korean there is no need to posit the existence of a functional category D; the head of the nominal phrase N' is N.

The decision to use Yoon's analysis over that of Fukui's in the present study is basically a pragmatic one. Overall, either analysis seems a reasonable one to adopt for the purpose of comparing the results of three different grammaticality evaluation methods. With one exception, both Fukui and Yoon make identical predictions about the grammaticality of the examples under investigation. This exception is that, in contrast to Yoon's analysis, Fukui's analysis (p. 230–236) allows for the co-occurrence of a determiner with one or more genitive-marked argument nominals as modifiers of the head noun (e.g., Fukui would accept as grammatical 2b'. and 3b'. above). I will be returning to this exception in depth in Chapter 6, Section 6.2. The selection of Yoon's analysis over Fukui's is based on the fact that Yoon's analysis, with a few small exceptions, explicitly accounts for each of the nominal types in the modification categories examined, MOD–NP, 2ARG–NP, and 3ARG–NP. As a bonus to the present investigation, it is possible that the data collected with respect to the exceptional case will indicate the validity of one analysis over the other; however, the main purpose of this study is to compare grammaticality evaluation methods, not competing theories of nominal phrase structure.

2.3. Another way to look at the data

All other things being equal, an ungrammatical string should be more difficult to parse than a grammatical one. This observation constitutes at least part of the original motivation for such grammaticality evaluation measures as elicited imitation and reaction time — these two measures are discussed in detail in the Chapter 4. Ungrammaticality can be viewed as introducing additional complexity to the task. There are other factors unrelated to grammaticality that can confound the task — these are also discussed in following chapters. Before leaving this chapter, however, I present here another way to analyse the data in terms of one particular confounding factor, namely, the length of the input as calculated in "chunks".

2.3.1. Chunking: Abney (1991)

Abney, in a 1991 article "Parsing by Chunks", outlines a design for a parser which operates in two steps; the first step is *chunking*, the second is *attachment*. *Chunking* captures the intuition that the determination of what constitutes a phrase, or "chunk", corresponds to an interaction among such elements as intonation, stress, and grammatical structure. *Attachment* refers to the relationships among phrases in a sentence, i.e., it describes the way in which chunks are assembled to form a completed tree structure for the sentence. Abney's description of this method of parsing is useful for the present study as an alternative way to look at the input sentences, that is, other than in terms of their grammaticality. The application of his parsing strategy to the input sentences may help to account for some of the variation among the data collected in this study via three different instruments.

Abney's formula for determining what is a chunk is an extension of earlier work by Gee and Grosjean (1983) who examined what they referred to as *performance structures* via a variety of experimental data which included pause duration in reading and naïve sentence diagramming. Abney found their unit of chunking, the *ϕ–phrase,* while basically sound, inadequate for his purposes in that it lacked syntactic structure. Abney's formula for determining a chunk, simplified somewhat for my purposes here, is as follows. Chunks are defined in terms of *major heads*. Major heads are all content words (i.e., members of lexical categories[11]) found in the input string. The exception is any content word which intervenes between a functional head (such as D) and the content word which it *f*-selects[12] (i.e., the N head of the obligatory NP complement of D).Therefore, in the phrase *the big blue textbook,* even though the adjectives *big* and *blue* are content words, because they intervene between the functional head of the DP, D (*the*), and the content word which it *f*-selects (*textbook*), these adjectives are not considered to be major heads. Thus, in Abney's system, *the big blue textbook* comprises one chunk. The sentence *Mary found the big blue textbook under the desk* totals four chunks (Figure 2–10)[13].

---

11 Prepositions, as functional heads of the prepositional phrase (PP), are not considered to be content words.

12 *F-selection*, the relation between a functional head and its complement, according to Abney "corresponds semantically to the 'passing on' of the descriptive content of the complement" (1987:55). For example, in a typical noun phrase, the functional head D f-selects its complement NP which passes up its semantic content to DP. Thus, although Det intuitively is an odd choice for the head of the noun phrase, this designation is a little easier to accept if we see Det as an essentially non-referential element that nevertheless has a very important job — it does the critical work of supplying information about definitiveness, number (e.g. English), gender (e.g., Spanish), and Case (e.g., German), for example, and passing this information, along with the semantic information supplied by its *f*-selected complement NP on to DP.

13 William O'Grady (personal communication) points out that Pritchett (1988) and Fodor (1978) might object to this system, claiming that it cannot account for relatively short (in terms of chunks) garden path sentences which nevertheless are difficult to process. I suggest that in the case of garden path sentences, processing is confounded due to ambiguity in the assignment of elements to membership in a particular chunk.

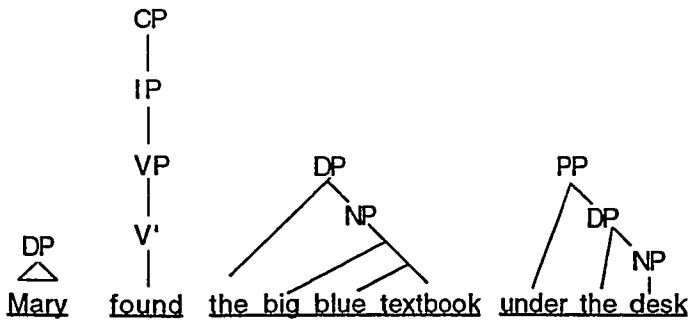


Figure 2–10: Sentence with grammatical MOD–NP, chunked

A problem arises, however, when an attempt is made to apply chunking to ungrammatical nominal phrases which include adjectives. Unfortunately, as in the case with most parsing mechanisms, Abney's parser is designed to handle only grammatical input. He supplies no specifics to suggest the analysis of the ungrammatical examples used in this study. To take an example, it is unclear how Abney would chunk the following nominal phrases (1d. and d'., repeated from above), taken from the set of input data.

(1) d. * big blue book the d'. * khun phulun chayk ku
big blue book the

In English, although it seems intuitive that *big blue book* constitutes a chunk and *the* constitutes another, Abney could object to this chunking on the grounds that the absence of a determiner-head of the DP "frees up" all three of the words *big, blue* , and *book* to represent major heads, that is, the adjectives *big* and *blue* are no longer intervening between the functional head (at least an overt functional head) and its *f*-selected content word, thus a count of four chunks is assigned to this nominal phrase. Alternatively, one might argue that a trace of a missing

determiner remains to bundle the phrase *big blue book* as one chunk (cf. *big blue books*, which surely would constitute a single chunk). Certainly, in Korean, *khun phulun chayk* 'big blue book' is a perfectly acceptable nominal phrase even without an overt determiner and would constitute a single chunk. Regrettably, too much of the analysis is left unresolved with respect to the MOD–NP examples to apply it to these; I return to these particular cases in the analysis of the data in Chapter 6, Section 6.3. The cases of the remaining input sentences (2ARG–NP and 3ARG–NP), however, are less ambiguous for both grammatical and ungrammatical examples, and in fact, Abney's chunking schema can be a valuable tool for analysing the input in such a way as to account for results of the data collection that are not explainable in terms of grammaticality value alone.

In Figure 2–11, a genitive noun, *Japan's*, modifies the head noun, *destruction*, which also takes a subcategorized PP, *of Seoul*. All three of these contain major heads which do not intervene between a functional head and its *f*-selected content word and therefore constitute separate chunks. In Figure 2–12, there are two genitive nouns, *Japan's* and *Seoul's*, modifying the noun, *destruction*, producing an ungrammatical nominal phrase; nevertheless, at the chunking level this phrase likewise consists of three chunks.

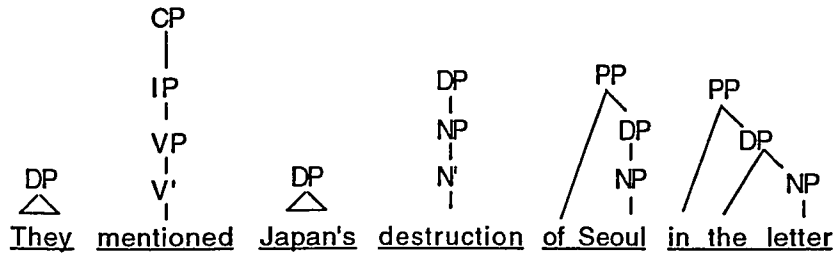


Figure 2–11: Sentence with grammatical English 2ARG–NP, chunked

Figure 2–12: Sentence with ungrammatical 2ARG–NP, chunked

The case of the Korean counterpart examples is similar, even though the judgments of grammaticality are opposite. In Figures 2–13 and 2–14, both arguments of the semantic head *phakoy* 'destruction' are expressed as genitive nouns, i.e., *ilpon-uy* 'Japan's' and *sewul-uy* 'Seoul's' (the internal structure has been simplified here). As above, all three of these contain non-intervening major heads, and therefore constitute separate chunks, regardless of their ordering.

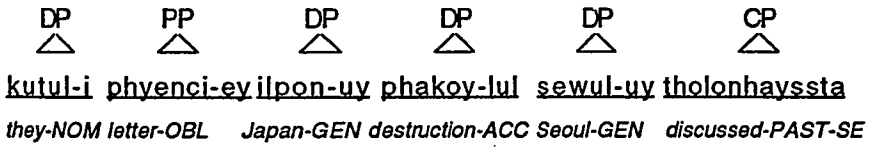


Figure 2–13: Sentence with ungrammatical Korean counterpart to Figure 2–11, chunked

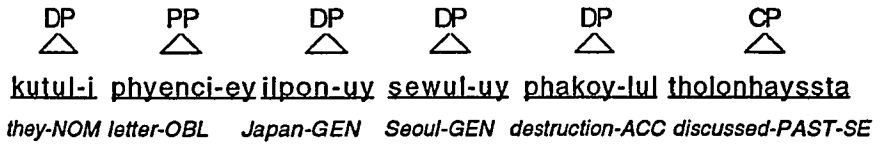


Figure 2–14: Sentence with grammatical Korean counterpart to Figure 2–12, chunked

The subject of the sentences in Figures 2–11 and 2–12, *they*, brings up the issue of the status of English pronouns. While Gee and Grosjean "sweep" them into adjacent chunks at the level of ϕ–phrases, Abney considers them to be "lexical noun phrases, and assign[s] them them the same status as orphaned words. At the level of chunks, they are orphaned words, belonging to no chunk" (1991, p. 260). For my purposes here, subject pronouns, as DPs, are considered as separate chunks. The total length in chunks for both the sentence in Figure 2–11, *They mentioned Japan's destruction of Seoul in the letter* and the sentence in Figure 2–12, **They mentioned Japan's Seoul 's destruction in the letter* is six, thus, at the chunking level, the grammatical and ungrammatical versions are equal in terms of numbers of chunks, the same is true for the Korean counterpart sentences in Figures 2–13 and 2–14. There exists a single ungrammatical version of this sentence in each language, i.e., *They / mentioned / Japan's / Seoul's / destruction / the / in the letter* (=7); *kutul-i / phyenci / ilpon-uy / sewul-uy / phakoy-lul / ku / iyakihayssta* (=7), which, because of the displacement of the determiner to the final position in the nominal phrase, is "longer" by one chunk than the above examples. A similar case can be made for sentences containing the 3ARG–NP examples. Because of the extra argument, each sentence is longer by one chunk. A summary of chunk assignment is shown in Table 2–1 below.

2.3.2. Attachment

The second step in Abney's parsing process is that of *attachment*, whereby attachment ambiguities are resolved and completed sentence-level trees are built. The completed parses of the sentences in Figures 2–11 and 2–12 are shown here in Figures 2–15 and 2–16.

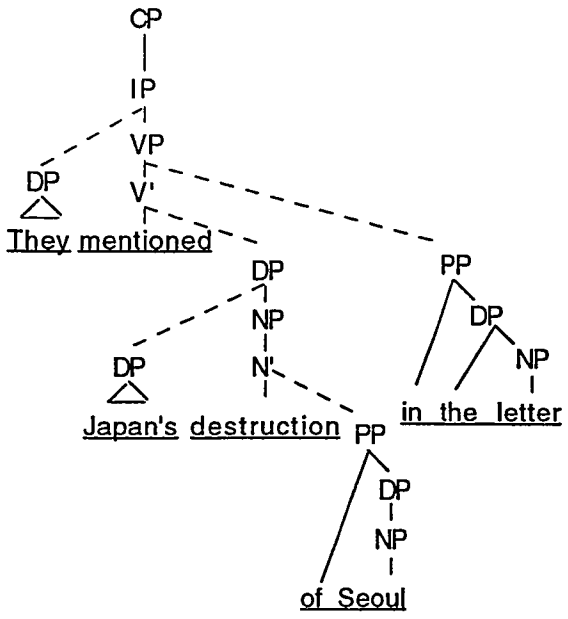


Figure 2–15: Sentence with grammatical 2ARG –NP, chunked and attached

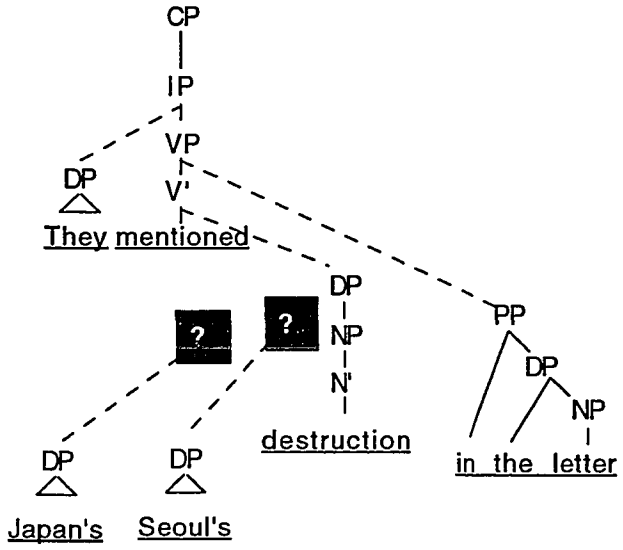


Figure 2–16: Sentence with ungrammatical 2ARG –NP, attachment FAILS

As pointed out above, Abney did not mention the issue of ungrammatical input in his account of the parser. I assume that in the case of ungrammatical input, as shown in Figure 2–16, complete attachment simply fails to occur. Possibly, one of the two DP candidates for attachment to the SPEC position of the DP, *destruction*, attaches leaving the other without an attachment site; alternatively, neither attaches. Although the precise manner in which this failure manifests itself is unclear, it seems reasonable to claim that complete attachment does indeed fail. Attachment is not pursued further here since, as a method of analysing the input sentences, it is completely redundant with the previous determinations of grammaticality predicted by the syntactic analyses.

2.3.3. Applying Abney's chunking to the input data

Based on the above discussion, I have assigned to each of the input sentence example types a number which indicates the length in chunks according to Abney's chunking scheme. The success or failure of attachment need not be coded as this information is completely congruent with the designation of "grammatical" or "ungrammatical" already assigned to the sentences based on the predictions of the syntactic analyses. Table 2–1 summarizes the sentence example types used in this study (see analysis in Chapter 5) and their length in chunks for the examples including 2ARG–NP and 3ARG–NP nominal phrases. The sentences of uncertain length in the MOD–NP category are analysed separately in Chapter 6, Section 6.3. The composition of the nominal phrases are indicated; the construction of the example sentences is covered in detail in Chapter 4, Section 4.1.1. English and Korean realizations of each sentence example type are shown. Sentence frames are held constant in terms of

grammaticality and length in chunks; ungrammaticality in a sentence always arises within the nominal phrase. A determination of ungrammaticality for a particular example is indicated by an asterisk preceding the length of chunks.

Table 2–1: Sentence example types — grammaticality and length in chunks

Sentence example type	*length in chunks*	
	English	*Korean*
2ARG–NP		
ARG NOUN ARG	6	*6
They mentioned Japan's destruction of Seoul in the letter.		
kutul-i phyenci-ey ilpon-uy phakoy-lul sewul-uy iyakihayssta.		
they-NOM letter-LOC Japan-GEN destruction-ACC Seoul-GEN talk about-PST-SE		
ARG NOUN ARG	6	*6
They mentioned Seoul's destruction by Japan in the letter.		
kutul-i phyenci-ey sewul-uy phakoy-lul ilpon-uy iyakihayssta.		
ARG ARG NOUN	*6	6
They mentioned Japan's Seoul's destruction in the letter.		
kutul-i phyenci-ey ilpon-uy sewul-uy phakoy-lul iyakihayssta.		
ARG ARG DET NOUN	*6	*6
They mentioned Japan's Seoul's the destruction in the letter.		
kutul-i phyenci-ey ilpon-uy sewul-uy ku phakoy-lul iyakihayssta.		
ARG ARG NOUN DET	*7	*7
They mentioned Japan's Seoul's destruction the in the letter.		
kutul-i phyenci-ey ilpon-uy sewul-uy phakoy-lul ku iyakihayssta.		
3ARG–NP		
ARG NOUN ARG ARG	7	*7
We described Steven's gift of flowers to Jenny to the group.		
wuli-ka salamtul-eykey chelswu-uy senmul-ul pohuy-eykey-uy kkoch-uy kiswulhayssta.		
we-NOM group-DAT [name]-GEN gift-ACC [name]-GEN flower-GEN describe-PST-SE		

DET NOUN ARG ARG ARG 7 *7

We described the gift of flowers from Steven to Jenny to the group.

wuli-ka salamtul-eykey senmul-ul chelswu-uy pohuy-eykey-uy kkoch-uy kiswulhayssta.

ARG ARG ARG NOUN *7 7

We described Steven's to Jenny of flowers gift to the group.

wuli-ka salamtul-eykey chelswu-uy pohuy-eykey-uy kkoch-uy senmul-ul kiswulhayssta.

ARG ARG ARG DET NOUN *7 *7

We described Steven's to Jenny of flowers the gift to the group.

wuli-ka salamtul-eykey chelswu-uy pohuy-eykey-uy kkoch-uy ku senmul-ul kiswulhayssta.

ARG ARG ARG NOUN DET *8 *8

We described Steven's to Jenny of flowers gift the to the group.

wuli-ka salamtul-eykey chelswu-uy pohuy-eykey-uy kkoch-uy senmul-ul ku kiswulhayssta.

* asterisk indicates ungrammaticality

2.4. Summary of modified nominal phrases in English and Korean

In this chapter I presented a description of the three categories of nominal phrases in both English and Korean. These were analysed both in terms of their grammaticality and in terms of their length in chunks. Example types from these nominal phrase categories constitute the input to the present study. As mentioned above, the main purpose of this study is to compare grammaticality evaluation methods, not competing theories of nominal phrase structure. The concept of grammaticality evaluation as a subset of metalinguistic performance is introduced in the following chapter; Chapter 4 details each of the methods under investigation.

CHAPTER 3: GRAMMATICALITY EVALUATIONS

3.1. Broad definition: Metalinguistic performance

Metalinguistic performance may be defined in its broadest sense as "any objectification of language" (Birdsong, 1989, p. 1). The range of metalinguistic performance is extensive, entailing not only the type of observation and abstract analysis of language that linguists engage in, but also more everyday enterprises, for example, conversation management, i.e., the ability to recognize communication breakdown resulting from inappropriate word choice, odd pronunciation, or structural ambiguity and consequently to repair utterances. Without the ability to step outside of language and its ordinary communicative function, we would miss out on the enjoyment of puns, word games, and poetry. All language users possess metalinguistic performance skills to some degree and it is upon this critical capacity that many linguists depend for supporting evidence of their claims.

For my purposes here, I develop and examine and discuss a narrow subset of metalinguistic behavior: grammaticality evaluation. First, as background, a general discussion of metalinguistic performance in the first language (L1), its development, and its variability across both individuals and groups is in order.

3.1.1. In children's developing L1

It is undisputed that children in general possess metalinguistic abilities that allow them to analyse both the language they hear and and the language they produce with respect to its congruence with the system of adult grammar they are acquiring. The end product of the acquisition process, the adult grammar of the first language, is widely assumed to be essentially equivalent among adult users. Bley-Vroman and Masterson (1989) refer to this as the *Uniformity of Resultant Systems* (from a phrase of Chomsky's (1965, p. 58)). The challenge to linguists is the characterization of these systems both for individual languages and for language in general. As Pinker (1984) proposes, "A theory of grammar for particular languages must be embedded in a theory of Universal Grammar that allows only one grammar...to be compatible with the sorts of sentences children hear." Thus, not only is the grammar of a particular language uniform across users of that language, but also there is a hypothetical set of principles that unify all languages. It does not seem unreasonable to suppose that inherent in these systems, both at the level of individual languages and at the level of Universal Grammar, mechanisms exist that enable metalinguistic awareness, which in turn facilitates the acquisition of language.

It is fairly well-accepted that such awareness acts as catalyst or at least has some supporting role in the acquisition of the adult grammar. Yet, claims of the precedence of the development of metalinguistic awareness over language acquisition appear circular: if a language user is aware of the acceptability of a particular language structure, then the evaluation based on this awareness is a reflection of previous knowledge of the structure rather than a guide to the acquisition of it. Furthermore, there is disagreement as to how crucial is the role

of these abilities (for example, see Marshall & Morton, 1978; Karmiloff-Smith, 1979; Pinker, 1989). Also apparent in the literature is the variability with which children display these skills. Reports from McNeill (1966) and Braine (1971), among others, cite examples of children who appear unable to perceive or repair deviant forms, even when such forms are the focus of the adult input. However, there are numerous studies (see Birdsong, 1989, p. 16, Table 1.1) recording children's relatively sophisticated skills in collecting, assessing, and rejecting or repairing instances of language. These skills appear fairly early, for example, deVilliers and deVilliers (1974) found that by the age of five, children are capable of recognizing and correcting the deviant word order within principal constituents, the focus of the present study. As children at this age are not literate, it seems reasonable to assume that literacy is not a necessary condition for the development of such skills.

3.1.2. In adult L1

Although all adults appear to possess metalinguistic abilities, the degree to which they do varies greatly. One of the most salient factors that can account for this variation is literacy. Studies done comparing literate and illiterate adults on such tasks as sensitivity to rhyme (Morias, Cary, Algeria, & Bertelson, 1979), speech segmentation (Read, Yun-Fei, Hong-Yin, & Bao-Qing, 1986), and phoneme addition and deletion (Morias, et al., 1979), among others all point to literacy as the critical element in accounting for performance differentials. This is not to claim, however, that there is an absolute causal link between literacy and the ability to observe and manipulate language nor between literacy and the ability to acquire language. As pointed out above, it is obvious that literacy is not a prerequisite to a child's acquisition of her first language; this holds true for an

adult's acquisition of a second language, as well. Scribner and Cole (1981) cite the example of Vai (Liberia) multilinguals, successful learners of more than one language, who are nevertheless unable to make most basic repairs to a sentence they recognize as deviant. Thus, possession of the type of metalinguistic skills associated with literacy in a native language is not necessarily a barometer of general success in learning the first language, nor a second: witness the low level of success evidenced by many university students, even after years of study.

It is apparent that a relationship exists between literacy and metalinguistic awareness; what is not clear is the nature of this relationship. Is literacy a prerequisite to certain kinds of metalinguistic awareness as indicated by the above studies, is the case vice versa as argued by Liberman, Shankweiler, Liberman, Fowler, and Fisher (1977), or does the answer lie somewhere in between (the "interactionist" position (Ehri & Wilce, 1980)? Though this remains an open question, it is indisputable that literacy exerts an influence on metalinguistic performance.

3.1.3. Biological component of language or learned skill?

Another question arising from a review of the literature is this one: to what extent is the ability to observe language innate and to what extent can performance be improved though instruction, familiarity, and practice? On the one hand, the early development of metalinguistic performance does not appear to be the result of training, formal or otherwise. Yet, many of the differences among adults are clearly linked to formal education. Thus, it seems metalinguistic ability is not a unitary faculty, but rather a "bundle" of many abilities, some developmental, some as a result of training.

A type of metalinguistic performance that is especially susceptible to improvement through training is what I refer to here by the general term "grammaticality evaluations". I expand on the definition of this term in Section 3.2.

3.1.4. Summary of metalinguistic performance

To summarize the discussion of metalinguistic performance in general, the ability to step outside of the use of language for communication and observe and manipulate it as an object in itself appears early on in a child's development of language and is present, to varying degrees in all adult language users. A major factor that appears to be closely linked with adult abilities to manipulate and evaluate language is that of literacy, though the question of whether such abilities are pre-requisite or artifact of literacy is not addressed here.

3.2. Narrow definition: Grammaticality evaluations

My focus here is on a subset of metalinguistic performance which I refer to by the general term *grammaticality evaluations*. These are most often referred to in the literature as "grammaticality judgments"; I choose here not to use the term "judgme for two reasons. First, I feel "judgment" implies a more conscious action than the term "evaluation". Although "evaluation" is not entirely neutral, it seems not so necessarily volitional. Second, I wish to use the term here to encompass all three of the instruments I examine, not to refer only to the written questionnaire. In the literature, the term "grammaticality judgment" is variously

applied to evaluation instruments in general and also to written questionnaires in particular in order to distinguish them from other types of instruments.

3.2.1. Direct versus indirect

In general, grammaticality evaluations take the form of participants' responses to instances of language with respect to grammaticality (or acceptability[1]). Grammaticality evaluations can be based on either implicit or explicit knowledge of the language. Bialystock (1981) cites this distinction when discussing sources for the formation of an evaluation stating that implicit knowledge is used to decide if a sentence is correct or incorrect (the typical grammaticality evaluation); explicit knowledge is accessed for further analysis required by introspection or production tasks. Her study showed that decisions made initially on an intuitive basis may or may not be supported by explicit knowledge. This suggests that reliance on explicit knowledge, often a product of formal education and therefore subject to being incorrect, could in theory reverse an evaluation.[2] It also raises one of the major issues of the present study, i.e., grammaticality evaluations as a faithful reflection of underlying competence.

I will be borrowing features of Bialystock's explicit/implicit distinction in categorizing types of grammaticality evaluations. I divide them into two main

1 Nunnally (1978:41) divides grammaticality judgments into two broad categories: judgments about grammaticality (theoretically there is one correct response) and sentiment (an expression of an attitude or opinion about use of language). This division is also often referred to in SLA literature, e.g., Chaudron (1983) discusses the distinction between judgment and sentiment.

2 Here is a classic example: A linguistics student was given a pencil-and-paper grammaticality evaluation task to complete at home. When he returned that completed test, he commented to the researcher that at first he wasn't so sure about his judgments, but once he realized that it was the Right Roof Constraint (Ritchie, 1978) that was the object of investigation, he was able to render what he felt were the "correct" judgments (Robert Bley-Vroman, personal communication). This situation not only points out the pitfalls of asking linguists for judgments, but also the need for distractor items.

categories: *direct* and *indirect* evaluations. These categories are related, though not congruent to Bialystock's explicit/implicit dichotomy. The terms *direct* and *indirect* are used here to indicate the extent to which the evaluation-giver attends to her evaluation of the grammaticality of an instance of language use. Both categories naturally presuppose implicit knowledge of the language in question.

Direct grammaticality evaluations are evaluations extracted by an experimental task which clearly focuses the attention of a language user upon consciously making a deliberate and considered evaluation and report of her feelings or intuitions about a structure in terms of its grammaticality or acceptability — variations on this theme are discussed further in Chapter 4, Section 4.2.1. The first of the three instruments employed in the present study, the written questionnaire, falls into this category.

On the other hand, *indirect grammaticality evaluations* are collected by employing experimental tasks which, though they may focus on language and the manipulation of it, nevertheless do not solicit direct assessments of the structure's grammaticality or acceptability. Such tasks may be in the form of production tasks, yet not require that the participant access explicit knowledge of a structure in order to complete the task. Indeed, in a task well-designed to gather indirect grammaticality evaluations, a participant is unaware of making evaluations at all. Examples of the use of indirect tasks include spontaneous conversation (Hilles, 1986), sentence-to-picture matching (Finer and Broselow, 1986; Shimura and Yoshino; 1988, Merzenich, 1989), and picture description (Parker and Chaudron, 1990). The remaining two instruments examined in depth in this study, elicited imitation and reaction time tasks, belong in the indirect category. The assumption here is that not only may these indirect instruments

provide cross-validation of the data collected via the direct task, but that they could prove useful in determining the uncertain cases, that is, the cases in which conscious introspection might tend to muddy the waters.

3.2.2. Performance as an indication of competence

An important step toward developing the explanatory adequacy of a linguistic theory is the characterization of of Chomsky's ideal speaker-hearer's underlying competence. Chomsky (1957, 1964, 1965) makes a distinction between grammaticality and acceptability and argues that these belong to different areas of study — *competence* and *performance* respectively. In the past, this distinction has been commonly misinterpreted when applied to the collection of evaluation data in that performance has erroneously been accepted as a faithful reflection of competence. Botha (1981), in his introduction to the methodology of generative grammar warns the researcher to examine carefully the data she collects:

> "...primary linguistic data should include only intuitive judgments which reflect the properties of linguistic competence...[these] represent genuine linguistic intuitions...Intuitive judgments which do not reflect the properties of linguistic competence represent spurious intuitions and should not be included in the corpus of primary linguistic data." (p. 70)

Similarly, Chaudron (1983, p. 343) interprets Chomsky's argument in this way: "linguistic theory must characterize the competence of a native speaker, and consequently, the speaker's competence must be used to test the theory." Chaudron does not say that the speaker's *performance* must be used to test the theory. Although the comments of Botha and Chaudron seem self-evident, it appears that all too often the latter interpretation is extracted, i.e., performance is

considered to be a direct "window on competence". Competence, by definition, is taken to be stable within the adult individual and essentially consistent throughout the L1 population, in other words, competence is the "resulting uniform system" of acquisition. In theory, grammaticality evaluations reflecting this competence would be likewise stable and consistent and could be used as measures by which to test the theory. In reality, however, grammaticality evaluations are instances of performance and therefore victim to all manner of interference, both psychological and physiological. They are notoriously unstable, as discussed below in Section 3.3.2. Nevertheless, the temptation to consider grammaticality evaluations as "a reflection of [a] learner's competence in the target language" (Arthur, 1980, see also Sharwood Smith, 1988) is strong. Ellis (1990, p. 4) points out that it is widely believed that grammaticality evaluations "...obviate the need for the learner to access the processing systems responsible for using the underlying grammar in actual performance. Grammaticality judgement tasks, therefore, are seen as providing a relatively direct window into competence." As a result of this temptation, the distinction between competence and performance frequently fails to be made. While competence is the object one may intend to examine in testing theory, it cannot be assumed that a language user's performance is necessarily a good reflection of competence.

Granted, as yet a way has not been found to tap directly into a language user's competence, still, there is a danger in the unexamined acceptance of the idea that a grammaticality evaluation, an instance of performance, can be equated with competence. A grammaticality evaluation as performance constitutes a cognitive act, and as such "must be removed from its object" (Itkonen, 1981). Gerkin and Bever (1966) focus on the cognitive aspect of

grammaticality evaluations, claiming that since grammaticality evaluations involve interaction between linguistic and cognitive systems they cannot be considered a faithful reflection of competence in isolation. Many studies have seriously questioned the reliability of data collected via grammaticality evaluation instruments, suggesting that grammaticality evaluations must be considered potentially unreliable as a self-contained measure; instead, they are more a reflection of the interaction between grammaticality, parsing processes, and other, non-language-related strategies.

Although the use of L1 baseline data is recognized as a critical ingredient in any program of research undertaken in the field of L2 acquisition, its value is often overlooked in first language studies; researchers often rely upon their own intuitions about the structures under investigation. Such intuitions are assumed to represent the language user's underlying competence, but the propriety of this biased practice is questionable. Intuition in this case, as in all cases of grammaticality evaluations, should not be mistaken for observation (Itkonen, 1981); unfortunately, in testing the adequacy of syntactic theory, it often is (but see Kuno, 1976; Mohan, 1977; and Carden & Dieterich, 1981, for exceptional examples). To illustrate this point: Spencer (1972) compares grammaticality evaluations collected from both linguists and non-linguists and finds that the non-linguists agreed among themselves 80% of the time but agreed with linguists' evaluations in only half the sentences. Bever (1970) and Labov (1975) find non-linguists as well may show great variability. Hill (1961) tested participants on sentences collected from Chomsky (1957) and found a great deal of variability among the participants and substantial disagreement with Chomsky's judgments.

It appears, then, that neither linguists nor non-linguists are particularly useful sources of grammaticality evaluations. How is a researcher to access the underlying system? Although it seems that users of a language are unreliable sources of data that might reveal features of the underlying system, one must not overlook the role of the data collection instrument itself. It is important here to recognize the distinction between a language user's having the facility to evaluate grammaticality and her facility in expressing that evaluation. A necessary condition for the evaluation of language is, of course, the underlying system. This, by itself, is not a sufficient condition. Even in the case that metalinguistic awareness is sufficiently developed for making grammaticality evaluations, the expression of this awareness requires both metalinguistic skills and a vehicle by which to convey them.

The first requirement, metalinguistic skills, like other cognitive skills, are highly variable. Inherent in the idea of the ability of a language user to make grammaticality evaluations, at least in the case of direct grammaticality evaluations, is the assumption that she is able to faithfully communicate her evaluations to another. Bley-Vroman and Masterson (1989) point out that the ability of adults to make evaluations is, in part, a trainable skill, but that the apparent inability to do so is not necessarily an argument against the Uniformity of Resultant Systems. Nevertheless, the facility for actually relaying such information — for this action certainly involves a separate step and hence a separate skill — may be overlooked in a discussion of the nature of grammaticality evaluations. It is important to keep in mind that at the point that a direct evaluation is expressed it is already two steps (the awareness to judge the

language plus the skill to express the evaluation) removed from the actual underlying competence.

The second requirement, the vehicle of expression, provides an opportunity for the introduction of all manner of noise into the experimental design through external environmental factors, internal design factors, and problems related to semantic context. Add to these factors the filter the researcher's interpretation of the data and it begins to look a bit chancy to accept grammaticality evaluations as an unpolluted reflection of linguistic competence.

Thus, there are two crucial elements to be taken into account in the use of grammaticality evaluations for getting a picture, though indistinct, of underlying structures and mechanisms. These are the participants who express the evaluations and the instrument though which the evaluations are expressed. Daunting as the management of these elements appears, a researcher equipped with a healthy scepticism may nevertheless find grammaticality evaluation instruments useful tools for getting a picture. Much current research, including the present study, is directed at "fine-tuning" that picture.

3.3. General methodological issues

3.3.1. Advantages to using grammaticality evaluations

3.3.1.1. Testing elusive structures

Snow and Meijer (1977) claim that while grammaticality evaluations (specifically syntactic as opposed to semantic intuitions) are an important source of information, they should be secondary to other types of language behavior.

They suggest that primary data are utterances and texts. While it is true that corpora represent actual, contextual language use, they tend to be extremely undirected and therefore a solely-corpus based approach is often impractical (but see a good use of a corpus in Bley-Vroman and Houng 1988). Despite the above-mentioned uncertainties, grammatical evaluations as a source of data have a number of practical advantages. For example, a researcher can target a specific and perhaps deviant or rarely occurring grammatical for investigation, rather than relying on it occurring naturally. It would be quite tedious, if not impossible, for example, to naturalistically collect a reasonable body of data on such objects of study as proper antecedent parameter (Shimura and Yoshino, 1988), governing category parameter (Otsu, 1981), Subjacency and Empty Category Principles (Bley-Vroman, Felix, & Ioup, 1988), adverb placement (White, 1989b), or head-direction parameter (Flynn & Espinal, 1985; Flynn, 1987). In addition, when referring to naturalistically collected data, the strongest claim a researcher can make for ungrammaticality is to claim that a structure is non-occurring. Non-occurrence is not particularly compelling evidence for ungrammaticality. Because by using grammaticality evaluations the researcher can zero in on a structure, the test instrument itself can be quite simple, often just a binary- or multiple-choice questionnaire. Such a questionnaire can be easily and systematically administered to large numbers of participants. Data collected from this type of questionnaire is homogeneous and lends itself well to complex statistical analysis and opens up the possibility for cross-validation with other methods as well as comparison across various studies.

3.3.1.2. The value of converging evidence

Any conclusion drawn from experimental evidence is significantly strengthened or weakened by the presence or absence of additional validating evidence collected through a different instrument. Such an instrument is designed to collect data about the same phenomenon, but by different means. The data for this study is collected using three different instruments — a written questionnaire, a reaction time task, and an elicited imitation task. If these measures are found to produce similar findings, these findings would reinforce the value of multiple measures as providers of converging evidence of the validity of the syntactic framework being tested. The value of converging evidence is widely recognized (Mohan, 1977; Singh, d'Anglejan and Carroll, 1982; Gass, 1983; among others). Chaudron (1983, p. 369) asserts, "It seems mandatory that, though they are indicators of participants' knowledge or values, [grammatical] judgements should be validated by other measures on the same or comparable items and subjects." Specifically referring to the use of an elicited imitation task in L2 research, Chaudron and Russell (1991, p. 12) remark, "What is needed are studies comparing participants' performances on elicited imitation tasks with performances on other tasks." And Bley-Vroman and Masterson (1989, p. 224) in a preliminary version of the present study find that "although both [written questionnaires and reaction time] methods are flawed, they are clearly flawed in different ways, so that the convergent results inspire a degree of confidence which neither set of results commands by itself."

3.3.2. Disadvantages to using grammaticality evaluations

There are also a number of practical disadvantages in addition to the metalinguistic disadvantages to grammaticality evaluations cited above. These can be divided into two types: those having to do with the design of a study which are, at least in theory, controllable and those having to do with the nature of dealing with human participants which ultimately are not.

3.3.2.1. Problems with task design

Environmental factors provide a morass of potential deterrents to the collection of noise-free data; these may include participants' fatigue, stress level, age, or state of physical fitness (Welford, 1980), the time of day[3], poor equipment, and many other types of other unforeseeable distractions. A unexpected example of the influence of environmental factors is found in a study done by Carroll, Bever, Pollack (1981) in which the researchers showed that grammaticality evaluations could be manipulated by altering the physical conditions under which they were presented. They found that the simple presence or absence of a mirror in the room where the participants were being tested affected grammaticality evaluations.

The design of the testing instrument itself is another area rife with pitfalls. Researchers have found that the outcome of a grammaticality evaluation task can be manipulated as a result of order and mode of presentation (Greenbaum, 1973; Carroll, Bever, Pollack, 1981), length (Quirk & Svartvik, 1966), complexity (ceiling and floor effects), the transparency of the task (see footnote 2), or even punctuation (Ellis, 1990) and spelling (Langendoen & Bever, 1973).

3 "Your short-term memory is said to be best at about 9 a.m. Long-term, at about 3 p.m. That's what the specialists report. Why memory varies in this way throughout the day remains unexplained." (Boyd 1992)

Problems that have to do with the administration of a testing instrument can include effects due to the participant's familiarity with the test administrator or inconsistencies arising from a testing situation in which more than one experimenter administers the test. Heringer (1970) observes that subtle clues from the researcher may led to self-fulfilling results, especially when the researcher has predicted a particular outcome.

Semantic interpretability is closely linked with participants' acceptance of language examples. Birdsong (1989) notes that the probability of a hypothetical event has bearing on its acceptability; similarly, Heringer (1970) finds that when "strange" sentences are given a context, evaluations of grammaticality increase greatly and concludes that a positive evaluation is linked to the participant's ability to find a context. Nagata's 1988 experiment shows that grammaticality evaluations may change with repetition. In Nagata's study, participants tended to judge more harshly the second time around — but only for sentences in isolation. Evaluations of sentences embedded in some context were more stable. Also, Levelt, Van Gent, Haans, and Meijers (1977) find "high imagery" content is more often judged to be grammatical.

3.3.2.2. Problems with participants

While the factors connected with the design of an experimental task mentioned in the previous section are difficult to control for, the problems explored in this section, those which are a function of having to use real people as participants, may appear insurmountable. Many have found participants' grammaticality evaluations to be unreliable (Snow, 1975), inconsistent across participants (Hill, 1961; Levelt, 1972; Greenbaum, 1977) and not stable on successive trials (Carroll, Bever, and Pollack, 1981; Birdsong, 1984; Nagata, 1988

(see above)). This is supported by Cowan and Abe (1991) and Ellis (1991); these studies collected inconsistent results on L2 participants via a test-retest procedures. Ellis (1991, p. 28), for example, notes the effect of repetition: "...when asked to judge identical sentences a second time, there is a good chance that the learners will respond differently." This is a possible example of where a participant, with more time and attention drawn to the sentence in question, applies Bialystock's explicit knowledge and thus reverses an evaluation originally based on implicit knowledge. In fact Lantolf, Goss, and Zhang reach this conclusion in their 1991 study of cognitive strategies employed by L2 learners. They assert that even if one is to find that learners' [and by extension L1 participants'] results on successive trials are consistent, this in itself is not sufficient to claim that their "judgments are predicated on linguistic principles rather than some other factors, such as consistent recall of memorized rules..." (p. 30).

Personality type and education level appear to be variables in the matter of grammaticality evaluations. A number of studies have shown that such factors as self-confidence and self-perception as "liberal" versus "conservative" (Ross, 1979) have an effect. Quirk and Svartvik (1966) find that grammaticality evaluations may vary according to a participant's college major — in the case of their study, English majors versus geography majors. This last finding brings up an interesting point. A high percentage of adult native speaker language studies are conducted within a highly homogeneous population, i.e., university students, many of whom are majoring in either psychology or language studies; very little has been done with other types of native speaker populations, non-literates, for example, who have been shown to exhibit vastly differing levels of metalinguistic awareness (Bialystock, 1986; Scholes and Willis, 1987a–c; Scribner and Cole,

1981). This doubtless has much to do with the availability of test populations; the present study is no exception.

Other interfering variables are closely related to the necessity of using participants from other cultures. Cultural biases may show up in the form of shyness, fear of failure, or anxiety about how the results are to be used. There could also conceivably be an effect due to perceived or promised payment or punishment as a result of performance. Members of a particular language or culture group may tend to employ similar strategies in answering questions about which they are uncertain. For example, Chinese students may have a tendency to accept sentences as grammatical more often due to a cultural bias and as a function of previous schooling (Oranit Limmaneeprasert, personal communication). Mohan (1977) identifies this as the "yea-saying factor". Such biases could significantly skew comparisons of participants of differing L1s.

It should be apparent by now that even if we accept the idea of grammaticality evaluations as a "window into competence", we have to admit that the view is considerably clouded by a great number of factors, some under our control, others not. The validity of grammaticality evaluations as a reflection of competence is filtered minimally through (a) the awareness to judge the language, (b) the skill to express the evaluation, (c) the testing environment, (d) the testing instrument, (e) inter- and intra-participant variation, and (f) the interpretation of the researcher. Birdsong (1989, p. 69) makes the following apt analogy, "Metalinguistic data are like 25-cent hot dogs: they contain meat, but a lot of other ingredients, too."

3.4. Summary of grammaticality evaluations

Clearly, grammaticality evaluations, as a type of metalinguistic performance, by no stretch of the imagination constitute a noise-free reflection of underlying competence. Nevertheless, they can be a useful tool for taking a peek through that "window on competence" if we remember not to expect an undistorted view. We need not discard grammaticality evaluation data out of hand if we can begin to understand the nature of that distortion. This, then is the focus of this study: to make and test predictions about the nature of data collected from native speakers of English and Korean, working with fairly uncontroversial input. Ideally, if these instruments are indeed measures of grammaticality, the data collected from them should be substantially similar.

Using three different instruments — a written questionnaire ("pencil-and-paper" or "P&P"), an elicited imitation task ("EI"), and a reaction time task ("RT") — I collected and examined data from 93 English and Korean participants about the status of the types of the nominal phrases described in Chapter 2 in the participants' internal representation of their native grammars. The three instruments are described in detail in Chapter 4.

CHAPTER 4: THE THREE INSTRUMENTS

4.1. Experimental design: Factors in common to all instruments

Before separately describing each of the three instruments, the written questionnaire ("P&P"), the elicited imitation task ("EI"), and the reaction time task ("RT"), I discuss of aspects of instrument design that are held in common by all three.

4.1.1. Input sentences

The input to all three tasks is in the form of sentences. I use the term *sentence* to refer to a string of words whose status as a grammatical sentence is at issue; the term *sentence* is not used to differentiate a grammatical string from an ungrammatical one. In creating the input sentences, every effort was made to guarantee that grammatically contrasting examples within a language were, as far as possible, equivalent in all other respects. Based on Eubank's (in press) findings (discussed below in Section 4.4.4. under "The input sentences") and on pilot studies of both the RT and EI tasks, I decided to make the input sentences fairly long. On all tasks the input was randomized to some degree (as described separately by task below).

Input sentences were generated according to a strict formula, controlling, when possible, for vocabulary and length, both in terms of syllables within a languages and length in chunks between languages. Whenever possible, the English and Korean sentences were near direct translations of each other. First names of people were alternated to reflect the language, for example, *Linda* in English became *Mica* in Korean. Cultural context was taken into account as well. Research has shown that an evaluation of grammaticality may often depend on a participant's ability to assign a context to it. In the group of input sentences used, for example, was a set involving the *destruction* of *Seoul* [Korea] by *Japan*. The use of this example was primarily motivated with Korean participants' knowledge of 20th century Korean history in mind. The hope was to heavily weight the possibility of *Japan* being considered as the agent of the deverbal noun *destruction*. Admittedly, cultural context was also the motivation for the gender-assignment to the arguments of *Steven* as the agent and *Jenny* as the goal in the set that involved *Steven's gift of flowers to Jenny*. The Korean set likewise involved *Changho* (a male) and *Pohuy* (a female) as agent and goal respectively in the corresponding set.

The number of syllables was much more difficult to control for between languages. For example, many of the Korean verbs were compounds of the type NOUN+*hata*, for example, *iyaki+hata* ['this story+do' ≈ 'talk about']. The nouns in these compounds were all of two to three syllables in length, thus a verb in Korean was necessarily four or five syllables as compared to the English verbs which were one, two, and four, for example, *found, mentioned, recommended*. There were trade-offs in the case of nouns as well, especially those that were used in the 2- and 3-ARG items, for example, English *criticism* (4 syllables) is translated in Korean as *piphan* (two syllables). Korean morphology added

syllables due to case marking as well. On the one hand — in the case of locative case marking — the added bound morphemes were roughly compensated for by English prepositions and obligatory articles, for example, English *at the party* (four syllables) versus Korean *phathi-twuyey* (four syllables). On the other hand, obligatory case marking on subject and object nouns in Korean nearly always results in the addition of at least one syllable to each noun. Thus, in general, the Korean sentences were substantially longer in terms of syllable length, even though in terms of words they were shorter, as shown in Table 4–1. The construction of the input sentences in terms of chunks is discussed in Section 4.3.4, below.

Table 4–1: Comparison of English and Korean input sentences in terms of length

category	*syllables*		*words**	
	English	*Korean*	*English*	*Korean*
MOD-NP	11–12	16–20	9	7
2ARG-NP	12–18	20-21	7–11	6–7
3ARG-NP	14–22	24–27	11–13	7–8

* includes bound morphemes

Each sentence conformed to a template, or *frame*, of the following form: *SUBJECT + VERB + nominal phrase + PP (English); SUBJECT + PP + nominal phrase + VERB (Korean)*. Into the nominal phrase slot of the frame was inserted a *vocabulary set*; the order of the elements in the vocabulary set was determined by the *example type*. The number of items in the vocabulary sets was small: the English set contained only 48 different lexical items, including nouns, verbs, articles, adjectives, and prepositions; the Korean set consisted of 46, if case

markers are considered separate lexical items. Example types consisted of various orderings of the elements found in the nominal phrase; the grammaticality of any particular ordering varied between the languages. Figure 4–1 shows an example of an English input sentence generated by the formula:

FRAME	Mary found _____ under the desk.
+ VOCABULARY SET	the, new, black, briefcase
+ EXAMPLE TYPE	*ADJ DET ADJ NOUN*
= INPUT SENTENCE	Mary found new the black briefcase under the desk.

Figure 4–1: Input sentence generation formula

There are two features to note in particular about the form of the frame. First, in both languages the nominal phrase slot appears in the third position, bracketed by constituents. This was done in order to "bury" the element of interest, the nominal phrase, in a non-salient position; this is of particular importance to the construction of the EI instrument as discussed below in Section 4.3.3.2 under "Recency/primacy effects". Second, the nominal phrase appears in the same relative position in both languages, although it was not possible to bracket the nominal phrase by constituents in the same order. This is due to English SVO versus Korean SOV order.

For each of the modification categories, MOD-NP, 2ARG-NP, and 3ARG-NP, there were three frames and three vocabulary sets. The number of input sentences varied according to the number of example types in each modification category. The set available for input to the three tasks consisted of 9 tokens of

each of the 16 sentence types for a total of 144[1] input sentences in each language. The overall ratio of grammatical to ungrammatical was 7:9 in both languages. Shown here is one grammatical and one ungrammatical example type from each modification category in each language (see Appendix A for the complete set):

MOD–NP category: a noun modified by a determiner and adjectives

grammatical examples:
Mary found the big blue textbook under the desk.

yenghuy-ka	chayksang-mitheyse	ku	khun	phulun	chayk-ul	palkyenhayssta.
[name]-NOM	*desk-LOC*	*the*	*big*	*blue*	*book-ACC*	*find-PST-SE*

ungrammatical example:
Linda saw small the red notebook on the chair.

mica-ka	uyca-eyse	cakun	ppalkan	kongchayk-ul	ku	poassta.
[name]-NOM	*chair-LOC*	*small*	*red*	*book-ACC*	*the*	*see-PST-SE*

2ARG–NP category: a noun modified by a genitive-marked nominal as external argument plus one subcategorized internal argument

grammatical example:
They mentioned Japan's destruction of Seoul in the letter.

kutul-i	phyenci-ey	ilpon-uy	sewul-uy	phakoy-lul	iyakihayssta.
they-NOM	*letter-LOC*	*Japan-GEN*	*Seoul-GEN*	*destruction-ACC*	*talk-about-PST-SE*

---

1 The actual set of input sentences consisted of 9 tokens of each of 21 types for a total of 189 sentences. The complete set included an additional modification category containing five example types:
POS-NP: a noun modified by a genitive-marked possessive nominal, which in turn is modified by a genitive-marked nominal.
grammatical example:
I stopped by the shop of Jenny's neighbor after class
ungrammatical example:
They arrived at David's teacher's the house before noon
The results on these sentences are not examined in this study.

ungrammatical examples:

We discussed Steven's of Linda photo at home.

wuli-ka	swuep-eyse	chelswu-uy	mica-uy	sacin-ul	ku	tholonhayssta.
we-NOM	*classroom-LOC*	*[name]-GEN*	*[name]-GEN*	*photo*	*the*	*discuss-PST-SE*

3ARG–NP category: a noun modified by a genitive-marked nominal as external argument plus two subcategorized internal arguments

grammatical examples:

We described Steven's gift of flowers to Jenny to the group.

wuli-ka	salamtul-eykey	chelswu-uy	pohuy-eykey-uy	kkoch-uy	senmul-ul	kiswulhayssta.
we-NOM	*group-DAT*	*[name]-GEN*	*[name]-DAT-GEN*	*flower-GEN*	*gift-ACC*	*describe-PST-SE*

ungrammatical example:

I recommended Mary's to Robert of money the donation in the letter.

kutul-i	phyenci-ey	changho-uy	mica-eykey-uy	ton-uy	kipu-lul	chuchenhayssta.
they-NOM	*letter-LOC*	*[name]-GEN*	*[name]-DAT-GEN*	*money-GEN*	*donation-ACC*	*recommend-PST-SE*

4.1.2. Participants

There were 49 English-speaking participants and 44 Korean-speaking participants. The 49 English-speaking participants were recruited from Psychology 100, 220, and 371 classes at the University of Hawai'i. Each earned $10.00 and 2% credit towards his or her final grade in return for participating. They ranged in age from 18 to 42, with a mean age of 23. All were native English speakers; there was a wide variety of second language experience among participants. Two participants claimed no knowledge whatsoever of a second language. While many had minimal experience with a second language (e.g., high school Spanish), others were exposed to other languages in the home from

childhood[2] — some professed good passive knowledge of these languages. The 44 Korean-speaking participants were all high school seniors from Haman Boys' High School in Haman, Gyeongsang Namdo, South Korea. Haman is a medium-sized agricultural community in the south of the country with a population of 50,000. There are 800 students at Haman Boys' High School. The students were not remunerated in any way for participating; in fact, they were required to do so. All 44 participants were in the same homeroom together and attended all the same classes together every day. This particular group was a "vocational track" group; that is, none of them had plans to apply to a university following graduation. Virtually all had grown up in Haman and had attended the same schools together. Their language learning history was equally similar: as is standard practice in Korea, all had begun instruction in English at the age of thirteen or fourteen and continued to the present time taking three to five hours a week. All had also studied German for two hours a week during the preceding year. None claimed exposure to any other language. They ranged in age from 17 to 19, with a mean age of 18.

4.1.3. Administration of the tasks

91 participants performed the three tasks in the same sequence: RT, EI, P&P; the remaining two, both Koreans, followed the sequence EI, RT, P&P due to scheduling constraints. The order of presentation was held constant as far as possible to mitigate the effects of repetition upon the consistency of evaluations (Nagata, 1988). The P&P was administered last to withhold the obvious grammaticality evaluation aspect of the task from the participants for as long as possible.

---

2 These included Japanese (3), Ilocano (3), Spanish (2), Korean (1), Tagalog (1), Hukunese (1), Chinese (1), Yapese (1), Pohnpeian (1), and Hawai'i Creole English (1).

The experiment using English-speaking participants was run in three separate rooms at the University of Hawai'i — the participants moved from one room to the next to complete the three tasks in sequence; the entire process took approximately 45 minutes to an hour per participant. The computer for the RT task was placed on a desk in an office with curtains obscuring the view of the outside. The elicited imitation experiment took place in a currently unoccupied graduate assistants' office. This room was without any distracting views and nearly sound-proof. The participant sat facing the opposite blank wall; the experimenter faced in the opposite direction. The pencil-and-paper task was administered after the other two experiments had been completed. Each participant completed the task while seated alone at a desk in a third windowless office attended by a researcher who was available to answer general questions or make clarifications about what was expected. For the English participants, each task was administered individually, each always by the same researcher, except for the EI task which was administered by two different researchers.

For the Korean participants, the experiment was run at Haman Boys' High School. The computer used in the RT experiment, was placed on a desk in a newly-built plant science laboratory which had not yet been put into service. The large, empty room was very quiet with curtains obscuring the view of the outside. In an adjoining laboratory, again not yet put into service, the elicited imitation experiment took place. The only furnishings were a desk, two chairs — one for the participant and one for the experimenter, and a large incubator. This room was without any distracting views and nearly sound-proof. Both the participant and the experimenter sat facing the same blank wall; behind was a window dividing the room from the adjoining room where the reaction time test

was taking place. The first two tasks were administered individually, always by the same two researchers, one for each task. The participants took the first two tasks consecutively which took each participant about 30 to 35 minutes. After all 44 participants had completed the other tasks, the P&P task was administered to them all at the same time in their classroom. The classroom was attended by their teacher and two researchers, one a native speaker of Korean who was available to answer general questions or make clarifications about what was expected. This researcher prefaced the task with a short explanation of grammaticality based on the instructions.

In the following sections, I describe each of the three instruments in this study in detail; first, I offer an overview of the method, then I discuss design considerations that are specific to the particular instrument under discussion.

4.2. Pencil-and-paper task: P&P

4.2.1. P&P: Overview

The pencil -and-paper grammaticality evaluation (P&P) task is in the category of *direct* grammaticality evaluations as I have defined it, in that it specifically focuses the participant's attention on making explicit judgments about language use and acceptability. P&P tasks may take take many forms; perhaps the most common is the type in which participants are asked to judge sentences, in isolation and without any additional context, in terms of "grammaticality" or "well-formedness ", "meaningfulness", "ordinariness", "naturalness", "comprehensibility", acceptability", "appropriateness", even "confidence in one's judgment" or "liberality". For an overview of first language

studies employing such measures, see Chaudron (1983, p. 352, Table 1). In an overview of second language acquisition research, Shimura (1990, p. 66) cites two main categories of direct tasks. The first is that referred to above which he designates as the *absolute grammaticality judgment*. Examples of previous use of tasks in this category may be found in the studies of Ioup and Kruse, 1977; Adjémian and Liceras, 1984; White, 1985, 1987, 1988, 1989a, 1989b; duPlessis, et al., 1987; Bley-Vroman, et al., 1998; Felix, 1988; Kui, 1988; Mazurkewich, 1988; Thom, 1988; Liceras, 1989; Merzenich, 1989; Renandya, 1989; Schachter, 1989a, 1989b; Suciadi, 1989. The second category, whereby the participant is asked to judge the grammaticality of one sentence relative to another, is the *comparative grammaticality judgment*; see, for example, the studies of McNeill (1966), Ritchie (1978), and White (1987, 1988, 1989a). A particularly ingenious method of extracting comparative grammaticality judgments is Sorace and Bard's (1990) use of magnitude estimation (Lodge, 1981), whereby the participant is presented with an initial example of language and then must assign numbers proportionally on a scale of their own device to subsequent exemplars. It has been shown that on retests of this type of task, participants tend to be consistent with their original estimations. Other tasks in the comparative category include rank-ordering tasks (Coleman, 1965; Mohan, 1977; Snow & Meijer, 1977; Kellerman 1978), think-aloud tasks (Ellis, 1991; Poulisse, et al., 1987), dyadic-problem-solving (Lantolf, Goss, and Zhang, 1991), passage correction (Cohen & Robbins, 1976; Jordens, 1977; Tucker & Sarofim, 1978; Arthur, 1980; Gass, 1983; Odlin, 1986a; Ellis, 1991). Direct grammaticality evaluations, both absolute and comparative, are also a feature of many miniature artificial language studies, for example, Moeser (1975, 1977), Nagata (1985), Byrne and Davidson (1985), Meier and Bower (1986), Goldstein, Angelo, and Wetherby (1987), Issidores (1987), and

Nayak, et al. (1990) — for a summaries of these, see McLaughlin (1980) and Masterson (1990).

The participant is usually given explicit instructions about what is expected of her, often including examples. One might hope that evaluations given under such conditions to be fairly straightforward and untainted by extraneous factors. Unfortunately, in a no-time-limit, focused language-use task such as P&P, as opposed to other types of tasks, specifically RT and EI, not only conscious knowledge of grammar is liable to be accessed, but also to varying extents other factors, such as attitudes about style, context, lexical selection, pragmatic circumstances, truth value, social acceptability, dialectal differences, and a participant's confidence in her own abilities to judge the input. Kellerman (1983) found his participants rejected grammatical sentences based on prescriptive notions of grammaticality or due to general markedness or perception of markedness. Birdsong, Johnson, McMinn, and Ingmundson (1985) found that even when provided with explicit instructions to base evaluations solely on syntactic structure, L2 learners of French were found to have based some evaluations on irrelevant morphological features. Likewise James Lantoff (personal communication) reported that, upon conducting informal post-task interviews of participants in an experiment, he found that some evaluations of "grammaticality" had actually been rendered on such extraneous features of the input as presence or absence of commas and periods. The type of knowledge accessed may be extremely subtle, for example, in an investigation of English native-speaker evaluations on verbs that do or do not participate in the dative alternation, Wolfe-Quintero (1992) suggests that a comparative grammaticality evaluation instrument will be sensitive to "strength of association", i.e., that an evaluation may be affected not only by the grammaticality of an item, but also by

a participant's perception of the prototypicality of the verb's involvement in dative alternation.

The challenge in constructing a P&P task which accesses primarily the feature one wishes to investigate is to design the task in such a way to access primarily the desired level of processing. This is true of the remaining two tasks as discussed below. In the case of the present P&P task, this means trying to focus the participant's attention, via the instructions, principally on the syntactic features of the items and controlling for potentially confounding features such as vocabulary and length (discussed above) and order of presentation.

4.2.2. Design considerations specific to P&P

Instructions. A particularly problematic area in the design of the P&P is the preparation of the instructions. First of all, the definition of "grammatical" can be extremely slippery, especially for so-called "naïve" participants. Additionally, an instruction section of a task which includes examples of the type of information the researcher seeks may contain subtle (or not so subtle) indications of researcher's expectations. Birdsong (1989, p. 64) warns against including examples of the structure under investigation: "the classic recipe for inducing response bias in grammaticality judgments involves demonstrating the required response procedure by offering answers to sample items similar to those under investigation" (see also the discussion in Rosenthal, 1966).

Much attention was given to the preparation of the full page of explicit instructions. Instructions developed in Bley-Vroman, Felix, and Ioup (1988) were used as a model; care was taken that the examples of grammatical and ungrammatical sentences were both unrelated to the sentences being tested and

as uncontroversial (in terms of grammaticality) as possible. Also included in the instructions was a survey to collect information on the participant's age, previous language experience, and degree of use of the left hand.

The native Korean linguist who acted as advisor for the Korean data collection, formerly a high school teacher in a school similar to the one from which the pool of participants was drawn, helped with the preparation of the instructions. He strongly objected to the literal translation of the portion of the language-experience survey that questioned the participant's previous contact with other languages, either through travel or residence in foreign countries. He submitted that to ask these questions of the particular type of participant being tested would be offensive — that it was so unlikely that any of them had had any outside language experience at all that to question them about it was to point out this fact to them in a deprecating manner. Upon his insistence, the instructions were altered on this point. See Appendix B for the full text of the instructions in both languages including the translation of the instructions in Korean.

Randomization of input. There were three separate versions of the instrument, each of which presented a different 63-sentence subset of the entire battery in a random order generated by a computer program. There were three tokens of each of the sixteen example types in the three modification categories, MOD–NP, 2ARG–NP, and 3ARG–NP, plus three tokens of the five example types in the POS-NP category (these are not analysed as part of the present study). Each subset was balanced internally and comparable externally to the other sets in terms of modification category, frame, example type, and vocabulary set. No repeats of any item were included either within a version or across versions.

Scoring the data. Previous studies have employed many types of scoring procedures such as binary, scalar, and ranking. I decided against the use of binary judgments, which force the participant to render an absolute judgment (or none at all). I was interested in collecting relative evaluations of grammaticality, especially for the purposes of comparison with the results from the RT task which scores the data in terms of time taken, a scalar measurement. For this study, the participants were directed to judge the sentences and rate them by using a 5-step Likert scale to assign a value of 1 through 5, 5 being "good", 1 being "bad".

4.3. Elicited imitation task: EI

4.3.1. EI: Overview

The Elicited Imitation (EI) task can be used to collect information about grammaticality without having to question the participant directly. In this respect, it is similar to the RT task, but unlike the P&P task. In a typical EI task, a participant is asked to listen to a language sample, usually a grammatical string, and repeat it verbatim. A review of the literature on the use of EI in language studies uncovers two main approaches to the design of the experiment and analysis of the resulting data — a participant-oriented approach and a theory-oriented approach.

Participant-oriented approach. The participant-oriented approach is adopted in the investigation of the state of a participant's knowledge of a language as indicated by his or her performance on an EI task. Overviews of this

use of EI in language studies are provided in Flynn (1987), Lust, Chien, and Flynn (1987), and Chaudron and Russell (1991). In such studies, there is assumed a projected norm against which a particular individual's or population's performance is measured. The projected norm is represented by the performance (either assumed by the researcher or supported by L1 control data) of the normal adult native speaker possessing a stable, mature grammar of the language. The populations tested in this type of study are, for the most part, L1 children, language-disordered L1 children or adults, and L2 learners. In other words, the participants are those presumed to possess either developing or deficient grammars.

In L1 studies, EI tasks often are used diagnostically as tests of grammatical competence with language-delayed or otherwise language disordered participants. Within this paradigm, there are two types of language behavior to be investigated — (1) global proficiency, whereby a variety of structures are used as test items to arrive at a general picture of the participant's ability to manipulate many types of language, and (2) behavior on specific grammar points, whereby a limited number of syntactic structures are used as test items in order to gauge the participant's control on certain isolated forms. In both paradigms, the participant's knowledge of *un*grammaticality is not usually of primary interest and therefore is not assessed via the inclusion of ungrammatical test items in the task. In principle, studies of normal adult native speakers may be done, but only as a control or to provide baseline data for studies investigating developing grammars, not in order to investigate questions of native competence.

EI tasks are used in L2 studies to measure the developing L2 grammar against a hypothetical L1 model. It is more common in L2 studies to find

investigations of specific grammar points, even so, it is rare to find ungrammatical items included in task (but see e.g., Munnich, Flynn, & Martohardjono, 1990). In L2 studies, model adult L1 performance is commonly regarded as a given, though there are exceptions. For example, Wolfe-Quintero (1992), Chaudron and Parker (1990) and Merzenich (1989) collected L1 adult data to use as native-speaker controls in EI studies on L2 populations.

Theory-oriented approach. In a theory-oriented approach, the EI data collected would be used to test the predictions of a particular syntactic framework. In this use of EI, measures of global proficiency would be too general to be useful; instead measures on a few specific grammar points would be analysed in order to determine if a participant's performance conforms with the predictions of the framework under investigation. Such a task should include both grammatical and ungrammatical test items. In principle, the syntactic framework under investigation could describe either the the mature L1 grammar, or an interim stage of a developing L2 grammar. There have been few EI studies designed for the purpose of confirming hypotheses about underlying adult L1 systems, in fact this appears to be a critical gap. Most available L1 data to date has been collected in the context of serving as baseline data for participant-oriented studies, thus the data was incidental to the main purpose of the studies and was not collected with the intention of testing a particular framework or theory of underlying grammar.

On the surface, the present study appears to fall into the latter category of experimental design, that is, it assumes adult native L1 performance as the standard by which the validity of the syntactic framework is gauged. In the present EI task, the input consists of both grammatical and ungrammatical sentences. By examining the relative degree of success the participant shows in

producing a correct repetition, I hope to draw conclusions about about the correspondence between the participant's internal grammar and the predictions made by the syntactic framework. If the framework successfully accounts for the participants' performance, this will tend to confirm the framework's validity, or at the very least, not disconfirm it.

However, in order to undertake the experiment as outlined, one must make the assumption that EI is indeed a measure of grammaticality. This cannot, unfortunately, be accepted at face value, as the discussion below is intended to demonstrate. I do not change the basic design of the experiment; rather, instead of assuming the validity of the elicitation measure, I am assuming the validity of the syntactic framework. By doing so, I investigate the viability of the elicited imitation task as a measure of the congruence of the participants' performance with the predictions of the framework. In this way, I am proposing the present experiment as a necessary preliminary step before either of the approaches described above can be undertaken with confidence.

4.3.2. What does EI measure?

One way to investigate whether EI measures grammaticality is to compare it with results of other grammaticality measurement instruments; this has been undertaken in a number of previous studies. Results from previous L1 studies employing elicited imitation have uncovered a general correlation with results obtained from other measures, but have been less successful in eliciting performance reliability on specific grammar points. Chaudron and Russell (1991) cite several, including several which employ P&P-type tasks as cross-validating measures. Among these are Mecham (1974), who finds no significant correlation between EI results and results of other grammaticality evaluation tasks with

language-delayed children; similarly, Fujiki and Willbrand (1982), in an investigation of informal methods of language evaluation, found no significant correlation among measurements of grammaticality obtained on P&P, EI, and sentence completion tasks on the same type of population. Fischer, Albertini, & Samar (1987, p. 16) conclude "...there [may be] limitations as to what kinds of populations this test will be useful for. In particular, it may be less useful as a diagnostic test for particular structures, though it may still be of use as a test of global English skills." Chaudron and Russell sum up the results in their review: "...these L1 attempts to compare EI performance with other tasks, while generally positive, lead to a certain degree of concern for the potential of EI measures to elicit performance reliably of *specific* grammar points that would correspond to other measures of language performance [emphasis Chaudron and Russell's]" (p. 14). They warn of remaining problems with diagnostic use and mixed populations such as learners or language-delayed children. They find the case of L2 studies more encouraging, however: "[some] studies have demonstrated not only a general validation of EI as a measure of productive competence (Gallimore & Tharp, 1981; Henning, 1983; Perkins, Brutten, & Angelis, 1986), but to a certain extent, validation on specific target points" (p. 16). Likewise they note "reasonable evidence of a strong relationship between performance on specific structures using EI and other tasks" (p. 20); this is supported by conclusions of Munnich, et al. (1992, p. 13) who found that "with respect to grammatical sentences, both the elicited imitation and grammaticality judgment tasks provide strikingly convergent results" (but see the discussion of their results in Chapter 5, Section 5.3.5. in light of the findings of the present study). It is not clear, however, to what extent of these findings relate to the present study as all of these studies, which use the participant-oriented approach, were done on populations which

are not clearly comparable to the one investigated here: normal adult native L1 speakers.

Overall, the body of EI research would lead one to accept that, when other factors are held constant to the extent possible, there is a tendency for grammatical strings to be easier to imitate than ungrammatical ones. However, as Bley-Vroman and Chaudron (in press) observe, linguistic knowledge is neither a sufficient (for various reasons, one may not be able to imitate a grammatical string) nor a necessary (one may be able to imitate a string, even though it is not licensed by the underlying grammar) condition for accurate imitation. From this observation they conclude "that EI and productive use of grammar must involve different things" (p. 2). Their views on the process involved are discussed in the following section.

4.3.3. The EI process

4.3.3.1. Levels of representation and the controlling level

Bley-Vroman and Chaudron hypothesize that elicited imitation works this way: first, a participant hears the input in the form of an elicitation and processes it, forming a representation on a number of levels. Next, the representation(s) is stored in short term memory (STM). Finally, the participant formulates an output sentence based on what can be retrieved from STM. Figure 4–2 is based on Bley-Vroman and Chaudron's description of the process.

Bley-Vroman and Chaudron stress that the comprehension system itself does not require STM in order to calculate representations, that "the language processor automatically and obligatorily produces representations of the input" (p. 4). Thus, the event of forming a representation of a string, something one expects to occur regardless of whether a participant intends to repeat the string,

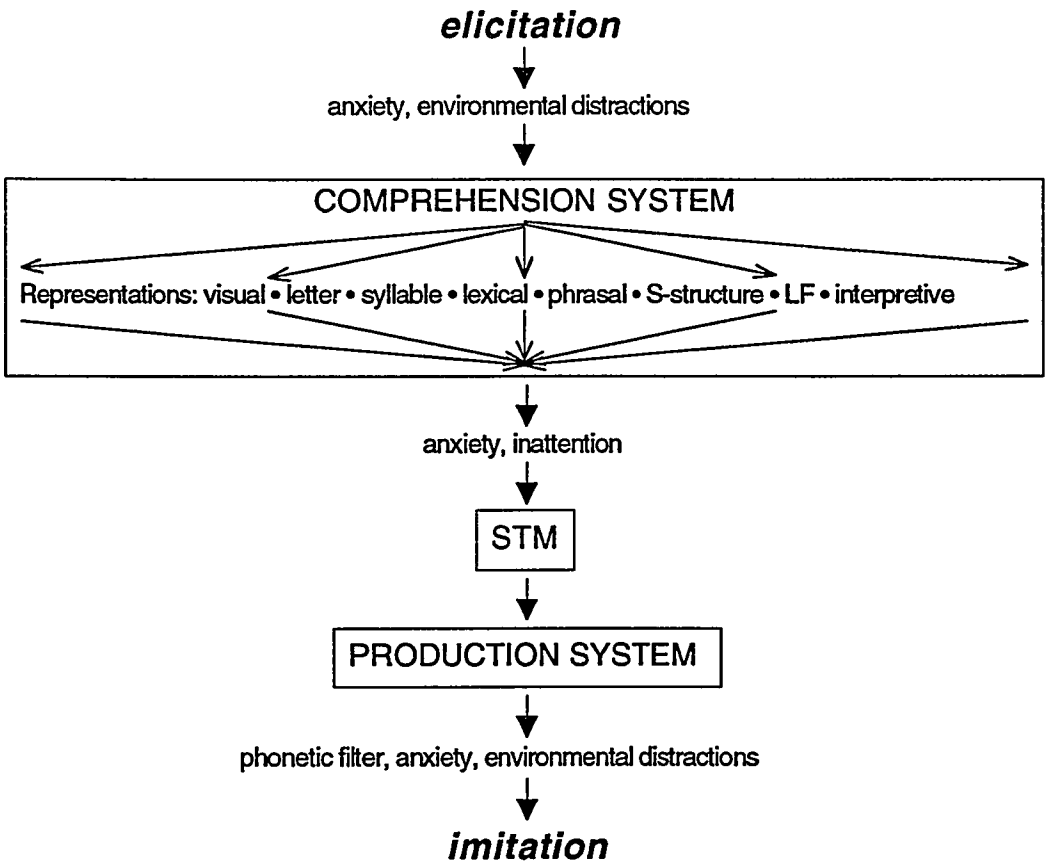


Figure 4–2: The EI process (after Bley-Vroman & Chaudron, in press)

does not "fill up" STM; rather the transfer of representation(s) to STM requires something more akin to intention. This allows the possibility of an interim stageat which factors such as inattention and performance anxiety may interfere. The representation must contain enough information about the string to allow an accurate repetition, but not so much that it cannot not be stored in and retrieved from STM; in other words, the less complex the representation stored in STM is (provided it is sufficiently complete), the easier the sentence will be to repeat.

For their model, Bley-Vroman and Chaudron adopt Forster's (1987) eight levels of representation . One or more of these come into play in encoding an

elicitation in the *comprehension system*. The lowest level is the *visual* level. It is not clear why this category of Forster's was adopted to explain the processing of aural input (perhaps *auditory* would have been more appropriate). However, even though, in theory, EI does not depend upon processing of visual input, some of the participants reported casually "picturing the sentence in my mind" in order to read it back out in imitation — a number of them closed their eyes during the task. It seems intuitive that this is also a possible strategy for repeating back strings of unrelated digits. In fact, it is possible in some cases to imitate a language string without processing it for sentence-level or even lexical meaning at all. Certainly many people can parrot back foreign language words, or even short phrases, without knowing their meaning. In such cases, the *letter* (or perhaps, *phoneme*) or *syllable* processing stages may take the major burden of passing on some sort of representation to STM.

The next level, the *lexical* level, has been the focus of a number of studies (for example, Lewandowsky & Murdock, 1989, discussed below). This is the level that is most likely accessed in the processing of lists of random words. From these lower level types of mental representation, the possibilities for representation levels range through to the interpretive, or "abstract message-meaning level". Lust, Chien, and Flynn (1987) describe processing that takes place at this level as "reconstruction". Evidence for this highest level is supplied by the observation that repetitions may be re-wordings of the input sentence, maintaining the interpretation, but not the syntax. Several examples of this were found in the data (see discussion Chapter 5, Section 5.3.5.).

In his 1987 paper, Forster asks the question: is there a "controlling" level and if so, which level fulfills this function? Naturally, the controlling level will vary with the task at hand. Short, simple input that requires no more than lexical

representation in theory need not operate at the S-structure or interpretive level, for example. The controlling level is susceptible not only to variation across tasks, but also across individuals — "There are no grounds for insisting that a particular level will always be the most efficient for all subjects on all occasions" (Forster, 1987, p. 74). Proposals have been put forth that there is a single, composite representation simultaneously encodes all available information (Frazier, 1987, p. 305, see also Pritchett, 1988; Levelt, 1989; Berwick, 1991; all cited in Bley-Vroman & Chaudron). Alternatively, there are proposals in which many levels are involved in parallel (see discussion in Levelt, 1989).

Bley-Vroman and Chaudron suggest the phrasal level as a candidate for a controlling level in the case of common input consisting of sentences. Their argument is based on the interaction between the need to maintain the number of elements that must be manipulated to a level that can be handled by STM (a relatively high-level representation) and the need to retain a sufficiently detailed representation to provide an accurate, word-for-word imitation (a relatively low-level representation). The application of Abney's chunking proposal (described in Chapter 2) may offer an explanation for as to why grammatical sentences should be easier to imitate than ungrammatical sentences: the former are usually composed of a few coherent chunks while the latter are composed of many disjointed phrasal fragments. The challenge to the design of an EI instrument which successfully accesses the level of syntactic processing is to create input sentences of the optimum length and complexity to encourage primary access to the theoretically relevant level among participants. For example, let us assume that, for example, "seven" is the average number of chunks that can be stored in and retrieved from STM and still provide the production system with a viable representation on which to formulate the output imitation. One would then

predict that, in general, participants could successfully repeat sentences of six chunks whether they were grammatical or not. On the other hand, one might expect a participant to find even grammatical sentences of more than seven chunks in length difficult to repeat verbatim, although she might be able to recast the meaning of the sentence in her own words, thereby indicating recourse to a higher (interpretive) level of representation. Alternatively, one might expect the participant to produce a partial imitation — an imitation which includes some, but not all of the chunks in the input string — or imitations with other errors such as transposed arguments or, in the case of ungrammatical sentences, grammatical "corrected" versions.

4.3.3.2. Interfering and competing factors

An understanding of the controlling level may not be sufficient to explain results obtained on an elicited imitation task. In the case of EI, similar to that of the P&P task, not only are there competing levels of linguistic knowledge involved, but also non-language-specific contextual and experimental factors. As mentioned above, both the original intake of the elicitation as well as the transfer of the representation of the input to STM may be impeded by performance anxiety. After navigating STM, a representation is filtered through the production system, in other words, a participant must formulate the imitation, modelling it on the representation formed, to the extent the representation is intact in STM. This involves the participant's phonetic realization of the string. Anxiety may act as a filter at any of these three points, along with environmental conditions, such as noise, fatigue, and hunger.

Certain effects discovered in serial recall experiments also have implications for EI. Three important considerations arising from the research of Lewandowsky and Murdock (1989) and others summarized in Lewandowsky and Murdock are: the effect of length, recency/primacy effects, and the sensitivity of the task to the nature of the individual items.

Effect of length. The first finding is that, not surprisingly, the longer the list of input items, the lower the success rate in recalling them all, and recalling them in order. Chaudron and Russell (1991, p. 7) cite two studies indicating that length is more critical than syntactic complexity or word familiarity (Montgomery 78) (Perkins 86). Similarity, Perfetti (1969, p. 720), in a study exploring lexical density and phrase structure depth as variables in sentence retention, claims, "..recall probability of individually presented sentences was not related to d[epth]."

Lewandowsky and Murdock's data revealed floor and ceiling effects, that is, the shortest lists were easy enough to assure almost total success on recall, while the longest lists were rarely reproduced correctly. From these findings, I predict imitation in the EI task to be accurate up to the point that STM begins to fail. i.e., the point at which it is required to hold too many items or *chunks* in order to provide an accurate repetition. Beyond this point, accurate repetition is much less likely. A number of small pilot studies were run to try to isolate the range in which imitation began to break down, indicating that the limits of STM were being taxed. This range is estimated to be from about six to seven chunks. In fact, Miller (1956) suggested the number "seven" as the approximate upper limit for the number of chunks that may be successfully stored and retrieved. Seven is also the mid-point of the range within which Lewandowsky and Murdock detected the most interesting recall effects, as discussed below.

Recency/primacy effects. Another important trend apparent in the research is the effect on recall of serial position; an overview and summary of this research is included in Lewandowsky and Murdock. It has been found that, in general, the last item in a list is the most easily retrieved, followed by the first. Items "buried" inside the list are much harder to retrieve; data from experiments generally forms a U-shaped curve. This effect has been referred to as the "recency/primacy" effect. This effect may help to explain, for example, results reported by Munnich, Flynn, and Martohardjono (1992). On a grammaticality task in which the input was presented as audio-taped examples of grammatical and ungrammatical sentences, they found that L2 participants performed significantly worse on sentences in which the ungrammatical element occurred sentence-medially than sentence-finally. These results were not significant, however, when the task was replicated using written input. Unfortunately, the conclusion they drew from this was that the sentence-final error examples were somehow more "preferred" than the sentence-medial error examples for these participants. Additionally, in EI versions of this experiment, participants were more likely to omit the ungrammatical sentence-final elements and thus "correct" the sentence than to perform this operation on the sentence-medial examples. This may indicate, rather than an increased sensitivity to the ungrammaticality of the sentence-final type errors, an increased insensitivity to errors occurring sentence-medially, attributable to recency/primacy effects.

Sensitivity of the instrument. In their examination of the data, Lewandowsky and Murdock found that in the middle range (lists of about five to nine items), success in recall depended on the nature of the individual item involved, for example, lists of random digits or letters were easier to recall than lists of random words, revealing an effect for some factor other than the mere

number of items in a list. Bley-Vroman and Chaudron (p. 10) suggest that "lexical items, being from a much larger set that requires more encoding, are more difficult to retain in short-term memory than letters or digits." Thus, differences may be attributed not only to the fact that lexical items belong to a set with many more members than do single digits or letters, but also to the possibility that a lexical item requires a more complex mental representation than does a digit or letter, involving multiple semantic representations plus information about category membership, morphological constraints, and anaphoric reference, for example. Semantic and syntactic relationships among lexical items in a phrase or sentences further complicate the picture. One might predict that a lexical item whose syntactic relations within in the string is uncertain or ambiguous, as would be the case in an ungrammatical string, would likewise require more encoding. In theory, this phenomenon should reveal itself in decreased accuracy in imitation for items which require more encoding within the sensitive middle range or *narrow band of sensitivity* as this range is described by Bley-Vroman and Chaudron. This narrow band is presented diagrammatically in Figure 4–3. The extremes that fall outside this band account for floor and ceiling represent the effects sometimes encountered in data collection via EI, that is, the cases in which the majority of the participants uniformly succeed or uniformly fail the task. For the items that fall inside the band, the EI instrument is sensitive to various types of complexity, thus effectively distinguishing between items that are identical in other aspects. It is the narrow band of sensitivity that I target with the EI task.

What types of "complexity" is this band sensitive to? The prediction of the present study is that significant effects will be found for two in particular: grammaticality and length (as measured in chunks).

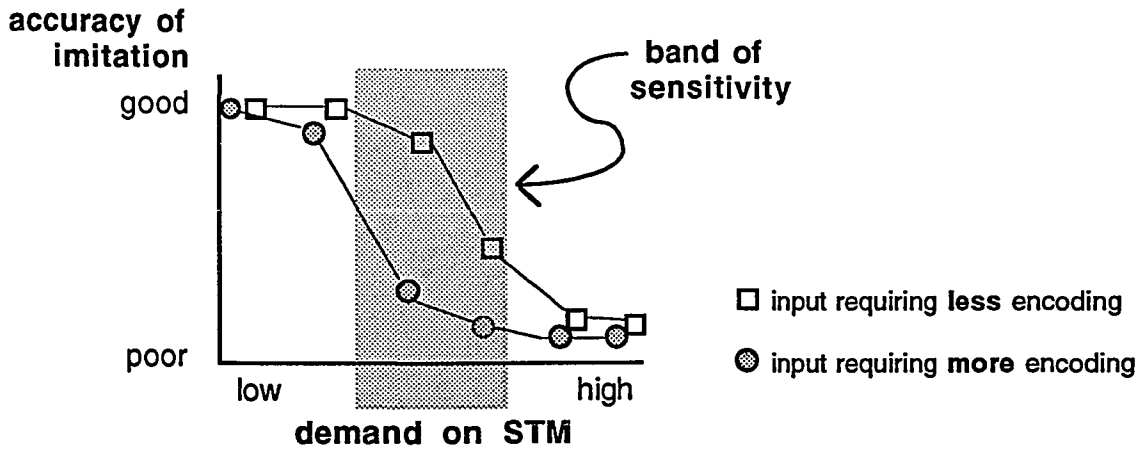


Figure 4–3: "Narrow band of sensitivity" (after Bley-Vroman & Chaudron, in press)

There is general agreement that in sentence string input, syntax has a role to play. Slobin and Welsh (1973, p. 496) suggest that imitation is "filtered through the individual's productive linguistic systems." Similarly, Bley-Vroman and Chaudron tentatively identify the complicating factor in the case of elicited imitation of sentences: "accuracy might be affected by details of syntactic structure" (p. 9). However, they seem to restrict the major role of syntax to a phrasal level, rather than a sentential level: "accuracy may often be relatively insensitive to high-level syntactic and semantic facts....but relatively sensitive to phrase level aspects of syntax" (p. 9). In the same paper they note the following two "*prima facie* facts about EI": "[1] other things being equal, it appears easier to imitate grammatical sentences than ungrammatical ones" and "[2] the longer a sentence is the less likely it is to be accurately repeated.....[this] suggests that some sort of limited capacity memory system is involved" (p. 2). These two observations, taken together, alert one to search the data from an EI task for evidence of the effects of and interaction between two aspects of input: grammaticality and length in chunks.

Grammaticality and chunks, though potentially related, are not congruent. Although it is necessary to hypothesize the syntactic structure of a string in order to calculate the number of chunks, this number in and of itself is not a measure of grammaticality. Grammaticality of a given string configuration varies according to language, thus the identical configuration (for example, ADJ–DET–NOUN) should have varying effects on the success of imitation according to the language in which it is realized. In Korean, such a string constitutes a single chunk, in English, more than one. On the other hand, the effect on imitation for length in chunks, as determined by the syntactic structure of a particular language, may be relatively non-language specific, having to do more with general, or universal, language processing strategies. There exist examples in the input set by which these two can be distinguished, for instance, in first pair of sentences below, the length in chunks is the same for both, but the grammaticality values are opposite. Conversely, in second pair of sentences, the grammaticality values are the same, but the length in chunks differs. Similar examples exist cross-linguistically as well.

(1)	gram.	chunks
They / talked about / Japan's / destruction / of Seoul / in the letter	+	6
They / talked about / Japan's / Seoul 's / destruction / in the letter	–	6
(2)		
We / described / Steven's / to Jenny / of flowers / the gift/ to the group	–	7
We / described / Steven's / to Jenny / of flowers / gift / the / to the group	–	8

Examples like these will be used to determine if there are significant effects for grammaticality and/or chunks, independent of one another. Data collected via all three tasks should reveal the presence or lack of these effects for both the

grammaticality, which is different for the two languages, and for the length in chunks, which is the same for the two languages.

4.3.4. Design considerations specific to EI

An examination of the existing studies incorporating EI tasks yielded the following list of factors, specific to EI tasks, that may affect the results of the data collection. These factors have been taken into account in the design of the present EI instrument. Before describing aspects of the design which are specific to the EI instrument, I describe the overall design and procedure for administering the EI instrument.

Overall design and procedure. The task administrator used a Sony TCM–5000EV tape recorder to play taped instructions for the participant which included four examples of model elicitations followed by imitations. In the recorded instructions it was stressed that the participant should do his or her best to recall as much of the sentence as possible, even if the imitation were incomplete (a transcript of the English instructions and the Korean instructions with a translation appear in Appendix C). The experimenter controlled the presentation of each of the taped elicitation sentences by pausing and resuming the elicitation tape. An identical tape recorder, constantly running, was placed in front of the participant to record the elicitations and the following imitations. Even though previous studies have imposed a time limit within which the participants were to answer (e.g., Munnich, Flynn, & Martohardjono, 1992 limited response time to ten seconds), I decided to allow unlimited time. This gave the participant an opportunity to catch her breath, and to pace herself. Participants appeared to appreciate the opportunity to recover after an especially

difficult elicitation. The experimenter did not speak except to ask if the participant were ready to begin, to thank the participant upon completion and, during the experiment, to offer a little encouragement when it seemed advisable. The entire instrument, including instructions, took from nine to fifteen minutes; the elicitation instrument for the Koreans was longer than the English one by about three minutes.

The input sentences. A number of steps were taken to design the EI instrument so as to take advantage of the effect of length, the recency/primacy effect, and Bley-Vroman and Chaudron's narrow band of sensitivity. First, based on the estimation of six to seven chunks as the optimum length for uncovering the effects I was looking for, I used sentences of from four to eight chunks in length. Additionally, the structure of interest was consistently "buried" in the same sentence internal position to thwart primacy/recency advantage, that is to avoid the sentence-initial and sentence-final positions which Lewandowsky and Murdock have shown to be more easily retrieved. Third, in order to take advantage of the sensitivity of recall to the nature, or complexity, of the item type, the input was set up to be comparable in all aspects except that of grammaticality or length in chunks which are the elements of interest.

Further, taking note of Keller-Cohen (1981), who found lexical complexity contributes to difficulty in recall, I avoided introducing additional complicating factors such as semantic or syntactic ambiguity, ambiguous anaphoric reference, or unfamiliar or unusual vocabulary as far as possible. In order to limit somewhat the set of items that made up the input sentences, the actual number of lexical items in the vocabulary set from which the sentences were constructed was constrained, as discussed above. This was done to provide some familiarity with the possible set of lexical items and to reduce the introduction of new items.

One result was, as a number of the participants commented, that the sentences sounded very similar one to another and that this may possibly have caused some confusion.

Randomization. Three different randomized versions of the input for each language were prepared. As with the P&P instrument, random order was generated by a computer program. Two dummy elicitations that preceded the actual input set were not included in the statistical analysis. No repeats of any item were included either within a version or across versions.

Pre-recorded elicitation. I felt it imperative to tape the elicitation sentences so that there would be no "experimenter effects" (Bley-Vroman & Chaudron, 1991, p. 3) due to inconsistent live presentation of the sentences — many of them were quite difficult to read smoothly.

Contextualization. Though it appears reasonable to assume that contextual enactment or visual clues would aid in comprehension and therefore strengthen the internal representation of a string to be imitated, experiments with children by Lahey, Launer and Schiff-Myers (1983) and Weber-Olsen, Putnam-Sims, and Gannon (1983) show varying effects for contextual support; furthermore, it is not clear how transferable the results of this study are to the present population. There is also the additional problem of how to provide a context for an ungrammatical string; in the present instrument, no contextual support was provided.

Comprehension component. Bley-Vroman and Chaudron (1990), in their review of Flynn's research on anaphoric co-reference, recommend the use of a comprehension task in order to provide evidence that the participants are indeed making the correct assumptions about the identity of the various pronouns in the input sentences. Flynn did, in fact, include such a component, although Bley-

Vroman and Chaudron doubted its validity as evidence of appropriate pronominal identification. Furthermore, Flynn was specifically investigating anaphoric reference, not grammaticality, thus, evidence of comprehension was critical to making claims based on the results. As the present study is testing native speakers on fairly straightforward sentences (in terms of simplicity of semantics and vocabulary), it seemed reasonable to omit the additional task. Furthermore, in the interest of both maintaining the overall battery of tasks at a manageable length and also making comparisons across tasks as straightforward as possible, I decided to forgo the addition of a comprehension component. Admittedly, the fact that the input sentences (at least the grammatical ones) should be easily processed for meaning is no guarantee that they will be. Nevertheless, I retreat to the position that previous studies have shown that, all other factors being equal, ungrammatical strings cause more difficulties (in this case, in terms of the possibility of a correct repetition of the input) than grammatical ones and thus a difference should be evident. Differences may even be attributable in part to the decreased comprehensibility of the ungrammatical strings.

Pre-imitation delay. Some EI studies, for example, McDade, Simpson, and Lamb (1982); and Grigg (1986), have imposed a specific delay before allowing the participant to produce an imitation of the input string. McDade, et al. found that although imitation success was greater following no delay, children's performance after a three-second delay correlated more strongly with other measures of grammatical development, indicating that, given time, the children had an opportunity to call upon their rule-based competence in production. These results were supported by the findings of Grigg in his study of mixed-L1 learners of English. Unfortunately, pilot studies of the input sentences indicated

that the input sentences used in the present study were already so difficult that the addition of an imposed delay before the participants' imitation made success unlikely. As shortening the input sentences would make them less comparable across tasks, I decided to not to include a pre-imitation delay in the design of the EI instrument.

Scoring the data. I did not follow the example of many previous studies which have employed a binary right/wrong scheme for scoring the resulting data of an EI task. In order to more finely discriminate among the data, Bley-Vroman and Chaudron (in press) recommend that the researcher score specific target forms and do error analyses; not simply evaluate the data for whole-sentence success or failure. Their suggestion was followed in part in that only the nominal phrase in question was considered when assessing success or failure of imitation; formal error analysis of the imitations will be left for future analyses of this data. In addition, any nominal phrase which contained lexical substitutions was not considered incorrect. There is a further discussion of lexical substitution plus examples of common types of substitutions made in Chapter 5, Section 5.3.5. After the nominal phrase was isolated and lexical substitution was allowed for, each imitation was given a score of either "1" for "successful", or "0" for "unsuccessful".

4.4. Reaction time task: RT

4.4.1. RT: Overview

The Reaction Time (RT) task, like the Elicited Imitation task has been used in both first and second language research programs as an indirect measure of

grammaticality. In a simple choice-reaction time task such as the one presented here, a participant is presented with a stimulus, in this case, an instance of language, to which she reacts by pressing one of two keys on a computer keyboard, striving for both speed and accuracy in making her choice. The idea behind the analysis of the data obtained is that varying latencies will reveal varying degrees of processing difficulty, and thus allow the researcher to draw inferences about the nature of the underlying system.

In reviewing the literature on the use of reaction time, "psychology's ubiquitous dependent variable" (Luce, 1986, p. 1), I find that of the three methodologies reviewed here, RT is perhaps the one that has produced the least consistent or clear-cut results in previous language studies. There has been less general agreement about the value of the data produced by this instrument both in terms of its consistency and of its value as an indicator of underlying mental representation. A number of studies which take a critical look at the uses and limitations of this methodology are reviewed here.

Robinson (1992), in his overview of the use of RT data in both first and second language studies, discusses the relationship between the uses of RT and the field of theoretical linguistics. Robinson finds that the scepticism on the part of mainstream linguists that discouraged the use of such methodologies in the sixties and seventies has abated somewhat and these methodologies are again being employed as cross-validating measures in studies motivated by Chomskyan theory (e.g., Freedman & Forster, 1985; Bever & McElree, 1988; Masterson, 1989, 1992; Eubank, 1991; Hagen, 1991).

Researchers in natural language processing and parsing performance have increasingly turned to the use of reaction-time measures in their studies. Robinson cites, as examples, Dowty, Kartunnen, and Zwicky (1985), Reyle and

Rohrer (1988), Iziwa (1989), and Pritchett (1992); see also Clahsen and Hong (1993). Though Robinson's emphasis is upon the use of RT data in second language acquisition research, he stresses the need for solid first language baseline data, "...any comprehensive theory of second language acquisition must have a psychological component which makes claims about 'representation', and supports its claims with reference to some data base." (p. 1). Along with growing interest in the use of RT both as a source of L1 baseline data and as a cross-validating measure is an increasing sensitivity to the complexity of the measure itself and of the data obtained. It appears that the story is more complex than previously supposed and thus requires a more critical examination of both the limitations on the use of reaction time and of the various levels of description which may be necessary in order to account for the processes at work.

Robinson points out two limitations of RT studies which are also frequently mentioned with respect to the other two tasks in this study. These are *incompleteness* and *indirectness*. Data obtained from an RT task, like data obtained from any single task, is *incomplete* in that the data is of limited value without converging evidence from other experimental measures. Secondly, RT data alone is *indirect* in that, as an output measure, RT furnishes data which may be used only to make inferences about the representation underlying the performance. Attention needs to be paid to physiological processes that may contribute to the outcome. One of the main purposes of this study is to address the issue of incompleteness. Other studies have succeeded in uncovering supporting evidence for RT's sensitivity to grammaticality, for example, Moore (1972) and Mistler-Lachman (1972) found a correlation between results from judgment tasks and reaction time tasks. I will be looking for such a correlation here. Eubank (1991) issued the challenge to investigate "whether [RT] might be

productively employed as a measurement device in applied areas in much the same way that one employs the more traditional grammaticality judgment task or the elicited imitation task." Data is compared in this study from precisely these three measures on the same participants, using the same input. As for the issue of indirectness, the limitations of the present study for monitoring both the *functional characteristics* and the *physiological processes* involved — a distinction made by Manicas (1982) — are discussed below in Section 4.4.3.2. In this study, I concentrate primarily on the suitability of RT as a valid source of converging evidence of the grammaticality of the input set of sentences.

4.4.2. What does RT measure?

The model for the RT instrument was provided in large part by the previous work of Freedman and Forester (1985) on Subjacency violations and the Specified Subject Condition. Their work was a continuation of earlier experimentation with various string matching tasks beginning with one by Chambers and Forster (1975) who found that participants responded more quickly to word pairs such as HOUSE/HOUSE than to "pronounceable" non-word pairs such as HOSUE/HOSUE. Likewise, participants responded more quickly to these than to unpronounceable non-word pairs such as HSEUO/ HSEUO. Chambers and Forster claimed that actual words had the advantage over non-words in that they could be encoded as a single element, and that pronounceable non-words had an encoding advantage in that they could be encoded as syllables. Unpronounceable non-words could only be encoded as strings of individual letters, significantly slowing response time on a reaction time task. To express this argument in terms of chunks, HOUSE would consist of a single word chunk, HOSUE, of two syllable chunks, and HSEUO of presumably five letter chunks. The smaller the

number of chunks, the quicker the processing, and thus the shorter the response time. In the next series of experiments, Forster (1979) found that, in a similar fashion, two-word pairs that could be represented sententially such as DOGS GROWL/DOGS GROWL required shorter reaction times that non-syntactic strings such as GROWL DOGS/GROWL DOGS. In other words, DOGS GROWL can be thought of as being dominated by a single syntactic node, e.g., "CP" in many current frameworks, and thus as forming a single chunk. This is in contrast to to GROWL DOGS which cannot be represented as being headed by a single node. The general conclusion drawn from these experiments is as follows: participants required less time to complete a string-matching task if the strings in question could be assigned a higher level representation, whether it be syllable, word, or syntactic constituent, all other factors being held constant. The observation of this phenomenon might provide a key to the form of the underlying grammar.

4.4.3. The RT process

There are four steps, according to Welford (1980), involved in reacting to a stimulus in a choice reaction context. These are first, reception of the signal; second, identification of the signal; third, choice of a corresponding response; and fourth, initiation of the action that constitutes that response. Welford contends that the majority of the time it takes to react is taken by the second and third steps, identification and choice. I assume that, as is the case with the EI process, factors such as anxiety, inattention, and environmental distractions may interfere. It is the contention of the Freedman and Forester study described above that variation in reaction time on a language task which occurs during the identification and choice steps has something to tell us about ways in which

language is processed. The RT process is presented diagrammatically in Figure 4–4.

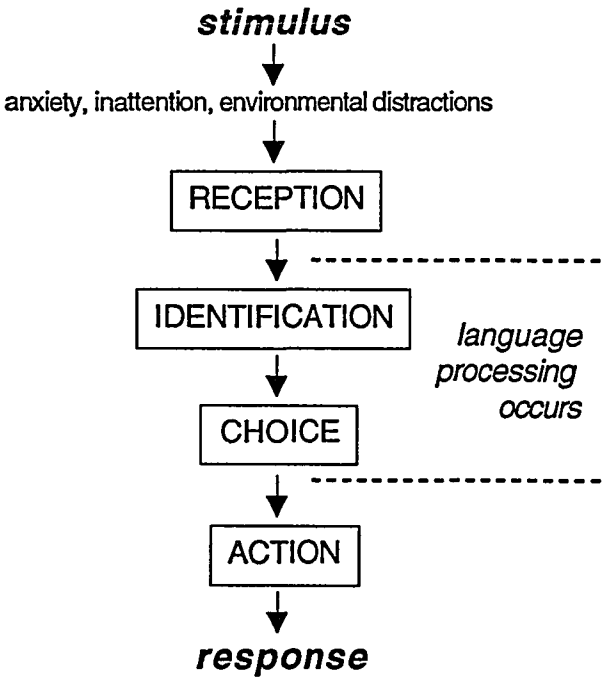


Figure 4–4: The RT process (based in part on Welford, 1980)

4.4.3.1. Levels of representation and the controlling level

The 1985 Freedman and Forester study explored the following question: at what level of representation did the sentence matching task take place? They focused on constraints on empty categories in *wh*-islands. Following Chomsky (1980), they assumed that sentences violating these constraints are initially overgenerated at the S-structure level, where they can be assigned a constituent structure, and that only at LF are such strings disallowed by the grammar, thus at S-structure, SSC violations, for example, **Who did the police believe that claim that John shot?*, are indistinguishable from grammatical sentences. On pencil-

and-paper grammaticality task participants evaluated these SSC violations as ungrammatical.

The results of their series of experiments led them to claim that it was indeed the case that the matching task utilized a level of representation at which such overgenerated sentences were processed similarity to grammatical ones. The matching times for these two types were essentially equivalent while times for ungrammatical strings such as those containing agreement violations (**Mary were writing a letter to her husband*), quantifier placement violations (**The baby ate up his cereal all*), phrase structure scrambles (**What to expect the judge told the thief*), and others were significantly longer. Thus, Freedman and Forester argued that at the level of processing that RT accesses, the SSC violations do not impede processing.

Crain and Fodor (1987) had a different explanation for the effects observed by Freedman and Forester. They claimed the issue was one of *correctability*, that is, the ability of participants to be able to easily correct some types of ungrammaticality, but not others. In their discussion of Freedman and Forester's study, they make an important distinction between sentences which violate constraints (those that patterned with grammatical sentences, such as SSC violations) and sentences that violate rules (such as the other types mentioned above). Pencil-and-paper pre-tests indicated that while constraint violations are difficult to correct, rule violations are not; native users of a language will generally agree on a precise rule violation correction. Under their proposal, when a participant is confronted with a rule violation, she mentally corrects it. Her representation is now at odds with the actual string, causing confusion and slowing reaction time. Such would not be the case with the overgenerated SSC violations as a structure that would facilitate the matching task would be

assigned at S-structure. An additional confounding factor which would be connected with the LF-level stage of interpretation, claim both the Freedman and Forster and the Crain and Fodor studies, the latter citing previous work by Murray and Ratcliff (1983) and Stevenson (1984), is the effect of semantic plausibility. Unlike Freedman and Forster who might claim that RT tasks can give us insight into competence, Crain and Fodor see them as more a measure of performance.

Eubank (1991), in a replication study, rejected both Freedman and Forester's and Crain and Fodor's characterizations of the processes at work. In fact, in his study, Eubank found that the SSC violations actually do result in longer latencies. Eubank concludes that RT tasks are sensitive to a variety of effects, among them is syntactic malformedness. He notes that all grammaticality evaluation procedures have drawbacks and states "...the sentence matching procedure would appear neither superior to, nor worse than, other typical measures" (p. 15).

Despite claims about what level is accessed by the RT task, the idea behind all these studies is similar. i.e., that by giving participants a matching task, information about grammaticality can be obtained indirectly, and some of the problems inherent in direct elicitation of grammaticality evaluations may be avoided. For purposes of this study, which focuses on a violation of the rule-violation type, it is not really important which of above explanations of the processing of ungrammatical strings, if any of them, is the correct characterization of the RT process as the outcomes are equivalent: ungrammatical sentences of the rule-violation type should require more processing time in the RT task as it has been designed.

There is one point of agreement among those who have incorporated RT tasks in their research programs: as Hagen (1991) comments, "...there has not yet been sufficient research on what such tests actually measure" (p .1), or as Eubank (1991) even more succinctly puts it, "...sentence matching is so poorly understood" (p. 22). This is the focus of the following section.

4.4.3.2. Interfering and competing factors

Apart from objections about the use of RT tasks because of linguistic-theoretical shortcomings, there is a body of literature objecting to their use on physiological processing grounds. For example, Manicas (1982), Luce (1986), Beretta (1991), and Crookes (1992) have questioned simplistic assumptions about the nature of the data collected through methodologies such as reaction time. Manicas (1982, p. 164) suggests that one consider psychological explanation as exhibiting two phases: *functional characterizations* and *physiological processes*. The first phase refers to analysis of the data from the perspective of the psychological processes which control a participant's performance, i.e., the participant's perception of the task. However, these are also physiological processes that come into play and effect the outcome. Recent studies support the idea of the involvement of many factors in the processing of language in an RT task — fortunately, work has begun on teasing apart the various complexities involved in performing these tasks. Under investigation are accompanying physiological aspects such as evoked brain potential (Kramer & Stayer, 1989), eye movement (Carrithers, 1985), the use of different hands and fingers (Welford, 1971; Heuer, 1981), and the influence of familial *sinistralis* (Cowart, 1985). Additionally, work has been done on the effects of fatigue, stress, and age; these are summarized in Welford (1980).

The present study is one which Manicas would characterize as a "phase-one study", that is, a study principally concerned with the functional characterizations of the data collected.

4.4.4.4. Design considerations specific to RT

The RT task design for the present study is based, as mentioned above, on the controlled presentation/timed-response technique originally developed by Freedman (1982; as reported in Freedman and Forster, 1985) and represents what Luce (1986) identifies as the *matching paradigm*. In the following description of the design, elements that are borrowed directly from the Freedman and Forster (1985) design are indicated by the notation "F&F". Before describing aspects of the design which are specific to the RT instrument, I describe the overall design and procedure for administering the RT instrument.

Overall design and procedure. The RT experiment was run on a Macintosh Powerbook 140 using a HyperCard program (Kim & Masterson, 1992) designed specifically for the task. A native-speaking researcher sat to one side, first to guide the participant through the instructions and practice session and to answer questions, and then, to observe the process. Both during the 5-pair practice session (see Appendix E for the set of practice sentences in each language) and during the actual task the priming sentence was displayed for 4160 msec. (Figure 4–5). The priming sentence then disappeared and a pointing finger appeared for 360 msec. to indicate in which of four random locations the pair sentence was to appear (Figure 4–6). This pointing finger, or *warning signal* in Luce's terminology, serves two functions: the first is to alert the participant that appearance of the pair sentence is imminent, encouraging a high level of attention; the second purpose is to force the participant to focus momentarily on

a new image and, it is hoped, to "erase" the visual image of the priming sentence, forcing the participant to rely more heavily on representations other than the visual. As soon as the match/mismatch pair sentence appeared (Figure 4–7), the participant was to indicate, as quickly as possible, if it matched exactly the priming sentence by pressing the red key ("f", covered with a red tag), or if the sentence differed in some way by pressing the blue key ("j", covered with a blue tag) (F&F). The was no time limit on how long the pair sentence remained on the screen. The participant initiated the next set of sentences by pressing the space bar. (Bever, 1988).

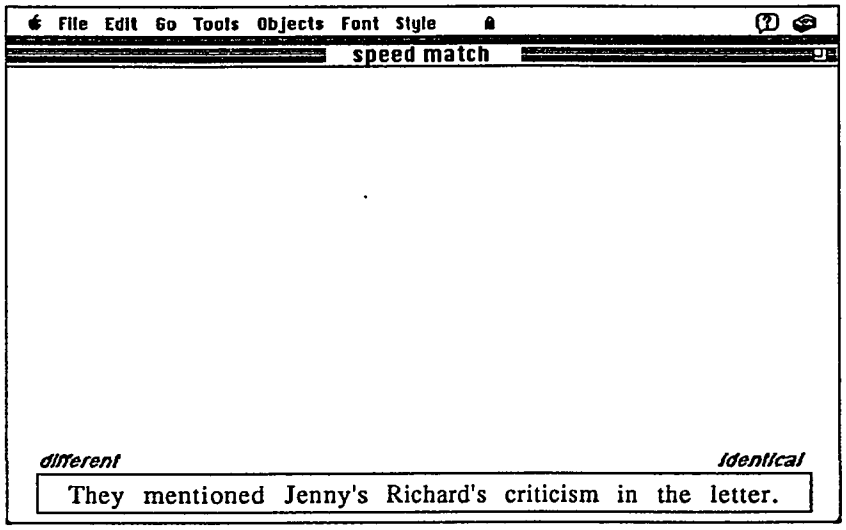


Figure 4–5: RT screen: Priming sentence

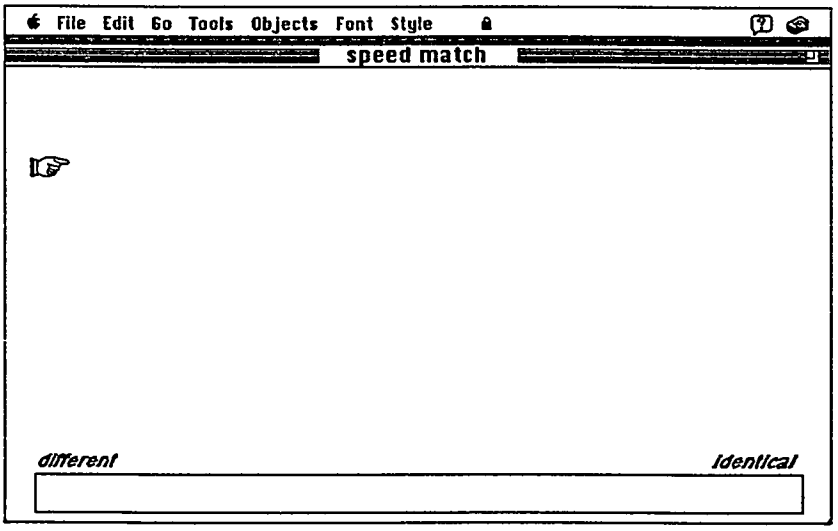


Figure 4–6: RT screen: Pointing finger

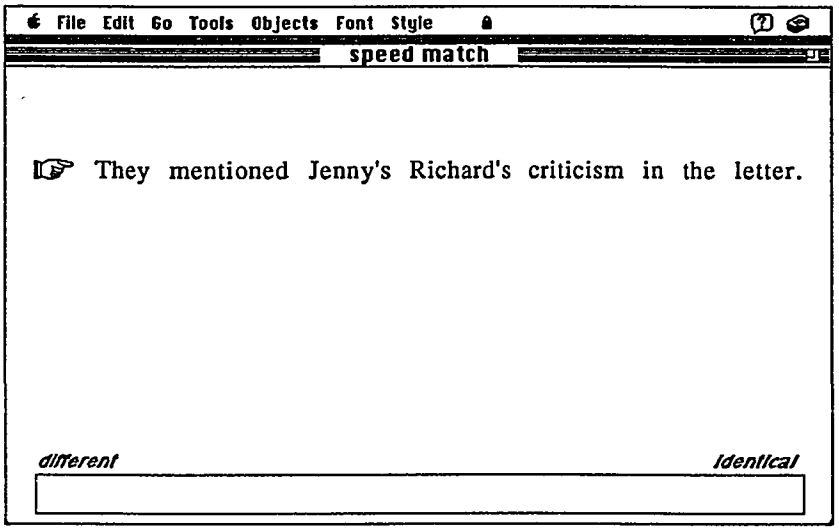


Figure 4–7: RT screen: Pair sentence

The program presented 63 randomly ordered pairs of sentences and recorded time between the appearance of second sentences in a pair (either a match or mismatch) and the participant's decision (key press). The resolution of

the system clock was 1/60 of a second (16 msec.), or one Macintosh "tick". 48 pairs included three tokens of each of the 16 example types from the three modification categories. Each pair consisted of either a sentence plus an identical repeat (match) or sentence plus a repeat that differed by one word, always a noun (mismatch). The ratio of matching pairs to mismatching pairs was 2:1. On average, the entire process, including instructions, practice session, and response to the elicitations took about fourteen minutes — approximately nine minutes just to respond to the elicitations.

Even though this RT instrument was based an an existing instrument, there were many specific design factors to be considered for inclusion, exclusion, or adjustment. These are described below.

Instructions. F&F do not explicitly refer to an instruction component in the report of their study. The instructions were fairly long and explicit, including an example of the appearance of an input sentence. The instructions were interactive, that is, they required active participation in the form of pressing specific keys on the keyboard. This was motivated in part by a desire to familiarize the participants with the computer key board; a number of the participants had not used one before. The task was introduced as a game called "Speed Match"; this name was repeated twice in the instructions to emphasize the importance of responding quickly. This was also stated explicitly in the instructions, as well as being mentioned by the researcher while guiding the participant through the instructions. The full transcripts of the English and Korean instructions, plus a translation of the Korean, are found in Appendix D. Following the instructions was a five-item practice session. At this point, the researcher offered to answer questions, or to replay the instructions or practice

session. When the participant indicated that he or she were ready, the researcher initiated the actual task.

The input sentences. All participants responded to the same subset of the input sentences. Following a suggestion of Miller (1988), I included the same number of trials (3) on each of the 16 example types to overcome the "familiarity factor", i.e., the unfair advantage given to a type which is over-represented in relation to others. The ratio of grammatical to ungrammatical was 1:1. I did not want to encourage an expectation of one type over the other. The ratio of matching to mismatching pairs was 2:1. Only the results on matching trials were to be considered for statistical analysis (see "Scoring the data", below). It was important to obtain a reasonable amount of data from each participant and so it was felt that a 1:1 ratio between matching and mismatching pairs either would "waste" too many responses by removing them from analysis or, alternatively, the task would have been too long.

Eubank (1991) — also taking F&F's sentence matching procedure as a model — on a series of studies testing grammaticality contrasts for both L1 and L2 speakers of German, had a great deal of trouble obtaining significant contrasts with the native speakers and found he was only able to do so by adding relative clauses to lengthen the input sentences, thus taxing short-term memory. Similarly, Cowan and Abe (1991) were able to obtain sharper contrasts by lengthening the input sentences. For this study, I experimented with added relative clauses and prepositional phrases, piloting at least four versions before arriving at what I felt to be the optimum length for both the RT and EI tasks.

A number of previous RT studies (e.g., F&F, Bever & McElree, 1988; Eubank, 1991) have included distractor items among the input. Cloitre (1984),

cited in Bever & McElree, 1988) claimed distractors to be an important element of instrument design and that participants can be extremely sensitive to obvious characteristics of the input that allow them to make predictions about the researcher's expectations. Distractors were considered especially important in the present study as the "real" input sentences were predicated on an extremely limited vocabulary using little structural variety. I found, upon running pilot studies, that participants commented on a perceived repetitiveness of the priming sentences (although, in reality, no two were identical). At this point, distractors of varying length and structure in each language were added to discourage the participants from developing and anticipating a "template" (see Appendix E for complete set of distractor items in each language). The ratio of distractors to the actual input sentences was 1:4; the 2:1 ratio of match/mismatch pairs was maintained among the distractors. The input to the RT instrument also included two filler items as the first two items. In total, the RT instrument contained 63 pairs of sentences — previous native-language studies have presented 60 to 90 items. Because the RT task was only one of three to be performed, it was felt that more pairs would be too tiring for the participants. There were no repeats of any priming sentence.

Randomization. Cowan and Abe presented their input sentences in random order; F&F do not provide this information about their experiment. The 63 pairs of the RT instrument were presented in a random order generated by the computer program — different for each participant.

Physical appearance of the input on the screen. The input sentences were formatted in standard upper and lower case. F&F chose to present their input all in upper case characters to confound the "skyline" effect which they suspected might aid visual matching. Nevertheless, I chose to follow Bever and McElree

(1988) and Eubank (1991), who employed "normal" orthography. I did this to maintain a more natural text appearance and to keep the presentation consistent with the P&P task presentation. In compensation, the location of the appearance of the pair sentence was randomized as described above. Eubank presented his priming sentence in the upper left-hand corner of the video monitor and the pair sentence at the bottom, offset to the right, presumably to confound visual matching. It is not clear exactly how F&F presented the input, though the formatting of examples in their paper indicates the possibility that they were stacked one over the other. In the present design, a priming sentence appeared centered at the bottom of the screen. It then disappeared and then, following the warning signal (pointing finger) and a 360 msec. delay, the pair sentence appeared randomly in one of four possible locations. In previous studies, Bever and McElree delayed the appearance of the single probe word by 900 msec.; Eubank delayed from between 1750 to 3000 msec. between the appearance of the first and second sentences. In his experiment, unlike the present one, both priming sentence and match/mismatch pair sentences remained on the screen simultaneously. In order to discourage visual comparison matching, this procedure was not adopted. The delay in the present study was shorter, but it was interrupted by the appearance of the pointing finger, as described above.

Use of a computer keyboard. As described above, participants reacted to the pair sentence by pressing a key, either red ("f") for "different", using the left hand, or blue ("j") for "identical", using the right hand. Welford (1971) ran a series of experiments on the use of different fingers and hands with a keyboard on reaction time tasks. Not surprisingly, he found that the right hand for most participants was significantly faster, as was the index finger over all others. The participants were not instructed as to which hand or fingers to use, nevertheless,

all chose to use the index finger of the left hand on the red key and the index finger of the right hand on the blue key. Participants alternately used a thumb or an index finger to press the space bar for the next pair, but there was more than sufficient time to reposition hands before the next "different" or "identical" key press was required. The fact that the right hand in general is superior in terms of speed posed no problem to the statistical analysis of the data as the accuracy of the reaction was not analysed, only the time it took react to the matching pairs. However, the discrepancy in speed between hands and among fingers is an important finding to keep in mind in designing RT instruments.

Comprehension component. In their 1992 study, Cowan and Abe employed both a math task and comprehension questions and recommended the practice of including intervening tasks in RT instruments. Bever and McElree (1988) have used comprehension questions as well as a means of keeping the comprehension process active and discourage a purely visual strategy. Takahashi and Roitblat (1992), in a study on sensitivity to literal versus intended meaning[3], preceded their input sentences with a "set-up" story. This design element was critical to assuring the correct reading of potentially ambiguous input sentences. I decided against including an intervening comprehension component for two reasons. First, I felt the task was already quite long in order to accommodate sufficient trials on the structures under investigation. The overall length of the data collection process was incremented also by the participants' required performance on two additional tasks. Second, the need for disambiguation was not a feature of the present study as it was in Takahashi and Roitblat.

---

3 For example, the difference between *Why don't you close the door?* (≈ What is the reason for your not closing the door?) and *Why don't you close the door?* (≈ Please close the door).

Feedback. Though F&F did not provide the participants with feedback regarding their on-going performance, in a later study Freedman and Forster (1987) included this practice. Testing for both conditions, Crain and Fodor obtained longer reaction times when providing no feedback. However, I chose to follow the original F&F, omitting feedback, as this would both lengthen the time required to complete the task and introduce another variable into the design. Following Eubank, a "beep" plus an on-screen error message were installed in the program to warn the participant when she pressed a key other than the blue or red one. In actual practice, such an occurrence was extremely rare past the training stage. Additionally, final statistics, though not volunteered, were furnished upon request to the participant after the task had been completed. Only seven of the participants requested the information, six English speakers and one Korean.

Scoring the data. In a matching paradigm, the input is partitioned into two sets: a positive set, and a negative set. On a trial, a participant is instructed to indicate which set an item presented is from. In the case of the present study, the pair sentences that matched the priming sentences represented the positive set, and the pair sentences that did not match represented the negative set.

Responses generate two types of data — data about reaction time and data about response accuracy. Only on items from the positive set (matching items) were included in the analysis, but these were analysed regardless of accuracy. What is of interest in here is the difficulty in processing as indicated by increased latency, not the accuracy of the response. All trials in which the pair sentences were drawn from the negative set were excluded from statistical analysis on the grounds that reaction times should be sensitive to the location the mismatch. The

mismatch could occur at any point in the sentence, even the first word; pre-experiment pilot studies indicated that reaction times to such trials elicited significantly faster reaction times than matching trials. This tendency is well documented in the RT literature (e.g. Silverman & Goldberg, 1975).

4.5. Summary of the three instruments

In this chapter, the three grammaticality evaluation instruments under investigation were examined. There was a discussion of their previous uses by language researchers and, based on the work of these researchers, speculation about what these instruments measure and the processes involved, covering levels of representation, the controlling level, and the role of interfering and competing factors. The design of each instrument was described in turn, and design considerations unique to a particular instrument were discussed.

The starting point for the comparison of the three instruments under investigation, P&P, RT, and EI, is the assumption that the sentences which constitute the set of input data either clearly conform to, or are uncontroversial syntactic violations of the language involved, as described by a widely-accepted syntactic framework. As stated before, the present study is not primarily intended as a test of the selected syntactic analysis; rather the intention is to examine the data obtained via these three instruments in terms of what the instruments actually measure. What is of particular interest to the analysis of data is the possible differentiation between two types of information to be obtained. The first type is information about the effect of grammaticality on the measurements. In the data examined here, the grammaticality differs between

the two languages. The question arises: will each instrument differentiate the set of grammatical and from the set of ungrammatical sentences for each language? The second type of information is concerned with the effect of length in chunks of the input sentences on the measurements obtained via the three instruments. Will each instrument likewise differentiate between sentences of different chunk lengths regardless of grammaticality? Will the effect for chunk length hold constant between the two languages? In the following chapter, the data collected is analysed with respect to these questions.

CHAPTER 5: PRIMARY ANALYSES OF THE DATA

The goal of this study is to contribute to the understanding of some of the functional characteristics of the three tasks under investigation. I have attempted to isolate three aspects of the input that the instruments were sensitive to, that is, aspects which produced significant effects. These were: the grammaticality of the input item, the length of the item in chunks, and language in which the task was presented.

Primary analyses were done on each of the three tasks in turn, the written questionnaire ("P&P", Section 5.2), the elicited imitation task ("EI", Section 5.3), and the reaction time task ("RT", Section 5.4). Findings are discussed with particular attention paid to evidence provided by the data of different levels of representation and the effects of interfering factors.

In all three tasks there was one dependent variable, which varies by task (Table 5–1), and three independent variables, (G)rammaticality, length in (C)hunks, and (L)anguage which remains constant across tasks.

Table 5–1: Dependent variables by task

task	*dependent variable*
P&P	Likert[1] rating of grammaticality
EI	imitation success score
RT	reaction time

Each participant responded to three tokens of each example type; the tokens differed only by vocabulary set. The responses for each example type were averaged and the mean was entered in a data table, resulting in one mean per example type per participant. A three-factor repeated measures ANOVA was done to measure the variance of the dependent variable in each case due to (G)rammaticality, (C)hunks, and (L)anguage. Means and ANOVA tables are presented in each section and significant effects are identified with particular attention paid to any significant interactions among the independent variables.

The example types 10 through 19, excluding 14, were analysed in the primary analyses. All belong to the categories 2ARG–NP and 3ARG–NP and thus vary between categories principally by the number of arguments. The length in chunks for each example type is equivalent between the languages — all types consist of either six or seven chunks. Example type 21 of the 3ARG–NP category which consists of eight was excluded. Also excluded were example types 14 and 20, even though they consist of six and seven chunks respectively. They are withheld from statistical analyses for the present because of their uncertain grammatical status in Korean (see Chapter 2, Section 2.2.4.); this pair is discussed separately in the following chapter. The MOD–NP category (example types 1

1 On a typical Likert scale, a participant indicates one of a number of response options which indicate varying degrees of agreement with or endorsement of the input stimulus. See, for example, DeVellis (1991, pp. 68–77).

through 4) was excluded in the primary analyses as the number of chunks to be assigned to English example types 3E* and 4E* of this category is not clear (see discussion in Chapter 2, Section 2.3.1.). This category is also dealt with separately in the following chapter.

For ease of reference, Table 5–2 below includes a sample of each example type within the two modification categories under analysis here (2ARG–NP, example types 10, 11, 12, 13, and 15; and 3ARG–NP, example types 16, 17, 18, and 19). A gloss for the first Korean sentence in each category is given; all remaining examples contain the same elements in various orders. The letter "E" or "K" following the sentence number indicates the language of the sample and the presence of an asterisk (*) indicates the ungrammaticality of the example type in that language. Example types 10 through 13 consist of six chunks; 15 through 19 consist of seven. In the following descriptions of the analyses, English-speaking participants are referred to as "ENGs" and Korean participants are referred to as "KORs".

Table 5–2: Sample of each example type included in the primary analyses

2ARG–NP sentences

10E They mentioned Japan's destruction of Seoul in the letter.

10K*	kutul-i	phyenci-ey	ilpon-uy	phakoy-lul	sewul-uy	iyakihayssta.
	they-NOM	*letter-LOC*	*Japan-GEN*	*destruction-ACC*	*Seoul-GEN*	*talk about-PAST-SE*

11E They mentioned Seoul's destruction by Japan in the letter.
11K* kutul-i phyenci-ey sewul-uy phakoy-lul ilpon-uy iyakihayssta.

12E They mentioned the destruction of Seoul by Japan in the letter.
12K* kutul-i phyenci-ey ku phakoy sewul-uy ilpon-uy iyakihayssta.

13E* They mentioned Japan's Seoul's destruction in the letter.
13K kutul-i phyenci-ey ilpon-uy sewul-uy phakoy-lul iyakihayssta.

↑ six-chunk sentences

↓ seven-chunk sentences

15E* They mentioned Japan's Seoul's destruction the in the letter.
15K* kutul-i phyenci-ey ilpon-uy sewul-uy phakoy-lul ku iyakihayssta.

3ARG–NP sentences

16E We described Steven's gift of flowers to Jenny to the group.

16K*	wuli-ka	salamtul-eykey	chelswu-uy	senmul-ul	pohuy-eykey-uy	kkoch-uy	kiswulhayssta.
	we-NOM	*group-DAT*	*[name]-GEN*	*gift-ACC*	*[name]-GEN*	*flower-GEN*	*describe-PST-SE*

17E We described Jenny's gift of flowers from Steven to the group.
17K* wuli-ka salamtul-eykey pohuy-ey-key-uy senmul kkoch-uy chelswu-uy kiswulhayssta.

18E We described the gift of flowers from Steven to Jenny to the group.
18K* wuli-ka salamtul-eykey senmul-ul chelswu-uy pohuy-eykey-uy kkoch-uy kiswulhayssta.

19E* We described Steven's to Jenny of flowers gift to the group.
19K wuli-ka salamtul-eykey chelswu-uy pohuy-eykey-uy kkoch-uy senmul-ul kiswulhayssta.

5.1. Research questions leading to the primary analyses

In order to obtain the information relevant to answering the following questions raised about the use of the three tasks, the present experiment was designed so that the three aspects of the input that are of interest — grammaticality, length in chunks, and languages of the participant could be investigated.

Effects of (G)rammaticality on:

P&P: Did grammatical sentences receive a higher Likert rating than ungrammatical sentences?

EI: Were grammatical sentences more often imitated successfully than ungrammatical sentences?

RT: Were reaction times on grammatical sentences faster than reaction times on ungrammatical sentences?

Effects of Length in (C)hunks on:

P&P: Did six-chunk sentences receive a higher Likert rating than seven-chunk sentences?

EI: Were six-chunk sentences more often imitated successfully than seven-chunk sentences?

RT: Were reaction times on six-chunk sentences faster than reaction times on seven-chunk sentences?

Effects of (L)anguage on all tasks:

Did the two language groups perform differently from one another on any of the tasks?

5.2. Pencil-and-Paper (P&P) task

There were three different randomized versions of the P&P instrument, each consisting of 63 sentences to be judged. The 63 sentences included three tokens of each of the 16 example types from the three modification categories under investigation (MOD–NP, 2ARG–NP, and 3ARG–NP) plus three tokens of

each of the five example types from the modification category POS–NP; data on the latter category was not included in the present study. There were no repeats among sentences either within a version or across versions. Each of the participants (ENG n=49; KOR n=44) rated each of the sentences in his or her native language for grammaticality on a five-point Likert scale, "5" representing "good", "1" representing "bad".

5.2.1. P&P data analysis

There were significant independent effects for grammaticality, chunks, and language as well as significant interactions between grammaticality and language and grammaticality and chunks. There was no interaction between chunks and language nor among grammaticality, chunks, and language. Each of the significant interactions are examined separately. The means are shown in Table 5–3; ANOVA results are shown in Table 5–4.

Table 5–3: P&P means table

	English		*Korean*	
	6 chunks	7 chunks	6 chunks	7 chunks
grammatical	4.77	4.30	3.36	2.55
ungrammatical	1.50	1.33	2.31	2.14

Table 5–4: P&P: Analysis of variance of Likert rating by (G)rammaticality, length in (C)hunks and (L)anguage

Source	df	Type III SS	MS	F-value	P-value
(G)ram.	1	326.68	326.68	859.5	**.0001**†
(C)hunks	1	13.15	13.15	51.54	**.0001**†
(L)anguage	1	15.69	15.69	18.13	**.0001**†
G x C x L	1	.32	.32	1.28	.2614
G x C	1	3.83	3.83	15.08	**.0002**†
G x L	1	1.35	1.35	3.54	**.0001**†
C x L	1	.32	.32	1.26	.2641

†Probability is significant ($p<.01$)

5.2.2. P&P: Interaction of grammaticality by chunks

Because there was a significant interaction of grammaticality and chunks, irrespective of language, the means for this interaction are shown in Table 5–5, along with the overall means for the independent effects for grammaticality and chunks.

Table 5–5: P&P: Grammaticality and chunks means table

	6 chunks	7 chunks	*Mean*
grammatical	4.06	3.48	3.78
ungrammatical	1.89	1.71	1.80
Mean	2.97	2.60	

Grammatical sentences were rated higher than ungrammatical sentences by all participants (3.78 for grammatical sentences versus 1.80 for ungrammatical

sentences). Sentences of six chunks in length were rated higher than sentences of seven chunks in length by all participants, regardless of grammaticality (2.97 for six-chunk sentences versus 2.60 for for seven-chunk sentences). Figure 5–1 represents the interaction between grammaticality and chunks. While the Likert rating was higher in general on the six-chunk sentences than on the seven chunk sentences, there was a greater difference for the grammatical sentences. This is shown in Figure 5–1. On a post hoc comparison test, the difference between the Likert ratings for grammatical sentences between six- and seven-chunk sentences was significant (t=7.83, p=.0001), but the difference between the ratings for ungrammatical sentences between six- and seven-chunk sentences was not (t=2.35, p=.0212). This shows that the P&P instrument distinguished chunks for grammatical sentences but not for ungrammatical sentences.

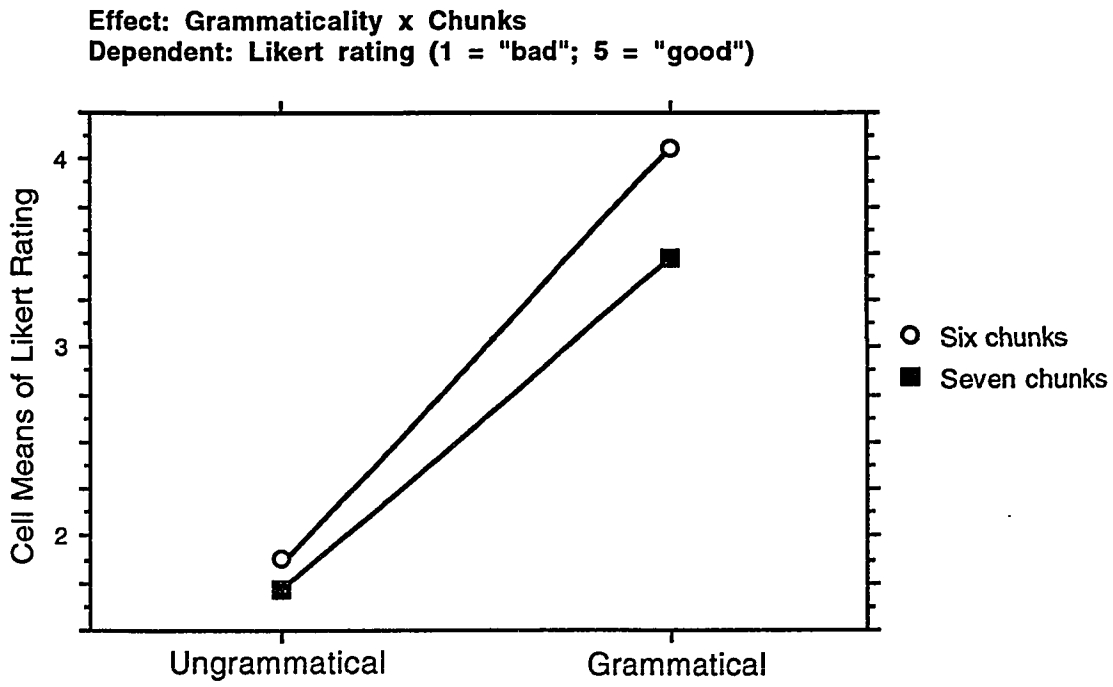


Figure 5–1: P&P: Grammaticality by chunks interaction plot

5.2.3. P&P: Interaction of grammaticality by language

Because there was a significant interaction of grammaticality and language, irrespective of chunks, the means for this interaction are shown in Table 5–6, along with the overall means for the independent effects for grammaticality and language.

Table 5–6: P&P: Grammaticality and language means table

	English	Korean	*Mean*
grammatical	4.53	2.90	3.78
ungrammatical	1.42	2.22	1.80
Mean	2.98	2.57	

As shown above, grammatical sentences were rated higher overall than ungrammatical sentences. As for the variable of language, ENGs rated sentences significantly higher overall than did the KORs (2.98 for ENGs versus 2.57 for KORs). While both language groups rating grammatical sentences higher than ungrammatical sentences, the ENG Likert mean ratings for grammatical and ungrammatical occupied a larger section of the possible 1 to 5 range than did the KOR mean ratings, i.e., ENGs rated grammatical sentences higher and ungrammatical sentences lower than did the KORs. This interaction is shown in Figure 5–2.

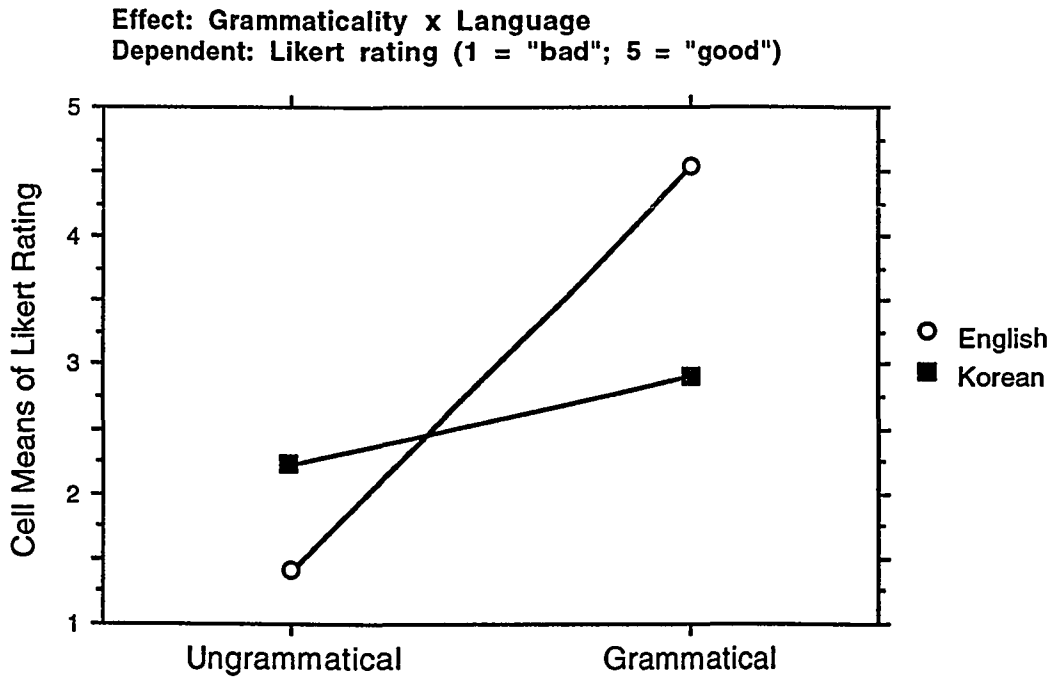


Figure 5–2: P&P: Grammaticality by language interaction plot

5.2.4. Summary and discussion of P&P results

The P&P instrument was shown to be sensitive to both grammaticality and, in certain circumstances, to length in chunks across both language groups. This was true for each language group individually as well (see Table 5–15 in the final summary, Section 5.5). The ENGs rated sentences significantly higher overall than did the KORs; additionally, their ratings for the sentences ranged more widely over the 1-through-5 scale provided. However, while P&P distinguished grammaticality at both the six- and seven-chunk level, it distinguished between six and seven chunks only for grammatical items. Thus, the results from the analysis of the P&P instrument indicate it was a more sensitive instrument for the detection of grammaticality than for chunks.

Evidence of levels of representation. The effect for length in chunks, though present, was not as strong as the effect for grammaticality on the P&P task; nevertheless, it seems curious that there should be an effect at all. When asked to judge sentences solely on grammaticality, the participants showed a tendency, upon encountering an increase in chunks to register a decrease in acceptability, though only for the set of grammatical items. This might be less unexpected if the differences in length among input sentences were extreme, but the fact that the P&P instrument detected differences between grammatical sentences of six and seven chunks should raise a red flag for researchers using the P&P task to make claims about the comparative grammaticality of sentences that differ in length.

Possible interfering factors. Ellis (1991) objected to P&P-type tasks on the grounds that success is based on test-performing strategies and thus performance is strongly tied to certain types of education or training. This claim may find support in the results of the present P&P experiment. Neither group was given more than a cursory explanation of the purpose of the experiment — participants were simply told that the researchers were interested in how native speakers felt about certain sentences in their language.

I felt the KORs appeared to be thoroughly unfamiliar with this type of testing and, judging by the types of questions they asked, less familiar than the ENGs with general concepts of psychological testing and the uses of the resulting data. This inequity may account in part for the apparent reluctance on the part of the KORs to venture into the extreme ends of the P&P Likert scale. Information from the participants themselves as to their perceptions of the purpose of the task and of their strategies might have provided helpful, unfortunately, post-task introspective interviews were not a component of the present study. Interviews

or other introspective components (e.g., Kellerman, personal communication (participants' self-critiques on video playbacks of their performances on a task); Marquer & Periera, 1990 (verbal reports by participants of their strategies); Lantolf, Goss, & Zhang, 1991 (dyadic-problem-solving)) can be extremely revealing, uncovering unexpected strategies and attitudes on the part of a participant. Also, as Vetter, Volovecky, and Howell (1979) found, in a replication of a study by MacLay and Streator (1960), there exists a wide variety of explanations and justifications for evaluations, some based on linguistic intuition, many others, not (also Ellis, 1990 (think-aloud task)). However, I had felt that, given the amount of time and effort required of the participants to complete the three-task battery, such an interview would have constituted an uncalled-for additional burden upon them. I now regret this decision.

Also, unfortunately, there were unequal treatments between language groups for administration of P&P task. If it is difficult to duplicate conditions of data collection among participants or on successive days, it is doubly difficult to do so for different languages in different countries. While the ENGs completed the P&P questionnaire individually, because of time constraints, this task was administered to all KORs at the same time in their homeroom class. The participants all sat very close to one another, allowing for possible covert collaboration between participants. Additionally, as the faster participants finished, they disturbed those who were still working, providing a type of pressure and distraction which was not part of the ENG collection.

5.3. Elicited Imitation (EI) task

There were three different randomized pre-recorded versions of the EI instrument, each consisting of 63 sentences to be judged. The 63 sentences included three tokens of each of the 16 example types from the three modification categories under investigation (MOD–NP, 2ARG–NP, and 3ARG–NP) plus three tokens of each of the five example types from the modification category POS–NP; data on the latter category was not included in the present study. There were no repeats among sentences either within a version or across versions. Each of the participants (ENGs n=49; KORs n=44) heard an elicitation in his or her native language and was allowed unlimited time to imitate it. In general, if the participant did not initiate a response within three or four seconds, likelihood of a correct response was very slight. Responses were tape-recorded and transcribed. The responses were assigned a score of either "1" or "0" according to the procedure described in Chapter 4, Section 4.3.4. under "Scoring the data".

5.3.1. EI data analysis

There were significant independent effects for grammaticality, chunks, and language as well as significant interactions between grammaticality and chunks, grammaticality and language, and among all three independent variables: grammaticality, chunks, and language. Only chunks and language showed no significant interaction. Each of the significant interactions are examined separately. The means are shown in Table 5–7; ANOVA results are shown in Table 5–8.

Table 5–7: EI means table

	English		*Korean*	
	6 chunks	7 chunks	6 chunks	7 chunks
grammatical	.721	.432	.428	.023
ungrammatical	.474	.172	.157	.036

Table 5–8: EI: Analysis of variance of Likert rating by (G)rammaticality, length in (C)hunks, and (L)anguage

Source	df	Type III SS	MS	F-value	P-value
(G)ram.	1	3.40	3.40	112.57	**.0001**†
(C)hunks	1	7.24	7.24	182.95	**.0001**†
(L)anguage	1	7.73	7.73	72.04	**.0001**†
G x C x L	1	.51	.51	26.08	**.0001**†
G x C	1	.43	.43	21.96	**.0001**†
G x L	1	.36	.36	11.94	**.0008**†
C x L	1	.02	.02	.58	.4466

†Probability is significant (p<.01)

5.3.2. EI: Interaction of grammaticality by chunks by language

The interaction among all three independent variables (represented in Figure 5–3) was evidently due to the performance of the KORs on seven chunk sentences (23% for the grammatical sentences versus 36% for the ungrammatical sentences). This is a good example of a "floor effect". While their performance on the six-chunk sentences conformed to the pattern of the ENGs on all sentences, on the seven-chunk sentences their rate of success was so low as to be unaffected by grammaticality. Thus, the performance of the KORs on the six-chunk

sentences and the overall performance of the ENGs were consistent, i.e., there were significant independent effects for both grammaticality and chunks.

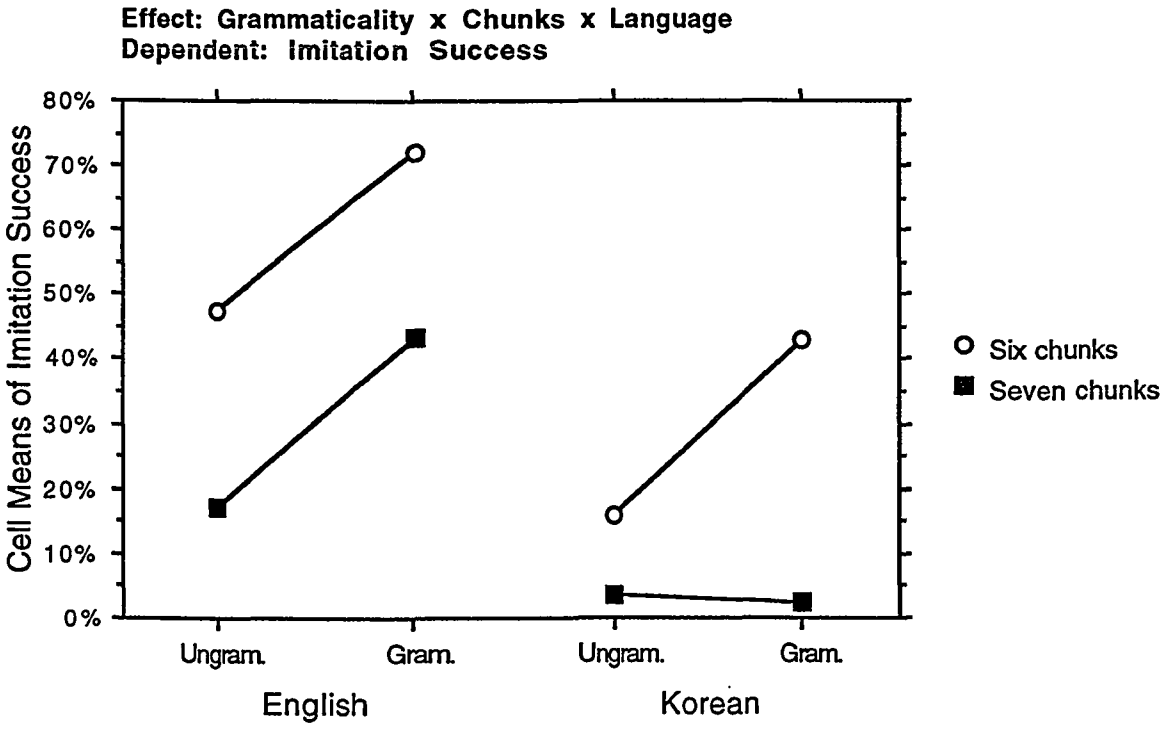


Figure 5–3: EI: Grammaticality by chunks by language interaction plot

5.3.3. EI: Interaction of grammaticality by chunks

There was a significant interaction of grammaticality and chunks, irrespective of language. The means for this interaction are shown in Table 5–9, along with the overall means for the independent effects for grammaticality and chunks.

Table 5–9: EI: Grammaticality and chunks means table

	6 chunks	7 chunks	*Mean*
grammatical	.583	.238	.410
ungrammatical	.324	.108	.216
Mean	.453	.173	

Not only were the participants more successful in imitating grammatical over ungrammatical sentences (41% rate of success for the grammatical sentences versus 21.6% rate of success for the ungrammatical sentences), but they were also more successful on the six- over the seven-chunk sentences (45.3% rate of success for six-chunk sentences versus 17.3% rate of success for the ungrammatical sentences). However, the spread between the success rates between grammatical and ungrammatical on the six-chunk sentences was wider than the spread on the seven-chunk sentences. This was due primarily to floor effects caused by the performance of the Korean on the seven-chunk sentences. The interaction is shown in Figure 5–4.

On a post hoc comparison test, the difference between grammatical and ungrammatical scores for both six- and seven-chunk sentences was significant ($t = -12.65$, $p = .0001$ and $t = -6.02$, $p = .0001$, respectively). This shows that the EI instrument distinguished chunks for both grammatical and ungrammatical sentences, with a greater difference for grammatical sentences.

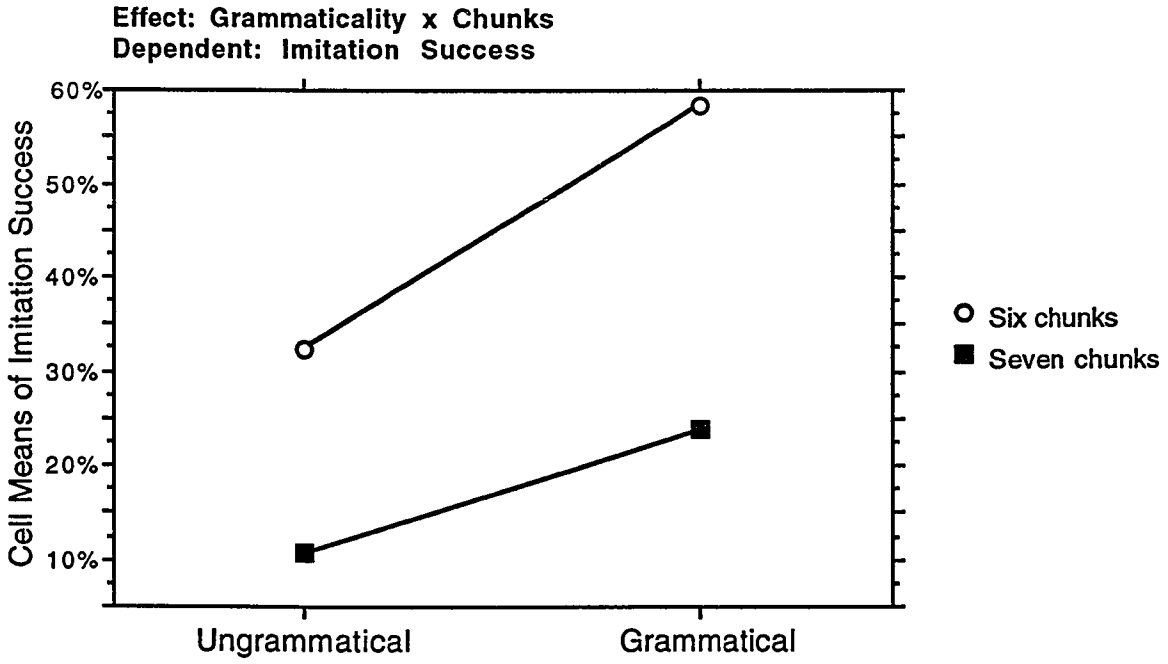


Figure 5–4: EI: Grammaticality by chunks interaction plot

5.3.4. EI: Interaction of grammaticality by language

There was a significant interaction of grammaticality and language, irrespective of chunks. The means for this interaction are shown in Table 5–10, along with the overall means for the independent effects for grammaticality and language.

Table 5–10: EI: Grammaticality and language means table

	English	Korean	*Mean*
grammatical	.577	.226	.410
ungrammatical	.323	.096	.216
Mean	.450	.161	

Although, as shown above, grammatical sentences were more often correctly imitated by all participants than ungrammatical sentences, the ENGs overall were more successful than the KORs on all sentences (overall 45% rate of success for the ENGs versus 16.1 % rate of success for the KORs). Not only were the ENGs more successful than the KORs in general, but there was a greater difference between grammatical and ungrammatical sentences. This is shown below in Figure 5–5.

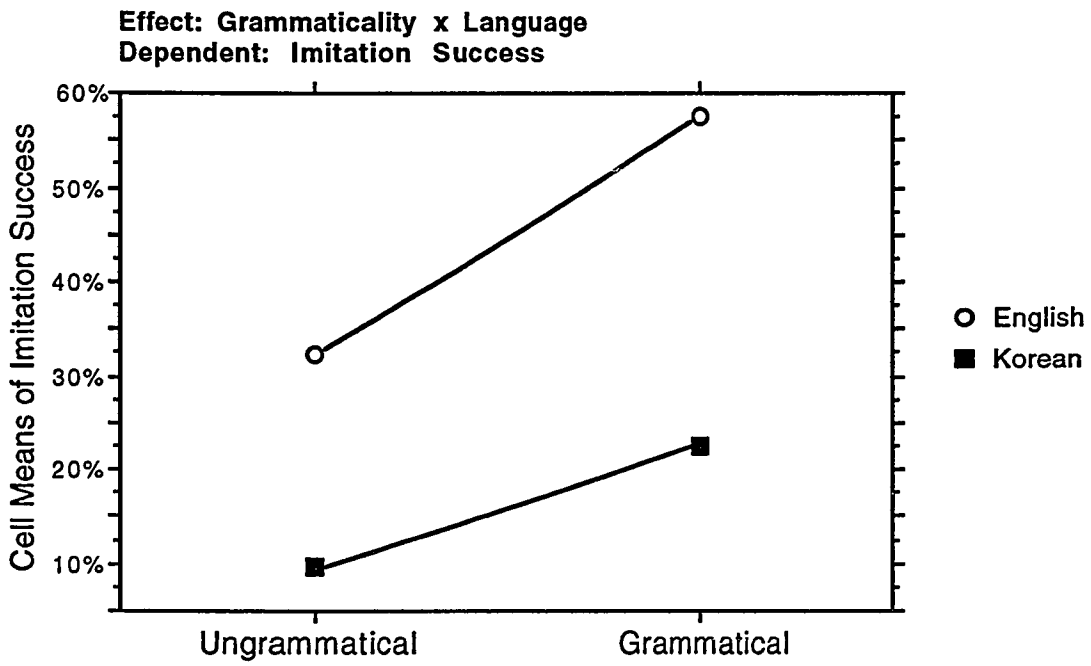


Figure 5–5: EI: Grammaticality by language interaction plot

5.3.5. Summary and discussion of EI results

In summary, the EI instrument was shown to be sensitive to both grammaticality and to chunks across both language groups. This is true for each language group individually as well (see Table 5–15 in the final summary, Section 5.5). Overall, the ENGs were more successful than the KORs on all

sentences. In addition, for the KORs, sensitivity to grammaticality diminished as length in chunks increased; there was no effect for grammaticality among the seven-chunk sentences for this group. The overall results from the analysis of the EI instrument show very similar effects for both grammaticality and chunks regardless of language.

Evidence of levels of representation. Although the present study does not include in-depth error analysis of the data generated by the EI instrument, even a cursory look at the types of errors yields anecdotal evidence for some of the processes at work. There was found, for example, evidence that processing of the elicitation was sensitive to syntactic constraints. The tendency to "correct" ungrammatical imitations has been documented in L1 research by O'Connell and Gerard (1985), although they found that the tendency to correct scrambled structures abates with age, e.g., 36-month-old children were able to repeat scrambled structures that 24-month-old children could not repeat (through the 24-month-olds could repeat canonical orders), and that 28-month-olds corrected. Many of the participants in this study, both ENGs and KORs, corrected the elicitation (E=elicitation; I=imitation), providing evidence that processing took place at a high, perhaps S-structure, or even interpretive level:

(1)

E: They recommended Mary's to Robert of money the donation in the letter
I: They recommended Mary's donation of money to Robert in the letter
(tape 7A; participant ENG38)

There was evidence as well of processing at the level of individual constituents or phrases. There were a number of instances in both languages of grammatical shuffling of phrases:

(2)
E: They recommended the donation of money from Mary to Robert in the letter
I: They recommended the donation of money to Robert from Mary in the letter
(tape 3B; participant ENG17)

In addition, as shown by the following example, the phrase *to the group*, a phrase which appeared in the previous elicitation, substitutes for another prepositional phrase, *at home*:

(3)
E: I described Steven's photo of Linda **to the group**
I: I described Steven's photo of Linda **to the group**
followed immediately by:
E: They discussed Linda's to Richard of a job offer the **at home**
I: They discussed Linda's offer of a job **to the group**
(tape 4A; participant ENG21)

As mentioned above, lexical substitution was accommodated in the scoring procedure. There was a fair amount of this, in both English and Korean. The following example is a single continuous group of responses from one of the ENGs. Note that both nouns and verbs were substituted with previously occurring items of the same grammatical category.

(4)

E: They arrived at the shop of **Jenny's** neighbor before noon
I: They arrived at the shop of Jenny's neighbor before noon

E: We **described** the donation of money from Mary to Robert to the group
I: We described the donation from Mary to Robert to the group

E: They recommended Linda's to **Richard** of a job the offer in the letter
I: They recommended of a job of Richard to **Jenny** in the letter

E: We **discussed** the the **photo** of Linda by Steven at home
I: We **described** the photo of Linda to **Richard** in a **photo** at home

E: They mentioned the destruction of Seoul by Japan in the letter
I: They **discussed** the destruction of Seoul to Japan in the letter

(tape 5B; participant ENG30)

The influence of previously encountered vocabulary even carried across tasks. The participant in the following example substituted the word *described* for *discussed* twice consecutively. Though *described* had appeared on the RT task a few minutes before, it had not yet occurred in the input to the EI task:

(5)
E: They **discussed** Steven's to Jenny of flowers gift the at home
I: They **describ- described** [laugh] Je-Jenny's the flower gift at home

E: They **discussed** Linda's to Richard of a job the offer at home
I: They **described** Jenny's to Richard the offer at home
(tape 7A; participant ENG38)

In addition to lexical substitution using lexical items from the vocabulary sets used in the task, many participants substituted words that, though not from the task, were semantically related. For example, one Korean participant consistently substituted *swuphamakhaytu* 'supermarket' for *kakey* 'shop'. Among

substitutions by ENGs were *picture* instead of *photo* and *after the test* in place of *after the class*. Even more intriguing were semantically-related substitutions that crossed grammatical categories, again indicating a high level, interpretive processing of the input. Note, for example, the following noun-to-verb and verb-to-nouns transformations:

(6)

NOUN => VERB

E: We described Linda's to Richard of a job **offer** to the group

I: We **offered** Linda to David job offer in the group

(tape 3A; participant ENG15)

VERB => NOUN

E: They **recommended** Mary's donation of money to Robert in the letter

I: They discussed the **recommendation** of money to Jenny in the letter

(tape 6B; participant ENG34)

Many of these "errors" lend credence to the claim that an interpretive level of processing is accessed in formulating an imitation. Perhaps the most telling evidence for processing at an interpretive level is the following: although four of the 49 ENGs switched the arguments *Japan* and *Seoul* in the sentences involving *Japan's destruction of Seoul,* none of the KORs did. I attribute this to the fact that KORs are more likely than the ENGs to know, and have strong feelings about the historical background[2] of the events to which these sentences refer:

(7)

E: They discussed the **destruction of Seoul by Japan** at home

I: They discussed the **destruction of Japan by Seoul** at home

(tape 9A; participant ENG47)

2 Marshall Pihl (personal communication) observes that there has not actually been any destruction of Seoul by Japan in recent history — not since the late 16th century and that perhaps this may have caused some confusion for the Korean students.

Possible interfering factors. As shown in Section 5.3.4., the ENGs were more successful in correctly imitating the sentences (45% rate of success for the ENGs versus 16.1 % rate of success for the KORs); in addition, there were strong floor effects for KORs on seven-chunk sentences. The question arises as to what factor or factors can account for this discrepancy. The two most obvious sources of difference between the two groups are the language of the task and the participants themselves.

The Korean sentences, though comparable in length to the English sentences when figured in chunks and actually containing fewer words, were nevertheless longer in terms of syllables. The elicitation tape for the Korean EI instrument was longer by about three minutes due to the fact that it took the speaker longer to say each of the 63 sentences. Thus, the real time it takes to produce the elicitation may effect the participant's ability to form and store a representation in STM.

The second source of difference between the two groups is the character of the participants themselves. As described in Chapter 4, Section 4.1.2. the KORs were younger, had received less formal education (although it is difficult to make comparisons between Korean and US educational systems), and probably less positively motivated than were the ENGs. The possible effects of age, fatigue, anxiety, (Section 5.4.5. under "Possible interfering factors") and level of education (Section 5.5.2.) are discussed with reference to all three tasks in detail below.

In the present study, the input sentences were constructed in such a way to control for recency/primacy effects, that is, in all the ungrammatical sentences, the source of the ungrammaticality occurred in the same sentence-medial location. Thus, the EI instrument generated no data by which to investigate these

effects (though they have been documented in previous studies as discussed in Chapter 4, Section 4.3.3.2.). However, I wish to point out here the danger of ignoring the effects of both recency/primacy effects and length in chunks in analysing data obtained via an EI instrument by citing a recent study by Munnich, Flynn, & Martohardjono (1992). In this study, Munnich, et al. measured and compared participants' performance on task sentences with relative clauses using both an EI task and a grammaticality judgment task. The relative clause structures were either Object-Object (OO) or Object-Subject (OS); there both grammatical and ungrammatical examples of each — the examples supplied in the study are repeated here.

grammatical sentences

A. Object-Object (OO)
The engineer answered the *gentleman* who the owner shoved *ø*.

B. Object-Subject (OS)
The diplomat answered the *man* who *ø* introduced the doctor.

ungrammatical sentences

C. Object-Object (OO)
*The man greeted the *doctor* who the actor introduced *him*.

D. Object-Subject (OS)
*The lawyer questioned the *student* who *he* answered the owner.

On an EI task, the L2 participants were found to more often omit the ungrammatical sentence-medial elements of the OS violations and thus "correct" the sentence than to perform this operation on OO violations. Based on this, Munnich, et al. claim that participants exhibited a "significant preference" for ungrammatical OS sentences, however, as the Chapter 4 discussion of

recency/primacy may indicates, rather than an increased sensitivity to sentence-medial errors, the opposite might actually be the case. It may be an increased *in*sensitivity to errors occurring sentence-medially that account for the differences between the performance on ungrammatical OS and OO sentences.

In addition, in the four examples of the stimulus sentences provided, the grammatical sentences were six chunks in length and the ungrammatical ones were seven. The results of the present data collection and analysis, precisely on sentences on six and seven chunks, can account for significant differences found between the grammatical and ungrammatical sentences without referring to their grammaticality at all. The effect for length in chunks alone can account for significant differences in imitation success.

5.4 . Reaction Time (RT) task

The input for the RT task consisted of 63 randomly ordered pairs of sentences; 48 of these pairs included three tokens of each of the 16 example types from the three modification categories (MOD–NP, 2ARG–NP, and 3ARG–NP; the POS–NP category was omitted entirely from the RT instrument). The 15 remaining pairs of sentences were unrelated distractors including two that were used as fillers at the beginning of the task; these were also excluded from analysis.

Each pair consisted of either a sentence plus an identical repeat (match) or sentence plus a repeat that differed by one word, always a noun (mismatch). Data from the negative set (consisting of the mismatch pairs) was not analysed — the total of data-generating pairs was 31 from the positive set. There was at least

one positive set pair for each example type in the English input set; unfortunately, positive set pairs of example types 12 and 17 were inadvertently excluded from Korean input, thus no usable data was generated on those two types.

Following a practice suggested by Snodgrass, Levy-Berger, and Hayden (1985, p. 52), reaction times that were more than two standard deviations from the mean for any participant were discarded. There were no repeats between pairs. Each of the participants (ENG n=49; KOR n=44) judged the 63 pairs and their reaction times were automatically recorded in Macintosh "ticks" (1 tick = 1/60 sec. or 16 msec.).

5.4.1. RT data analysis

In the RT data, there were significant independent effects for chunks and language, but not for grammaticality. There were less strong, but nevertheless significant effects for the interaction between grammaticality and chunks and among all three independent variables: grammaticality, chunks, and language. There was no interaction between grammaticality and language nor between chunks and language. The means are shown in Table 5–11; ANOVA results are shown in Table 5–12.

Table 5–11: RT means table

	English		*Korean*	
	6 chunks	7 chunks	6 chunks	7 chunks
grammatical	165.97	178.95	201.54	238.04
ungrammatical	155.18	198.18	198.98	233.79

Table 5–12: RT: Analysis of variance of Likert rating by (G)rammaticality, length in (C)hunks, and (L)anguage

Source	df	Type III SS	MS	F-value	P-value
(G)ram.	1	15.57	15.57	.02	.9023
(C)hunks	1	93891.90	93891.90	81.94	**.0001**†
(L)anguage	1	175613.50	175613.50	38.07	**.0001**†
G x C x L	1	5820.80	5820.80	6.64	**.0116**†
G x C	1	4653.37	4653.37	5.31	**.0235**†
G x L	1	1346.92	1346.92	1.31	.2554
C x L	1	1360.41	1360.41	1.19	.2788

†Probability is significant (p<.01)

5.4.2. RT: Interaction of grammaticality by chunks by language

The lack of significance for grammaticality (p=.9023) should be interpreted in light of the interaction of grammaticality by chunks by language. The effects for grammaticality were not consistent across categories. The interaction between grammaticality and chunks (Table 5–13 below) was due to the fact that while, as expected, participants took less time to react to grammatical seven-chunk sentences than to ungrammatical seven-chunk sentences (215.03 ticks versus 206.91 ticks), the results were the opposite for the six-chunk sentences (175.91 ticks for grammatical versus 182.80 ticks for ungrammatical). This unexpected result holds true for both the ENGs (165.97 ticks for grammatical versus 155.18 ticks for ungrammatical) and the KORs (201.54 ticks for grammatical versus 198.98 ticks for ungrammatical). Note, however, that as shown on Table 5–11 above, the expected result on the seven-chunk sentences was due to the performance of the ENGs only (178.95 ticks for grammatical versus 198.18 ticks on ungrammatical). Again, contrary to expectation, the pattern for the KORs on

seven-chunk sentences, though not significant, was the opposite (238.04 ticks for grammatical versus 233.79 ticks for ungrammatical). While the difference was not significant for the KORs on any of the sentences (six-chunk: t= –.60, p=.5526; seven-chunk: Korean t= –.36, p=.7193), the difference was significant between the ENGs' reaction times for grammatical and ungrammatical seven-chunk sentences (t=3.67, p=.0006.) — the ENGs reacted more quickly to grammatical seven-chunk sentences than to ungrammatical ones. Surprisingly, the ENGs reacted in the opposite manner on the six-chunk sentences, that is, they reacted more quickly to ungrammatical six-chunk sentences than to grammatical ones. However, the difference was not significant (t= –2.06, p=.0452;). This interaction is shown in Figure 5–6.

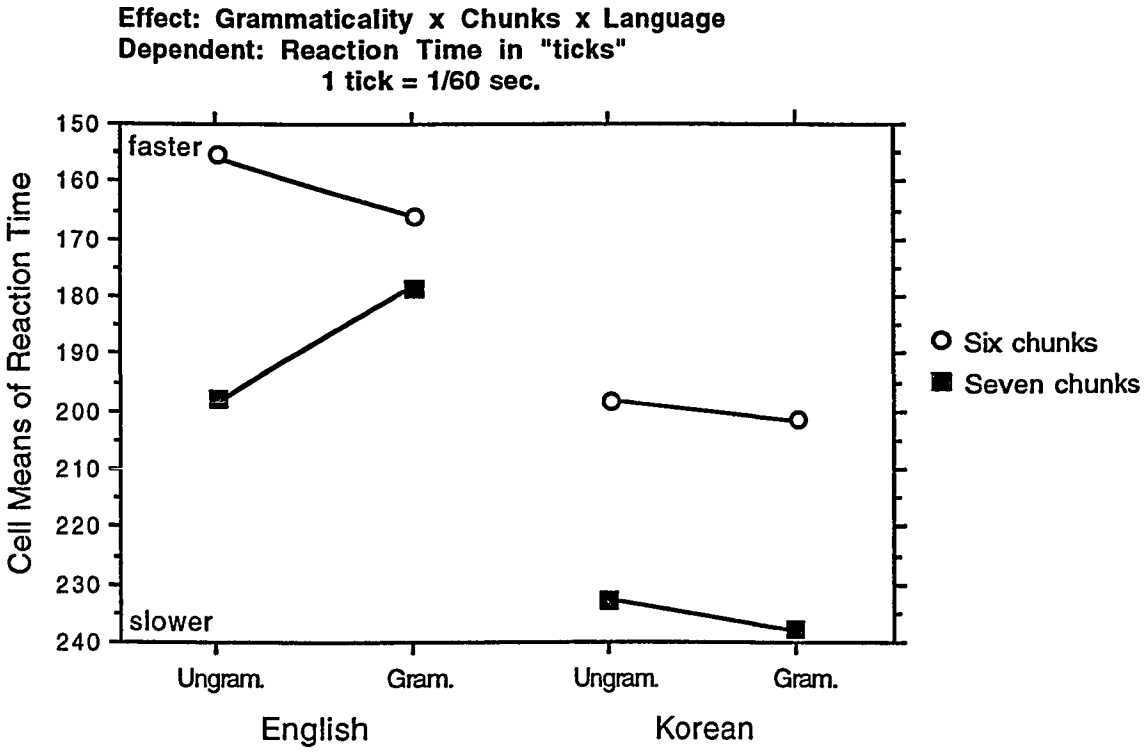


Figure 5–6: RT: Grammaticality by chunks by language interaction plot

In other words, the only significant grammaticality by chunks interaction found was that of the ENGs (p=.0002 versus p=.8678 for the KORs). On the seven-chunk sentences, where significant differences were found, the ENGs performance conformed to expectations, i.e., other things being equal, reaction time was faster for grammatical than for ungrammatical sentences. The overall conclusion to be drawn here is that RT was a generally ineffective measure of grammaticality.

5.4.3. RT: Interaction of grammaticality by chunks

There was a significant interaction of grammaticality and chunks, irrespective of language; the means for this interaction are shown in Table 5–13, along with the overall means for the independent effects for grammaticality and language.

Table 5–13: RT: Grammaticality and chunks means table

	6 chunks	7 chunks	*Mean*
grammatical	182.80	206.91	194.85
ungrammatical	175.91	215.03	195.47
Mean	179.35	210.97	

There was no significant independent effect for grammaticality (194.85 ticks ticks for the grammatical sentences versus 195.47 ticks for the ungrammatical sentences); the mean reaction time for all participants was 195.16. While the reaction time on seven-chunk sentences was faster for grammatical than for ungrammatical sentences (206.91 ticks ticks for the grammatical sentences versus 215.03 ticks for the ungrammatical sentences), the opposite effect held for the six

chunk sentences (182.80 ticks ticks for the grammatical sentences versus 175.91 ticks for the ungrammatical sentences). This interaction is shown in Figure 5–7.

On a post hoc comparison test, the difference between the reaction times for neither six- nor seven-chunk sentences between grammatical and ungrammatical sentences was found to be significant (t= –1.53, p=.1284 and t=1.72, p=.0883, respectively). This shows that the RT instrument did not distinguish grammaticality at either chunk level.

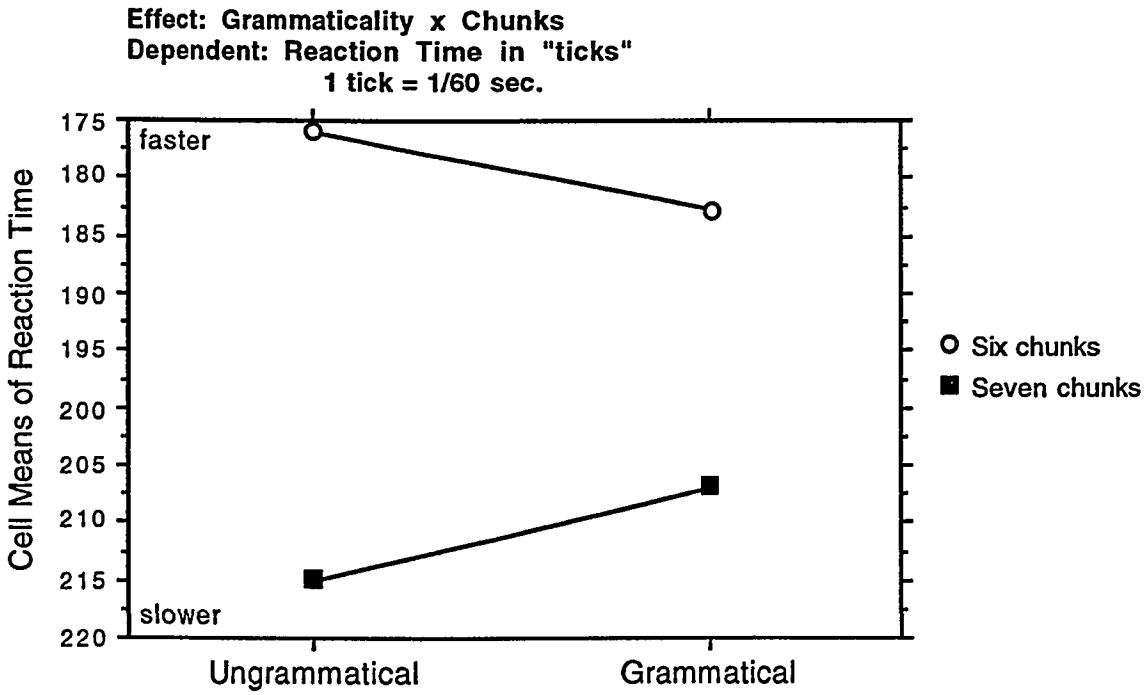


Figure 5–7: RT: Grammaticality by chunks interaction plot

5.4.4. RT: Interaction of chunks by language

While there were significant independent effects for chunks and for language, there was no significant interaction between these variables. The means for these are shown in Table 5–14 to make a point about the effectiveness

of the RT instrument in measuring length in chunks. As shown in the Table, the ENGs were significantly faster on RT than were the KORs (174.57 mean reaction time in ticks for the ENGs versus 218.09 mean reaction time in ticks for the KORs). Nevertheless, it is clear that all participants reacted significantly much more quickly to the six chunk sentences than to the seven chunk sentences (179.35 ticks for the six-chunk sentences versus 210.97 ticks for the seven-chunk sentences). There was no significant effect of language on chunks.

Table 5–14: RT: Chunks and language means table

	English	Korean	*Mean*
6 chunks	160.58	200.26	179.35
7 chunks	188.57	235.91	210.97
Mean	174.57	218.09	

5.4.5. Summary and discussion of RT results

The main significant effects on the RT instrument were for length in chunks. The participants, regardless of their native language (see Table 5–15 in the final summary, Section 5.5) or the grammaticality of the sentences, were able to react more quickly to six-chunk sentences than to seven-chunk sentences. Even though there was a significant interaction between grammaticality and chunks, the difference between the reaction times for grammatical and ungrammatical at either chunk level was not found to be significant. The only significant effects found for grammaticality were those for ENGs, and only on the seven-chunk sentences. In addition, the two language groups performed differently; the ENGs were faster overall than in the KORs.

Thus, the RT instrument, as designed, accessed a narrow band of sensitivity within which at least some aspect of the input, specifically, length in chunks, produced significant effects. The overall results from the analysis of the RT instrument suggest that it may be a suitable for the detection of length in chunks, but not for grammaticality, as the following example demonstrates. The RT measure did not distinguish between example types 10E and 14E*, which differ in grammaticality, but not chunks (mean reaction time on 10E=149.31; mean reaction time on 14E*=158.27; p=.9010). However, example types 14E* and 15E*, which differ by number of chunks, but not grammaticality, are distinguished (mean reaction time on 14E*=158.27; mean reaction time on 15E*=197.16; p=.0001). The more chunks, the longer the reaction time.

		gram.	chunks
10E	*I / described / Steven's / photo / of Linda / to the group*	+	6
14E*	*I / described / Steven's / Linda's / the photo / to the group*	–	6
15E*	*I / described / Steven's / Linda's / photo / the / to the group*	–	7

Evidence of levels of representation. In the Chapter 4, Section 4.4.3.1. discussion of RT the following conclusion was reached — that "ungrammatical sentences of the rule-violation type should require more processing time in the RT task as it has been designed." The data from the present study did not bear out this conclusion, however. Effects for the grammaticality of the sentences were either not significant (as in the case of the KORs) or contradictory (opposite effects according to length in chunks for the ENGs).

What can account for these results? One possibility is that the syntactic analysis selected does not reflect the underlying competence of the grammar. Yet, significant results for grammaticality were obtained by both the P&P and EI

instruments on the same data with the same participants. Another possibility is that the instrument was badly designed. Perhaps RT would be sensitive to grammaticality on input sentences that are either longer or shorter, in other words, maybe this particular RT instrument did not successfully access the narrow band of sensitivity within which significant effects for grammaticality were in evidence. A third possibility is that there were interfering factors that simply overcome the instrument's sensitivity to what was being tested — these are touched on below.

One must not make the claim, however, that non-significance for grammaticality, can be equated with insensitivity to syntactic structure, for it is the analysis of the syntactic structure that yields the number of chunks, for which there was a significant effect. Perhaps it would be more reasonable to suspect that RT did not access a level of representation high enough to include a check on the viability of attachment (see Chapter 2, Section 2.3.2.) at the CP level. This claim is supported by findings of Eubank (in press), discussed below, and is also in line with Bley-Vroman and Chaudron's (in press) suggestion that the phrasal level may act as the controlling level on EI tasks.

Possible interfering factors. Again, as with the previous two tasks, the performance of the KORs differed significantly from that of the ENGs; in the case of RT, the KORs took more time to react. As far as possible, interfering factors such as use of different fingers and hands, environmental distractions, and anxiety due to poor understanding of the task or lack of pre-task practice were controlled for.

Welford (1980) discussed the sensitivity of reaction time to the factors of age, fatigue, anxiety (or stress). The majority of studies done on RT and age are done on participants over 20 years old; of the studies cited in Welford (1977),

those that included participants between the ages of 15 and 20 lumped them together in categories with an upper limit of from 34 to 40, thus there are no available statistics regarding any possible differences between 18-year olds (the Korean participants' mean age) and 23-year olds (the English speaking participants' mean age). In the majority of the studies cited, effects for age do not show up until after the age of 40.

Results of research (Singleton, 1953) summarized by Welford demonstrate that fatigue showed up as an overall slowing of reaction times, which can be attributed in part to the increasing occurrence of *blocks* (Bills, 1931) on a long task. A *block* is an unusually long reaction time in an otherwise rapid performance. In contrast, Eubank (1991) reports the opposite relationship between reaction time and order of presentation: the participants had a tendency to speed up near the end. It is possible that unfamiliarity with task would show up as a higher initial error rate or that fatigue would show up as increased inaccuracy near the end of the task. However, correlations calculated on the present data for reaction time and order (R^2=.01) and for reaction time and error (R^2=.05) showed no significance for either of these. Figure 5–8 is a scattergram showing the lack of correlation between reaction time and the order of presentation; each participant was presented the input sentences in a different random order.

There were few instances of blocks. A block is defined as any reaction time more than two standard deviations from the mean. These were removed from analysis following the suggestion of Snodgrass, Levy-Berger, and Hayden (1985, p. 52) (but see Heathcote, Popiel, & Mewhort, 1991). There were no repeats between pairs. Presumably the task was short enough to escape significant effects of either fatigue or familiarity. Or perhaps both effects were present and simply

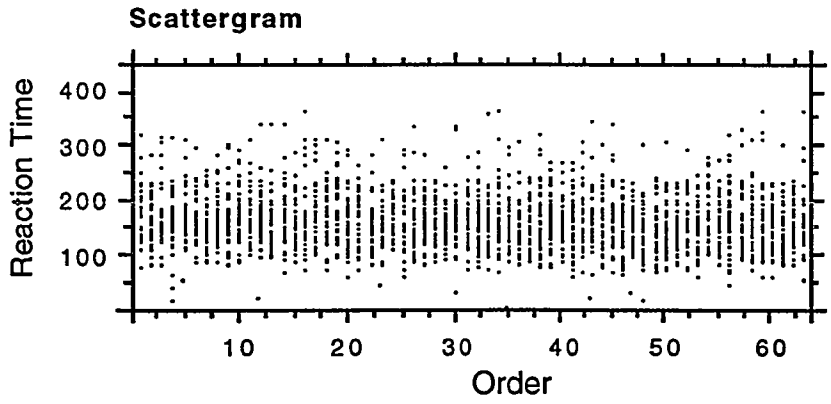


Figure 5–8: RT: Scattergram representing the correlation between reaction time and the order of presentation

cancelled each other — Schmidtke (1976) in researching the effects of varying degrees of choice in an RT task found that the benefits of practice were negated by the fatiguing effects of mental work. Welford finds that it is difficult to separate the effects of stress from those of fatigue. Also, there are different kinds of stress which have different effects; Welford divides stress into two main types — there is stress which tend to increase a participants' level of arousal, for example, loud noises, pressure for speed, physical or mental effort; and stress which lowers a participants' level of arousal, for example, monotony, idleness, and lack of aim or reward. There were both kinds of stress evident in the RT task.

5.5. Summary of primary analyses of the data

Table 5–15 summarizes the findings of this study with respect to grammaticality and length in chunks for each language individually. On all tasks for both languages there was a significant independent effect for the length of the

input sentences in chunks. Table 5–16 summarizes all significant effects cross-linguistically for the three tasks examined in this study. On all tasks the language of the participant significantly affected the outcome. While there was a significant effect due to the grammaticality of the sentences for the P&P and EI tasks, there was no such effect for the RT task. The interaction between grammaticality and chunks in the P&P task was significant for both language groups; the interaction between grammaticality and chunks in the EI task was significant only for the KORs, while in the RT task it was significant only for the ENGs.

Table 5–15: Summary of three tasks: Analysis of variance of dependent variables by (G)rammaticality and length in (C)hunks for each language individually

	P&P		EI		RT	
	Eng	*Kor*	*Eng*	*Kor*	*Eng*	*Kor*
(G)rammaticality	**.0001**	**.0001**	**.0001**	**.0001**	.3587	.4882
(C)hunks	**.0001**	**.0001**	**.0001**	**.0001**	**.0002**	**.0001**
G x C	**.0184**	**.0052**	.7512	**.0001**	**.0002**	.8678

Table 5–16: Summary of three tasks: Analysis of variance of dependent variables by (G)rammaticality, length in (C)hunks, and (L)anguage

	P&P	EI	RT
(G)rammaticality	**.0001**	**.0001**	.9023
(C)hunks	**.0001**	**.0001**	**.0001**
(L)anguage	**.0001**	**.0001**	**.0001**
G x C x L	.2614	**.0001**	**.0116**
G x C	**.0002**	**.0001**	**.0235**
G x L	**.0001**	**.0008**	.2554
C x L	.2641	.4466	.2788

The consistent effect for length in chunks on all tasks is important to the discussion of controlling levels that follows. This effect is not surprising when viewed from the following angle: the of number of chunks is calculated according Abney by referring to the syntactic structure of a string at the level of phrases and, while the number of chunks is not a direct measure of grammaticality, it may nevertheless be a good indicator of the grammaticality of the string. Concretely, the number of chunks in the grammatical ordering of a string of words is generally smaller than the number in an ungrammatical ordering of that same string of words. The second step in Abney's system, that of attachment, depends upon the viability of the S-structure, i.e., the various chunks must be able to attach one to another in such a way as to form an acceptable S-structure in the language in question. In order for a sentence to be grammatical, both steps must be successfully completed. Thus, Abney's system may be used to tease apart effects due to processing that takes place at the phrasal level from that which takes place at the S-structure level.

5.5.1. Controlling levels

The analyses of the data obtained by the three tasks instruments suggest that processing on the three types of grammaticality evaluation tasks accesses multiple levels of representation. In addition, there is evidence that different tasks access different controlling levels.

Grammaticality evaluation tasks access multiple levels of representation. In order to answer the question of whether a task accesses more than one level of representation, one must first consider what would constitute evidence of access to the various levels. The case of the EI task is the most interesting of the three in terms of such consideration. This is due to the nature of the EI task in which participants are not confined to a scale (as in the P&P task) or a binary choice (as in the RT task) by which to express their reaction to the input string. Virtually anything goes in EI, which produces a rich data base to be interpreted by the researcher. I found a number of indicators of access to various levels. For example, to begin at the visual level, as reported previously, various participants reported casually "picturing the sentence in my mind" in order to read it back out in imitation. Among both language groups there were even cases of a participant tapping out the rhythms of the sentences on his or her leg or on the table, presumably in an attempt to capture for replay the prosodic structure, or "melody" of the string. As for letter-level (or phoneme-level) processing, it is possible that the frequent substitution between *described* and *discussed* and also between *Richard* and *Robert* indicates the influence exerted by a particular phomene in a salient position. The abundance of lexical substitution (Section 5.3.5., examples (4) and (5)) provides evidence of lexical-level processing; prepositional phrase substitution (3) as well as phrase-shuffling (2) provides

evidence of phrasal-level processing. Complete corrections (1) indicate a higher, S-structure or perhaps interpretive level of processing; this latter level is supported also by the examples involving cross-category transformations (6).

The arguments in favor of access to multiple levels of representation are more difficult to make for the remaining two tasks, P&P and RT, because the evidence for representations at individual levels is less obvious. On the P&P task, the effect for grammaticality was stronger than that for chunks. There were no differences among input tokens of an example type other than the order of the elements within the nominal phrase, thus, there is no obvious evidence of effects for differences in lexical-level representation[3]. The reason that ungrammatical sentences were rated lower than grammatical ones might be attributed to the participants' perception of them as aberrant at either the phrasal level (in the sense that there were incomplete phrases or orphaned elements) or at the S-structure level (in the sense that attachment failed), or as merely less interpretable. Rating differences could furthermore be due to an interaction of some or all of these levels. Perhaps the relative[4] lack of effect for length in chunks lends weight to the argument that evaluations on the P&P task were more dependent upon the highest, i.e., S-structure and interpretive, levels of representation.

In contrast to the P&P task, on the RT task, it was differences among the length in chunks, not the grammaticality of the strings, that accounted for the significant effects found. This does not, however eliminate the phrasal level as

---

3 Murray (1982) reports shorter reaction times on sentences with plausible meanings, such as (a) than on those with implausible meanings such as (b):
(a) *The dentist extracted the tooth.*
(b) *The clergy advertised the bones.*
This finding suggests an interaction of the lexical level (subcategorization) with the interpretive level (semantic plausibility).

4 Table 5–6 shows that while difference between grammatical and ungrammatical sentences was 3.78 versus 1.80, the difference between six-and seven- chunk sentences was only 2.97 versus 2.60

candidate as an important level of representation. As pointed out previously, the calculation of chunks crucially depends upon the analysis of the input as to syntactic structure at the phrasal, but not at the S-structure level. If support for a particular level is weakened by the RT data, then it is the S-structure level, a position which receives support in the findings of Eubank (in press). Similar to the effect of interpretability on the P&P task, lack of interpretability might also been a factor slowing reaction time.

In summary, it seems likely that all three of the tasks access more than one level of representation during processing; the following section deals with the evidence for a controlling level for each of the tasks.

Different tasks access different controlling levels. On the P&P task, as discussed above, while it is possible to attribute rating differences to processing difficulties at either the phrasal, S-structure, or interpretive levels, there is no clear evidence pointing to any one of these as the controlling level. Robert Bley-Vroman (p.c.) makes an analogy between a participant processing input in order to render a judgment and a computer processing input in order to yield a parse. He suggests that to look for an answer to the question of at what level of representation processing takes place, one might consider the level at which the computer must suspend rules in order to yield a parse of ungrammatical input — probably it should be assumed that, in a bottom-up process, this suspension of rules takes place at the lowest feasible level. By analogy then, upon encountering the lowest level violation in an ungrammatical sentences, the participant is able to judge the input as ungrammatical. In the case of the present study, the phrasal level is the lowest level at which violations occur. By this interpretation, the controlling level in a P&P task may be strongly influenced by the nature of the input. This is supported by results of data collection on P&P tasks in previous

studies discussed in Chapter 4, Section 4.2., for example, input sentences which contain primarily semantic anomalies may encourage a chiefly interpretive strategy on the part of a participant, whereas an input set designed to test purely syntactic features may cause the participant to concentrate, though not necessarily consciously, more on phrase level analyses. The present study is of this latter type.

There is evidence in the EI data of at least limited access to virtually all the representational levels discussed here. As outlined in Chapter 4, Section 4.3.3., Bley-Vroman and Chaudron have suggested the phrasal level as a candidate for the controlling level for EI in the case of input consisting of sentences. There were two kinds of evidence, phrase-substitution and phrase-shuffling, found in the data to support access to this level, though it is not clear what kind of evidence would constitute the claim of the precedence of one level over another as the controlling level. In the EI data there is also evidence of successful attachment, e.g., the elements within phrases may be reordered, chunked, and attached to form grammatical sentences, as well of indications of access to other levels. Although I suspect the successful formation of a representation at the phrasal level is quite influential on the EI process, I can find no compelling reason to claim that the phrasal level is the controlling level in EI.

As mentioned, the effect for chunks on all tasks is not surprising; what is surprising is the existence of an instance in which there is an effect for chunks in the absence of an effect for grammaticality. Specifically, I refer to the results found for both languages on the RT task. I do not claim that the participants' analysis of the strings was conscious, however, I suggest that as part of the RT process, syntactic processing went on to the extent that chunks were calculated

by the participants, that is, the participants were able to analyse the string to the point of identifying the syntactically acceptable constituents, or phrases, of the string. Yet, in contrast to both P&P and EI, there is no evidence that the analysis completed the next step, that of attachment, thus forming a higher level, or S-structure, representation, as grammatical and ungrammatical sentences required virtually the same amount of time to react to. From this, I draw the tentative conclusion that on the particular RT task incorporated in the present study, the phrasal level acts as the controlling level. It remains to be seen if these results are generalizable to other RT tasks.

5.5.2. Interfering or competing factors

As shown in Table 5–16, there was a significant independent effect for the variable of language on all three tasks. In all three cases this effect was manifested as a less "successful" performance by the Korean participants, i.e., they were less discriminating on the P&P task, less successful in imitating the EI sentences, and slower on the RT task than the English-speaking participants.

Returning to the point made in the summary of the EI findings in Section 5.3.5., there are two possible sources of difference between the two groups — the language of the task and the participants themselves. Some of the differences between the English and Korean input sentences are summarized in Chapter 4, Table 4–1, which shows that while sentences in both languages within an example type are equal in length when measured in chunks, the Korean sentences, which actually contain fewer words, average about five more syllables per sentence than their English counterparts. The additional syllables in Korean may overload STM, making less likely the retention of a representation from which to produce an imitation on the EI task, or by which to determine if a pair

sentence matches on the RT task. While this explanation seems reasonable, it does not address the question of why the Korean participants were less discriminating in their evaluations on the P&P task where they are given unlimited time and access to the sentences. The fact that the Koreans were consistently less successful across all three tasks indicates that there may be complicating factors having to do with the participants themselves and not the language of the task.

Both the EI and the RT tasks, and to perhaps a lesser extent, the P&P task, depend upon the participants' ability to take in a signal, form a representation on one or more levels, and take some action based upon the representation. A critical factor in this process for both tasks is the capacity of STM. It seems reasonable to speculate that the effects of age, fatigue, and stress on reaction time discussed above are generalizable to the EI task as well.

A fourth factor which may affect all three tasks is the participants' level of education. As mentioned in the summary of the EI analysis in Section 5.3.5., although it is difficult to make comparisons between products of the Korean and US educational systems, it seems a reasonable assumption that the KORs, as high school students on a vocational track, had achieved a generally less broad education than had the ENGs, all University of Hawai'i students. As discussed in Chapter 3, Section 3.1.2, degree of literacy can account for variation in metalinguistic awareness. Of course, all the participants in the present study were literate in their native language, but there are many types of "literacy". Two types that may have had effects here are literacies borne of familiarity with the fields of computer technology and of psychology. As for the first of these, the KORs appeared to be less familiar with computers and the use of a computer keyboard which could account for some slowing of their reactions. In addition,

the KORs may have been less familiar with design and use of psychological testing in particular — this is in contrast to this particular group of ENGs who were had relatively sophisticated knowledge of this type of activity. They had discussed it in their psychology classes, and in fact, for many of the participants, this was the fourth experiment in which they had voluntarily participated during the current semester. It is also less likely that their anxiety level was as high as that of the Koreans, as they were being rewarded for their participation rather than being required to participate. The KORs may very well have been suspicious of how the results of the experiments were to be used and have felt anxious about the ability of the instrument to fairly represent an individual's abilities. Such effects have been documented in a study (Oh, 1992) of Korean students being tested on their knowledge of English. Oh found that performance suffered significantly as the students' unfamiliarity with a task and their anxiety about the results increased. It is also possible that the Koreans in general had a lower self-perception about quality of their overall performance — as mentioned earlier, these students were "vocational track" students, without plans for additional education past high school graduation. More than one person on the staff at the high school questioned my choice of this group, claiming that it would not be representative, that these students could not be expected to perform as well as their college-bound peers.

As for the interfering factors discussed here, I make no specific claims about their effects on the data collected. I simply enumerate them and include my observations in hopes that these will not only suggest pitfalls to avoid in designing and undertaking an experiment aimed at accessing underlying competence, but also to indicate the complexities of analysing the data obtained.

CHAPTER 6: SECONDARY ANALYSES OF THE DATA

The excluded example types 14K and 20K and the MOD–NP category are examined here in the secondary analyses. In the first of these analyses (Section 6.2.), conclusions from the previous chapter about way in which the written questionnaire ("P&P") and the elicited imitation task ("EI") work can be used to test the controversial example types 14K and 20K for clues about their grammatical status. In the remaining analysis (Section 6.3), I return to the MOD–NP category to see if, based on what can be concluded about the suitability of the reaction time task ("RT") for investigating length in chunks, any inferences can be made about the length in chunks of the uncertain MOD–NP example types.

6.1. Research questions leading to the secondary analyses

Based on the primary analyses of the data, it is now possible to formulate the following specific questions. The answers to these questions should shed light both on the the grammatical status of the controversial example types 14K and 20K and on the uncertain length in chunks of the example types form the MOD–NP category.

Grammaticality of 14K and 20K

> Did the data on example types 14K and 20K pattern with data on grammatical items or with data on ungrammatical items?

Chunk length of 3E* and 4E*

> Did the data on example types 3E* and 4E* from the MOD–NP category (example types of uncertain length in chunks) pattern with data on example types 1E or data on 2E* (for which length in chunks is known)?

6.2. Grammaticality of 14K and 20K

As mentioned in Chapter 2 (Section 2.2.4.; also footnote 10), there is a fair amount of disagreement about Yoon's designation of "ungrammatical" for the example types, 14K and 20K; tokens of these example types are shown here. It is because of this disagreement that these examples were withheld from the primary analyses.

2ARG–NP

14E* They mentioned Japan's Seoul's the destruction in the letter.

14K? kutul-i phyenci-ey ilpon-uy ku phakoy-lul sewul-uy iyakihayssta.
they-NOM letter-LOC Japan-GEN destruction-ACC Seoul-GEN talk about-PAST-SE

3ARG–NP

20E* We described Steven's to Jenny of flowers the gift to the group.

20K? wuli-ka salamtul-eykey chelswu-uy pohuy-eykey-uy kkoch-uy ku ku senmul-ul kiswulhayssta.
we-NOM group-DAT [name]-GEN [name]-GEN flower-GEN the gift-ACC describe-PST-SE

As shown by the primary analyses of the three tasks in Chapter 5, it can be claimed with a degree of confidence that for sentences of six to seven chunks in length, both the P&P and the EI tasks are reasonably sensitive to the grammaticality of a sentence. In this secondary analysis, sentences 14K and 20K, for which the length in chunks but not the grammatical status is known, are compared with the remaining example types in the same modification categories. This is done to determine how these controversial cases pattern with respect to example types for which both grammaticality and length in chunks are known.

6.2.1. Grammaticality of 14K

In order to infer the grammatical status of 14K, which is six chunks in length, the KORs' mean Likert ratings (P&P) and imitation success scores (EI) on this sentence are compared to those on the other six-chunk sentences, both grammatical and ungrammatical, in the same modification category, 2ARG–NP. Although example type 15K* is a member of this category, it is omitted from analysis here because it is seven chunks in length. For ease of exposition, these example types with English literal translations of each of the example types in this category are provided; chunks are indicated:

2ARG–NP

10K* kutul-i phyenci-ey ilpon-uy phakoy-lul sewul-uy iyakihayssta
they / in the letter / Japan's / destruction / Seoul's / talked about

11K* kutul-i phyenci-ey sewul-uy phakoy-lul ilpon-uy iyakihayssta
they / in the letter / Seoul's / destruction / Japan's / talked about

13K kutul-i phyenci-ey ilpon-uy sewul-uy phakoy-lul iyakihayssta
they / in the letter / Japan's / Seoul's / destruction / talked about

14K? kutul-i phyenci-ey ilpon-uy sewul-uy ku phakoy-lul iyakihayssta
they / in the letter / Japan's / Seoul's / the destruction / talked about

To find the pattern among these sentences, least squares means tables (Tables 6–1 and 6–2) were calculated for these four example types on both P&P and EI.

Table 6–1: P&P least squares means table for example types 10K*–14K?

ex. type	vs. ex. type		means	t-Test	P-value
10K*			2.49		
	≠	11K*	2.14	2.72	**.0074**†
	≠	13K	3.24	–5.91	**.0001**†
	≠	14K?	3.00	–3.65	**.0004**†
11K*			2.14		
	≠	13K	3.24	–8.62	**.0001**†
	≠	14K?	3.00	–6.36	**.0001**†
13K			3.24		
	≈	14K?	3.00	2.26	.0251

†Probability is significant ($p<.01$)

The P&P table shows that, as expected, both ungrammatical sentences 10K* and 11K* were significantly different from the grammatical 13K. An unexpected result was the significant (though less so) difference between ungrammatical 10K* and ungrammatical 11K*. The mean rating for 14K was not as high as that of 13K. As the bar graph below shows, this lower rating may be attributed in part to the large number of tokens of 14K that received a rating of "1" ("bad").

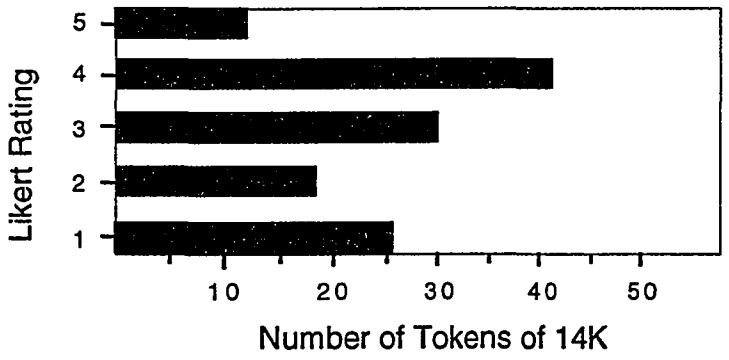


Figure 6–1: Likert ratings of all tokens of 14K?

What might account for this group of outliers? Both Ho-min Sohn and William O'Grady (personal communication) suggest that the differences may be due to dialectal differences among the participants. Overall, however, the sentence in question, 14K, patterned with 13K and against 10K* and 11K*, tentatively supporting the conclusion that for these KORs as a group, 14K was grammatical.

The EI table below shows that among the example types for which grammaticality is known, the significant effects were as expected. The ungrammatical 10K* patterned with the ungrammatical 11K*; these were both significantly different the grammatical 13K. The results for the sentence in question, 14K, though close to reaching significance, do not fully support the results found with the P&P instrument, therefore, the conclusion that 14K is grammatical for this group of KORs is weaker than if the EI results had been clearer. Nevertheless, the EI results are not inconsistent with the P&P results.

Table 6–2: EI least squares means table for example types 10K*–14K?

ex. type	vs. ex. type		means	t-Test	P-value
10K*			.186		
	≈	11K*	.159	.53	.5991
	≠	13K	.442	–4.69	**.0001**†
	≈	14K?	.296	–2.19	.0301
11K*			.159		
	≠	13K	.442	–5.22	**.0001**†
	≠	14K?	.296	–6.36	**.0073**†
13K			.442		
	≈	14K?	.296	2.51	.0131

†Probability is significant (p<.01)

6.2.2. Grammaticality of 20K

In an similar test of the grammatical status of 20K, a sentence of seven-chunks in length, the KORs' mean Likert ratings on this sentence was compared to the ratings on the other seven-chunk sentences in the same modification category, 3ARG–NP. For ease of exposition, these example types with English literal translations of each of the example types in this category are provided; chunks are indicated:

3ARG–NP

16K* wuli-ka salamtul-eykey chelswu-uy senmul-ul pohuy-eykey-uy kkoch-uy kiswulhayssta
we / to the group / Steven's / gift / flowers' / Jenny's / described

18K* wuli-ka salamtul-eykey senmul-ul chelswu-uy pohuy-eykey-uy kkoch-uy kiswulhayssta
we / to the group gift / flowers' / Steven's / Jenny's / described

19K wuli-ka salamtul-eykey chelswu-uy pohuy-eykey-uy kkoch-uy senmul-ul kiswulhayssta
we / to the group / Steven's / Jenny's / flowers' / gift / described

20K? wuli-ka salamtul-eykey chelswu-uy pohuy-eykey-uy kkoch-uy ku senmul-ul kiswulhayssta
we / to the group / Steven's / Jenny's / flowers' / the gift / described

To find the pattern among these sentences, least squares means tables were calculated for these four example types on both P&P (Table 6–3) and EI. The P&P table shows results similar to those found earlier. Among the example types for which grammaticality is known, the ungrammatical sentence 16K* was not significantly different from the ungrammatical 18K* but was significantly different the grammatical 19K. Ungrammatical 18K* was also significantly different the grammatical 19K. The sentence in question, 20K, clearly patterns with 19K, against 16K* and 18K*, thus it is tentatively concluded that for this group of KORs, 20K was grammatical.

Table 6–3: P&P least squares means table for example types 16K*–20K?

ex. type	vs.	ex. type	means	t-Test	P-value
16K*			1.98		
	≈	18K*	1.72	1.86	.0653
	≠	19K	2.56	–4.16	**.0001**†
	≠	20K?	2.39	–2.95	**.0036**†
18K*			1.72		
	≠	19K	2.56	–6.01	**.0001**†
	≠	20K?	2.39	–4.81	**.0001**†
19K			2.56		
	≈	20K?	2.39	1.21	.2287

†Probability is significant (p<.01)

The table for the EI least squares means for example types 16K* through 20K is not shown. There was no significant differences between any two example types ($p<.01$). From this, I conclude, as I did in the primary analysis of the data from the EI task, that the task was simply too difficult at the seven-chunk level for the KORs to obtain any significant results.

6.2.3. Summary of grammaticality of 14K and 20K

In the cases where there are significant results, example types 14K and 20K pattern with the grammatical example types in their categories. This contradicts the analysis of Yoon but supports both Fukui's analysis (Chapter 2, Section 2.2.4) and the intuitions of a number of Korean informants who claim that the example types represented by 14K and 20K are acceptable sentences in Korean.

6.3. Chunk length of 3E* and 4E*

In Chapter 2, Section 2.3.3., there was a discussion of the application Abney's chunking scheme to the input sentences. The system was deficient for my purposes in that there was insufficient information to determine how certain ungrammatical English input sentences should be chunked. These sentences occurred in the MOD–NP category. The uncertainty arose due to the "wandering" determiner (the functional head) which *f*-selects the content word to the right of it only and, in effect, orphans the preceding elements. What is not clear is if these orphaned elements can band together to form some kind of chunk on their own or if each one constitutes a separate chunk. For example, does the

nominal phrase in 4E*, *big blue book the*, consist of two chunks, *big blue book* and *the*, or of four, *big* , *blue* , *book* and *the*[1]? The English example types 1E through 4E* are repeated here in Table 6–4; the wandering determiner *the* is shown in italics, known chunks ("/") and possible boundaries ("?") are indicated.

Table 6–4: English example types 1E through 4E* with length in chunks

	example type	*chunks*
1E	Mary / found / *the* big blue book / under the desk.	4
2E*	Mary / found / big / *the* blue book / under the desk.	5
3E*	Mary / found / big ? blue ? *the* book / under the desk.	?
4E*	Mary / found / big ? blue ? book / *the* / under the desk.	?

According to the primary analyses of the data from the RT instrument in the previous chapter, this instrument can be considered to be relatively sensitive to length in chunks, at least within the narrow band that includes sentences of six and seven chunks. In this analysis, I return to the MOD–NP example types to see if, based on what can be concluded about the suitability of RT for investigating length in chunks, any inferences can be made about the length in chunks of the uncertain MOD–NP example types. If the RT instrument is sensitive to sentences of fewer chunks in length, then, based on the clear cases, namely 1E (4 chunks) and 2E* (5 chunks), I can look for clues as to the length in chunks of the remaining example types. If mean reaction times for 3E* or 4E* differ significantly from 1E, but not from 2E*, one might in principle take this as evidence that the two example types in question also consist of the five chunks

1 Freedman and Forster (1985:104), in their discussion of the reaction time process, make the claim that "a phrase such as *the young girl* can be matched as a single unit, whereas *girl the young* must be matched as three separate units". However, they offer no support for this claim.

and thus the orphaned elements are able to form some sort of headless chunk. If, on the other hand, the reaction time on these were found to still be significantly longer than that on the five chunk example type, 2E*, one might conclude that each of these elements constitutes a separate chunk unto itself.

First, it must be determined if the instrument is indeed sensitive to length in the four- to five-chunk range. If the RT instrument remains sensitive to chunks at this lower level, then the mean reaction time on 1E (four chunks) should be significantly faster than that on 2E* (five chunks). A least squares means table was calculated for the four example types to determine if reaction time on any of the example types differs significantly from that on any other; see Table 6–5. In fact, no significant differences resulted between any two example types. From this, it seems that this RT instrument is not sensitive to length in chunks for sentences of less than six chunks, i.e. six chunks represents the ceiling for input on this task.

Table 6–5: RT least squares means table for example types 1E–4E*

ex. type	vs. ex. type	#chks.	means	t-Test	P-value
1E		(4)	156.75		
	≈ 2E*	(5)	157.98	–0.05	.9576
	≈ 3E*	(?)	166.71	–0.43	.6677
	≈ 4E*	(?)	202.04	–1.96	.0527
2E*		(?)	157.98		
	≈ 3E*	(?)	166.71	–0.37	.7067
	≈ 4E*	(?)	202.04	–1.90	.0594
3E*		(?)	166.71		
	≈ 4E*	(?)	202.04	–1.52	.1298

†Probability is significant ($p<.01$)

Even if the results had been significant at the lower chunk level, the interpretation of the results would have been problematic due to the fact that there is no sure way to know whether a difference in reaction times should be attributed to an effect for grammaticality (1E is grammatical, the others not) or length in chunks (1E is four chunks, the rest are longer) or an interaction between the two.

As support for the claim that the RT instrument is insensitive to length a least squares means table was also calculated for the Korean data in the MOD–NP category to see if the non-significance for length generalizes across language groups; see Table 6–6. For the Korean counterpart sentences, unlike the English sentences, the chunks length of all example types is known as shown in Table 6–6.

Table 6–6: Korean example types 1K through 4K* with length in chunks

	example type	*chunks*
1K	yenghuy-ka / chayksang-mitheyse / ku khun phulun chayk-ul / palkyenhayssta. *[name]-NOM desk-LOC the big blue book-ACC find-PST-SE*	4
2K	yenghuy-ka /chayksang-mitheyse /khun ku phulun chayk-ul / palkyenhayssta.	4
3K	yenghuy-ka /chayksang-mitheyse / khun phulun ku chayk-ul / palkyenhayssta.	4
4K*	yenghuy-ka /chayksang-mitheyse / khun phulun chayk-ul / ku / palkyenhayssta.	5

Once again, as shown in Table 6–7 all results on the Korean example types were non-significant, thus neither length in chunks nor grammaticality had an effect at the lower chunk levels.

Table 6–7: RT least squares means table for example types 1K–4K*

ex. type	vs. ex. type	#chks.	means	t-Test	P-value
1K		(4)	293.85		
	≈ 2K	(5)	193.41	1.09	.2795
	≈ 3K	(5)	237.18	0.06	.9504
	≈ 4K	(5)	208.21	0.74	.4605
2K		(5)	193.41		
	≈ 3K	(5)	237.18	–1.02	.3078
	≈ 4K	(5)	208.21	–0.35	.7299
3K		(5)	237.18		
	≈ 4K	(5)	208.21	0.68	.4990

†Probability is significant ($p<.01$)

Thus, although in principle the RT instrument can be used to provide an indication of relative length in chunks, in fact, the range to which the instrument is sensitive is too narrow to allow the sensitivity evidenced on the sentences of more than five chunks to be generalized to shorter input sentences. The non-significant results are not surprising in light of Bley-Vroman and Chaudron's *narrow band*. The ceiling effects found in this analysis together with the floor effects found in the primary analyses of the data set the narrow band at the six- to seven-chunk range for the instrument used in this study. This is true for both language groups.

6.4. Summary of secondary analyses of the data

In the input sets there were a pair of sentences in Korean, 14K and 20K, whose grammaticality was in question and a pair of English sentences, 3E* and 4E*, whose length in chunks was in question. Using both the P&P and EI instruments as indicators of grammaticality, I showed that the controversial Korean sentences patterned with grammatical counterparts. Using the RT instrument as an sensor of length in chunks, I tested the controversial English sentences, but uncovered no strong significant independent effects that would indicate how these items pattern with respect to counterpart sentences of known length in chunks. In fact, no significant independent effects were found for the Korean counterpart sentences, for which length in chunks is known, leading to the conclusion that sentences of fewer than six chunks in length fall below the limit of the range to which the RT instrument is sensitive.

CHAPTER 7: SUMMARY AND CONCLUSION

7.1. Summary

In this final chapter, I summarize the main findings of the comparison of the three grammaticality evaluation instruments. Following the summary, I identify and discuss what I feel are the three main themes arising from the investigation of their use. The first is the importance of understanding what a specific grammaticality evaluation instrument measures. Second, I point out the value of converging evidence, and last, I emphasize the need for independent data in validating linguistic theory.

In Chapter 2, I described three categories of nominal phrases, and presented example sentences in both English and Korean. These were analysed in terms of their grammaticality, referring to the DP analysis of Abney (1987) and the complementary DelP analysis of Yoon (1990), both formulated within the Government and Binding framework. Sentences used to exemplify the nominal phrases were also analysed in terms of their length in chunks, adopting a system developed by Abney (1991). Although the overlap between (a) grammatical versus ungrammatical orderings of elements in a sentence and (b) orderings that are analysed as being comprised of fewer versus more chunks is considerable, it is not complete. This difference was exploited in this study as a tool to to investigate

factors at work in the processing of the sentences. Example sentence types from the three nominal phrase categories described in this chapter served as the input to the present study.

The topic of Chapter 3 was the nature of metalinguistic awareness, i. e., the ability to observe language from the outside, which is present to varying degrees in all language users. The main purpose of this study was not to test the validity of the syntactic framework outlined in Chapter 2, but to use the framework in the construction of a research program designed to compare methods of obtaining grammaticality evaluations from English and Korean native-speaking participants. I proposed that the ability of a participant to render a grammaticality evaluation, either consciously or not, depends upon her level of metalinguistic awareness, her ability to access this awareness, and her ability to express the evaluation. It was stressed that, although existing instruments for obtaining grammaticality evaluations do not provide incontrovertible data that faithfully reflects a language user's underlying competence, grammaticality evaluation instruments can nevertheless be considered useful tools for not only gaining insights into the systems underlying language competence, but also into non-linguistic processes at work.

In Chapter 4, I gave a detailed description of the overall research design and of each of the three different instruments being investigated. The instruments were a written questionnaire ("pencil-and-paper" or "P&P"), a reaction time task ("RT"), and an elicited imitation task ("EI"). There was a discussion of previous uses of these three including speculation about what these instruments measure and the processes involved, covering levels of representation, controlling levels, and the role of interfering and competing factors. The design of each instrument was described in turn, and design considerations unique to each particular instrument were discussed.

Chapter 5 contained the primary analyses of the data obtained. The main focus was to investigate both the functional characteristics of making evaluations and the physiological processes involved, with an emphasis on the former. The primary analyses of the data obtained by the three tasks instruments suggested that in processing the input sentences on the three types of grammaticality evaluation tasks, the participants accessed multiple levels of linguistic representation and that different tasks accessed different controlling levels. Specifically, with respect to the issue of sensitivity to grammaticality versus length in chunks, it was found that the P&P instrument was primarily, though not exclusively, sensitive to the grammaticality of the input sentences. The EI instrument was sensitive to grammaticality and length in chunks in more or less equal measure, while the RT instrument was sensitive to length in chunks alone. There was also evidence of a number of non-linguistic interfering or competing factors, such as age, fatigue, stress, and the level of education of the participants.

In Chapter 6, I returned to two issues raised in Chapter 2. The first of these was the disagreement between the analysis of Yoon, on the one hand, and the analysis of Fukui, supported by the native speaker intuitions, on the other, with respect to the grammatical status of Korean nominal phrases in which the determiner intervenes between modifiers and the head noun, represented by example types 14K and 20K. The data collected supported Fukui's analysis which predicts such structures are grammatical. The second issue arose from the inadequacy of Abney's chunking system for calculating length in chunks for some of the ungrammatical English input sentences. By using the RT instrument, which had been shown to be sensitive to length in chunks on a different set of the input sentences, I looked for evidence of the relative length in chunks of the sentences in question. The results of the data analysis were inconclusive on this point,

however, the results did help to zero in on the "narrow band" to which the RT instrument is sensitive.

7.1.1. The importance of understanding what a specific grammaticality evaluation instrument measures

In referring to native language user data in support of a linguistic theory, it is important to determine, to the extent possible, what precisely the data collection instrument being used measures, and to identify and control for factors likely to interfere with a clean collection of the elements of interest — this is what I referred to earlier as "fine-tuning". Although it is unlikely that one can identify all the factors that contribute to, or interfere with the production of a measurement of a phenomenon, it is possible, through examination of previous uses of an instrument, and through experimentation and piloting of that instrument, to arrive at a better understanding of the nature of both the instrument itself and of the behavior being measured in order to make the appropriate adjustments.

Perhaps one of the clearer and, at the same time, more unanticipated findings of the present study was the extent to which the length of a sentence as calculated in chunks influences what had been assumed to be a measure of its grammaticality. For the sake of argument, suppose a researcher relied solely on the rating obtained on the P&P instrument employed in this study as an indication of grammaticality in English[1]. In that case, she would surmise that

---

1 Although all differences between pairs are significant, the raw Likert rating scores indicate that 16E is more like 1E than it is like 3E*.

ex. type	vs. ex. type	means	P-value
16E		4.41	
	1E	5.00	.0001
	3E*	1.62	.0001
1E	3E*		.0001

the sentences 1E and 16E were essentially equivalent in terms of grammaticality for native speakers, in contrast to the ungrammatical sentence 3E*.

1E Mary found the small red notebook under the desk.
16E They discussed Steven's gift of flowers to Jenny at home.
3E* Mary found small red the notebook under the desk.

If, on the other hand, the researcher relied only on the results obtained by the EI task[2], she might conclude that sentence 16E above was considerably worse than the 1E as evidenced by the lower imitation success score. In fact, the imitation success score on 16E appears roughly equivalent that on 3E*, a sentence that intuitively seems much worse.

This is not to say that either the P&P or the EI instrument is inherently inaccurate, only that an understanding of each measure is crucial to the analysis of the results that it yields. The P&P instrument was found to primarily evaluate grammaticality as opposed to length in chunks for sentences of six to seven chunks in length; the anecdotal evidence presented here suggests that this holds true even when comparing these to sentences of as few as four chunks in length, like 1E. In contrast, the EI task was shown to be sensitive to both grammaticality and length in chunks within the range of six to seven chunks. As the above example shows, when the input includes four-chunk sentences such as 1E as well, the effect for length in chunks is more even more salient. The moral of this story,

2 There is a significant difference between 16E and 1E, even though both are grammatical; there is no significant difference between the grammatical 16E and the ungrammatical 3E*.

ex. type	vs. ex. type	means	P-value
16E		.556	
	1E	.947	.0001
	3E*	.539	.9970
1E	3E*		.0001

it seems, is that in order to use the EI instrument as a measure of grammaticality with confidence, one must very strictly control for length in chunks.

In addition to the effect of length in chunks evident in this study, the investigation of interfering and competing factors uncovered effects that were tentatively attributed to the differences between the two language groups relating to anxiety and level of education.

7.1.2. The value of converging evidence

The multi-method approach to measuring the object of interest is by no means a new concept. Denzin (1970), for example, has proposed the concept of "triangulation" which is commonly used in qualitative research; the term refers to desirability of the use of participant observation, interviews, and document analysis as well as having more than one participant and one observer. Unfortunately, the collection and analysis of converging evidence is still fairly rare in our field, especially in the area of L1 studies. Nevertheless, such an approach is necessary to promote the validity and generalizability of experimental findings.

Converging evidence from various data collection can be used in making claims about data in support of theory. If a researcher can capture a reasonably reliable indication of a language user's competence through the use of metalinguistic evaluation, she then has a standard by which to measure the validity of a theory. Although there is always a danger that a set of results which conforms to a given theory does not necessarily confirm it, this uncertainty can be mitigated to some extent by supplying converging evidence from cross-validating measures. Furthermore, effects due to defects in task design can be diminished to some extent via cross-validating results from a different (and possibly equally

defective) instrument. For example, Bley-Vroman and Masterson (1989, p. 224), in a discussion of an early version of the present study, comment that "although both [P&P & RT] methods are flawed, they are clearly flawed in different ways, so that the convergent results inspire a degree of confidence which neither set of results commands by itself."

7.1.3. The need for independent data in validating linguistic theory

As Carlson and Tanenhaus (1989, p. 24) point out, "Experimentation provides the facts upon which the field is built; the validity of theories based on them depends on their being facts to begin with." Any scientific endeavor must base theory validation upon objective observations of the phenomenon under investigation; in the case of the field of linguistics, the phenomenon is language. Independent data is critical for validating predictions of grammaticality made by a theory. Instruments intended to objectively observe and measure language behavior must thus be atheoretical. A linguist's subjective evaluation of her own personal intuition about a structure, while perhaps useful in constructing a theory, cannot in turn be used as evidence of its validity.

The present study demonstrates a case in point. Yoon (1990) has constructed a convincing theory of nominal phrases in Korean. Unfortunately, a rather major point in his theory (in fact all of his theory) is supported by what are apparently his personal intuitions about what is, and is not, a grammatical structure in Korean. Many of the judgments he makes are so obvious as to seem self-evident. However, the data collected from Korean native speakers on example types 14K and 20K in particular supplied compelling evidence for the grammaticality of these structures, undermining an important portion of Yoon's theory.

7.2. Conclusion

Much remains to be done with respect to understanding both the psychological and the physiological processes which are involved when we attempt to tap the metalinguistic awareness of language users. In this study, I identified a number of factors that must be taken into account when designing instruments to collect grammaticality evaluations. There are task-internal factors to consider such as the preparation of the input, pre-task instruction, recency/primacy effects, the semantic and contextual considerations, and the analysis of the resulting data. There are also task-external factors such as the syntactic framework employed, the language of the task, and the effects of age, fatigue, stress, and level of education of the participants.

One of the more interesting findings of this study was the effect of length of the input sentences in chunks upon the participants' reaction to the input, whether that reaction be measured as a judgment of grammaticality, as on the P&P task; as imitation success, as on the EI task; or as reaction time, as on the RT task. Only two of the three, the P&P and the EI instruments yielded data that showed significant effects for the grammatical status of the input. The question remains as to what other previously-overlooked aspects of the input to grammaticality evaluation instruments may have equally consequential effects on the data collection via these and other instruments.

ENGLISH

MOD–NP category: noun modified by a determiner and adjectives

vocabulary set:

DET	ADJ	NOUN
the	big, blue	textbook
the	small, red	notebook
the	new, black	briefcase

frames:

David lost ____ in the classroom
Linda saw ____ on the chair
Mary found ____ under the desk

example type: 1E (DET ADJ ADJ NOUN)

David lost the big blue textbook in the classroom
David lost the small red notebook in the classroom
David lost the new black briefcase in the classroom
Linda saw the big blue textbook on the chair
Linda saw the small red notebook on the chair
Linda saw the new black briefcase on the chair
Mary found the big blue textbook under the desk
Mary found the small red notebook under the desk
Mary found the new black briefcase under the desk

example type: 2E* (ADJ DET ADJ NOUN)

David lost big the blue textbook in the classroom
David lost small the red notebook in the classroom
David lost new the black briefcase in the classroom
Linda saw big the blue textbook on the chair
Linda saw small the red notebook on the chair
Linda saw new the black briefcase on the chair
Mary found big the blue textbook under the desk
Mary found small the red notebook under the desk
Mary found new the black briefcase under the desk

example type: 3E* (ADJ ADJ DET NOUN)

David lost big blue the textbook in the classroom
David lost small red the notebook in the classroom
David lost new black the briefcase in the classroom
Linda saw big blue the textbook on the chair
Linda saw small red the notebook on the chair
Linda saw new black the briefcase on the chair

Mary found big blue the textbook under the desk
Mary found small red the notebook under the desk
Mary found new black the briefcase under the desk

example type: 4E* (ADJ ADJ NOUN DET)

David lost big blue textbook the in the classroom
David lost small red notebook the in the classroom
David lost new black briefcase the in the classroom
Linda saw big blue textbook the on the chair
Linda saw small red notebook the on the chair
Linda saw new black briefcase the on the chair
Mary found big blue textbook the under the desk
Mary found small red notebook the under the desk
Mary found new black briefcase the under the desk

POS-NP category: a noun modified by a genitive-marked possessive nominal, which in turn is modified by a genitive-marked nominal

vocabulary set:

DET	GEN	NOUN
the	David, teacher	house
the	Jenny, neighbor	shop
the	Father, friend	office

frames:

We went to ____ after the party
I stopped by ____ after class
They arrived at ____ before noon

example type: 5E (GEN GEN NOUN)

We went to David's teacher's house after the party
We went to Jenny's neighbor's shop after the party
We went to Father's friend's office after the party
I stopped by David's teacher's house after class
I stopped by Jenny's neighbor's shop after class
I stopped by Father's friend's office after class
They arrived at David's teacher's house before noon
They arrived at Jenny's neighbor's shop before noon
They arrived at Father's friend's office before noon

example type: 6E (DET NOUN GEN GEN)

We went to the house of David's teacher after the party
We went to the shop of Jenny's neighbor after the party
We went to the office of Father's friend after the party
I stopped by the house of David's teacher after class
I stopped by the shop of Jenny's neighbor after class
I stopped by the office of Father's friend after class

They arrived at the house of David's teacher before noon
They arrived at the shop of Jenny's neighbor before noon
They arrived at the office of Father's friend before noon

example type: 7E (DET NOUN GEN GEN)

We went to the house of a teacher of David's after the party
We went to the shop of a neighbor of Jenny's after the party
We went to the office of a friend of Father's after the party
I stopped by the house of a teacher of David's after class
I stopped by the shop of a neighbor of Jenny's after class
I stopped by the office of a friend of Father's after class
They arrived at the house of a teacher of David's before noon
They arrived at the shop of a neighbor of Jenny's before noon
They arrived at the office of a friend of Father's before noon

example type: 8E* (GEN GEN DET NOUN)

We went to David's teacher's the house after the party
We went to Jenny's neighbor's the shop after the party
We went to Father's friend's the office after the party
I stopped by David's teacher's the house after class
I stopped by Jenny's neighbor's the shop after class
I stopped by Father's friend's the office after class
They arrived at David's teacher's the house before noon
They arrived at Jenny's neighbor's the shop before noon
They arrived at Father's friend's the office before noon

example type: 9E* (GEN GEN NOUN DET)

We went to David's teacher's house the after the party
We went to Jenny's neighbor's shop the after the party
We went to Father's friend's office the after the party
I stopped by David's teacher's house the after class
I stopped by Jenny's neighbor's shop the after class
I stopped by Father's friend's office the after class
They arrived at David's teacher's house the before noon
They arrived at Jenny's neighbor's shop the before noon
They arrived at Father's friend's office the before noon

2ARG–NP category: nominal phrase containing an external argument plus one subcategorized internal argument

vocabulary set:

DET	GEN	NOUN
the	Japan, Seoul	destruction
the	Jenny, Richard	criticism
the	Steven, Linda	photo

frames:

We discussed ____ at home

I described ____ to the group

They mentioned ____ in the letter

example type: 10E (GEN (ext. arg.) NOUN GEN (int. arg.))

We discussed Japan's destruction of Seoul at home
We discussed Jenny's criticism of Richard at home
We discussed Steven's photo of Linda at home
I described Japan's destruction of Seoul to the group
I described Jenny's criticism of Richard to the group
I described Steven's photo of Linda to the group
They mentioned Japan's destruction of Seoul in the letter
They mentioned Jenny's criticism of Richard in the letter
They mentioned Steven's photo of Linda in the letter

example type: 11E (GEN (int. arg.) NOUN GEN (ext. arg.))

We discussed Seoul's destruction by Japan at home
We discussed Richard's criticism by Jenny at home
We discussed Linda's photo by Steven at home
I described Seoul's destruction by Japan to the group
I described Richard's criticism by Jenny to the group
I described Linda's photo by Steven to the group
They mentioned Seoul's destruction by Japan in the letter
They mentioned Richard's criticism by Jenny in the letter
They mentioned Linda's photo by Steven in the letter

example type: 12E (DET NOUN GEN (int. arg.) GEN (ext. arg.))

We discussed the destruction of Seoul by Japan at home
We discussed the criticism of Richard by Jenny at home
We discussed the photo of Linda by Steven at home
I described the destruction of Seoul by Japan to the group
I described the criticism of Richard by Jenny to the group
I described the photo of Linda by Steven to the group
They mentioned the destruction of Seoul by Japan in the letter
They mentioned the criticism of Richard by Jenny in the letter
They mentioned the photo of Linda by Steven in the letter

example type: 13E* (GEN (ext. arg.) GEN (int. arg.) NOUN)

We discussed Japan's Seoul's destruction at home
We discussed Jenny's Richard's criticism at home
We discussed Steven's Linda's photo at home
I described Japan's Seoul's destruction to the group
I described Jenny's Richard's criticism to the group
I described Steven's Linda's photo to the group
They mentioned Japan's Seoul's destruction in the letter

They mentioned Jenny's Richard's criticism in the letter
They mentioned Steven's Linda's photo in the letter

example type: 14E* (GEN (ext. arg.) GEN (int. arg.) DET NOUN)

We discussed Japan's Seoul's the destruction at home
We discussed Jenny's Richard's the criticism at home
We discussed Steven's Linda's the photo at home
I described Japan's Seoul's the destruction to the group
I described Jenny's Richard's the criticism to the group
I described Steven's Linda's the photo to the group
They mentioned Japan's Seoul's the destruction in the letter
They mentioned Jenny's Richard's the criticism in the letter
They mentioned Steven's Linda's the photo in the letter

example type: 15E* (GEN (ext. arg.) GEN (int. arg.) NOUN DET)

We discussed Japan's Seoul's destruction the at home
We discussed Jenny's Richard's criticism the at home
We discussed Steven's Linda's photo the at home
I described Japan's Seoul's destruction the to the group
I described Jenny's Richard's criticism the to the group
I described Steven's Linda's photo the to the group
They mentioned Japan's Seoul's destruction the in the letter
They mentioned Jenny's Richard's criticism the in the letter
They mentioned Steven's Linda's photo the in the letter

3ARG–NP category: nominal phrase containing an external argument plus two subcategorized internal arguments

vocabulary set:

DET	GEN	NOUN
the	Steven, Jenny, flowers	gift
the	Linda, Richard, job	offer
the	Mary, Robert, money	donation

frames:

They discussed ____ in class
We described ____ to the group
I recommended ____ in the letter

example type: 16E (GEN (ext. arg.) NOUN GEN (int. arg.) GEN (int. arg.))

They discussed Steven 's gift of flowers to Jenny at home
They discussed Linda's offer of a job to Richard at home
They discussed Mary's donation of money to Robert at home
We described Steven 's gift of flowers to Jenny to the group
We described Linda's offer of a job to Richard to the group
We described Mary's donation of money to Robert to the group

I recommended Steven 's gift of flowers to Jenny in the letter
I recommended Linda's offer of a job to Richard in the letter
I recommended Mary's donation of money to Robert in the letter

example type: 17E (GEN (int. arg.) NOUN GEN (int. arg.) GEN (ext. arg.))

They discussed Jenny's gift of flowers from Steven at home
They discussed Richard's offer of a job from Linda at home
They discussed Robert's donation of money from Mary at home
We described Jenny's gift of flowers from Steven to the group
We described Richard's offer of a job from Linda to the group
We described Robert's donation of money from Mary to the group
I recommended Jenny's gift of flowers from Steven in the letter
I recommended Richard's offer of a job from Linda in the letter
I recommended Robert's donation of money from Mary in the letter

example type: 18E (DET NOUN GEN (int. arg.) GEN (ext. arg.) GEN (int. arg.))

They discussed the gift of flowers from Steven to Jenny at home
They discussed the offer of a job by Linda to Richard at home
They discussed the donation of money from Mary to Robert at home
We described the gift of flowers from Steven to Jenny to the group
We described the offer of a job by Linda to Richard to the group
We described the donation of money from Mary to Robert to the group
I recommended the gift of flowers from Steven to Jenny in the letter
I recommended the offer of a job by Linda to Richard in the letter
I recommended the donation of money from Mary to Robert in the letter

example type: 19E* (GEN (ext. arg.) GEN (int. arg.) GEN (int. arg.) NOUN)

They discussed Steven's to Jenny of flowers gift at home
They discussed Linda's to Richard of a job offer at home
They discussed Mary's to Robert of money donation at home
We described Steven's to Jenny of flowers gift to the group
We described Linda's to Richard of a job offer to the group
We described Mary's to Robert of money donation to the group
I recommended Steven's to Jenny of flowers gift in the letter
I recommended Linda's to Richard of a job offer in the letter
I recommended Mary's to Robert of money donation in the letter

example type: 20E* (GEN (ext. arg.) GEN (int. arg.) GEN (int. arg.) DET NOUN)

They discussed Steven's to Jenny of flowers the gift at home
They discussed Linda's to Richard of a job the offer at home
They discussed Mary's to Robert of money the donation at home
We described Steven's to Jenny of flowers the gift to the group
We described Linda's to Richard of a job the offer to the group
We described Mary's to Robert of money the donation to the group
I recommended Steven's to Jenny of flowers the gift in the letter
I recommended Linda's to Richard of a job the offer in the letter
I recommended Mary's to Robert of money the donation in the letter

example type: 21E* (GEN (ext. arg.) GEN (int. arg.) GEN (int. arg.) NOUN DET)

They discussed Steven's to Jenny of flowers gift the at home
They discussed Linda's to Richard of a job offer the at home
They discussed Mary's to Robert of money donation the at home
We described Steven's to Jenny of flowers gift the to the group
We described Linda's to Richard of a job offer the to the group
We described Mary's to Robert of money donation the to the group
I recommended Steven's to Jenny of flowers gift the in the letter
I recommended Linda's to Richard of a job offer the in the letter
I recommended Mary's to Robert of money donation the in the letter

KOREAN

MOD–NP category: noun modified by a determiner and adjectives

vocabulary set:

DET	ADJ	NOUN
그	큰, 푸른	책
그	작은, 빨간	공책
그	새, 검은	가방

frames:

영수가 교실에서 _____ 잃었다
미자가 의자에서 _____ 보았다
영희가 책상밑에서 _____ 발견했다

example type: 1K (DET ADJ ADJ NOUN)

영수가 교실에서 그 큰 푸른 책을 잃었다
영수가 교실에서 그 작은 빨간 공책을 잃었다
영수가 교실에서 그 새 검은 가방을 잃었다
미자가 의자에서 그 큰 푸른 책을 보았다
미자가 의자에서 그 작은 빨간 공책을 보았다
미자가 의자에서 그 새 검은 가방을 보았다
영희가 책상밑에서 그 큰 푸른 책을 발견했다
영희가 책상밑에서 그 작은 빨간 공책을 발견했다
영희가 책상밑에서 그 새 검은 가방을 발견했다

example type: 2K (ADJ DET ADJ NOUN)

영수가 교실에서 큰 그 푸른 책을 잃었다
영수가 교실에서 작은 그 빨간 공책을 잃었다
영수가 교실에서 새 그 검은 가방을 잃었다
미자가 의자에서 큰 그 푸른 책을 보았다
미자가 의자에서 작은 그 빨간 공책을 보았다

미자가 의자에서 새 그 검은 가방을 보았다
영희가 책상밑에서 큰 그 푸른 책을 발견했다
영희가 책상밑에서 작은 그 빨간 공책을 발견했다
영희가 책상밑에서 새 그 검은 가방을 발견했다

example type: 3K (ADJ ADJ DET NOUN)

영수가 교실에서 큰 푸른 그 책을 잃었다
영수가 교실에서 작은 빨간 그 공책을 잃었다
영수가 교실에서 새 검은 그 가방을 잃었다
미자가 의자에서 큰 푸른 그 책을 보았다
미자가 의자에서 작은 빨간 그 공책을 보았다
미자가 의자에서 새 검은 그 가방을 보았다
영희가 책상밑에서 큰 푸른 그 책을 발견했다
영희가 책상밑에서 작은 빨간 그 공책을 발견했다
영희가 책상밑에서 새 검은 그 가방을 발견했다

example type: 4K* (ADJ ADJ NOUN DET)

영수가 교실에서 큰 푸른 책을 그 잃었다
영수가 교실에서 작은 빨간 공책을 그 잃었다
영수가 교실에서 새 검은 가방을 그 잃었다
미자가 의자에서 큰 푸른 책을 그 보았다
미자가 의자에서 작은 빨간 공책을 그 보았다
미자가 의자에서 새 검은 가방을 그 보았다
영희가 책상밑에서 큰 푸른 책을 그 발견했다
영희가 책상밑에서 작은 빨간 공책을 그 발견했다
영희가 책상밑에서 새 검은 가방을 그 발견했다

POS–NP category: a noun modified by a genitive-marked possessive nominal, which in turn is modified by a genitive-marked nominal

vocabulary set:

DET	GEN	NOUN
그	영수, 선생님	집
그	보희, 이웃	가게
그	아버니, 친구	사무실

frames:

우리가 파티 뒤에 _____ 갔다
내가 수업 뒤에 _____ 들렀다
그들이 오전에 _____ 도착했다

example type: 5K (GEN GEN NOUN)

우리가 파티 뒤에 영수의 선생님의 집을 갔다
우리가 파티 뒤에 보희의 이웃의 가게에 갔다
우리가 파티 뒤에 아버님의 친구의 사무실에 갔다
내가 수업 뒤에 영수의 선생님의 집을 들렀다
내가 수업 뒤에 보희의 이웃의 가게에 들렀다

내가 수업 뒤에 아버님의 친구의 사무실에 들렀다
그들이 오전에 영수의 선생님의 집을 도착했다
그들이 오전에 보희의 이웃의 가게에 도착했다
그들이 오전에 아버님의 친구의 사무실에 도착했다

example type: 6K* (DET NOUN GEN GEN)

우리가 파티 뒤에 그 집을 영수의 선생님의 갔다
우리가 파티 뒤에 그 가게에 보희의 이웃의 갔다
우리가 파티 뒤에 그 사무실에 아버님의 친구의 갔다
내가 수업 뒤에 그 집을 영수의 선생님의 들렀다
내가 수업 뒤에 그 가게에 보희의 이웃의 들렀다
내가 수업 뒤에 그 사무실에 아버님의 친구의 들렀다
그들이 오전에 그 집을 영수의 선생님의 도착했다
그들이 오전에 그 가게에 보희의 이웃의 도착했다
그들이 오전에 그 사무실에 아버님의 친구의 도착했다

example type: 7K* (DET NOUN GEN GEN)

우리가 파티 뒤에 그 집을 선생님의 영수의 갔다
우리가 파티 뒤에 그 가게에 이웃의 보희의 갔다
우리가 파티 뒤에 그 사무실에 친구의 아버님의 갔다
내가 수업 뒤에 그 집을 선생님의 영수의 들렀다
내가 수업 뒤에 그 가게에 이웃의 보희의 들렀다
내가 수업 뒤에 그 사무실에 친구의 아버님의 들렀다
그들이 오전에 그 집을 선생님의 영수의 도착했다
그들이 오전에 그 가게에 이웃의 보희의 도착했다
그들이 오전에 그 사무실에 친구의 아버님의 도착했다

example type: 8K (GEN GEN DET NOUN)

우리가 파티 뒤에 영수의 선생님의 그 집을 갔다
우리가 파티 뒤에 보희의 이웃의 그 가게에 갔다
우리가 파티 뒤에 아버님의 친구의 그 사무실에 갔다
내가 수업 뒤에 영수의 선생님의 그 집을 들렀다
내가 수업 뒤에 보희의 이웃의 그 가게에 들렀다
내가 수업 뒤에 아버님의 친구의 그 사무실에 들렀다
그들이 오전에 영수의 선생님의 그 집을 도착했다
그들이 오전에 보희의 이웃의 그 가게에 도착했다
그들이 오전에 아버님의 친구의 그 사무실에 도착했다

example type: 9K* (GEN GEN NOUN DET)

우리가 파티 뒤에 영수의 선생님의 집을 그 갔다
우리가 파티 뒤에 보희의 이웃의 가게에 그 갔다
우리가 파티 뒤에 아버님의 친구의 사무실에 그 갔다
내가 수업 뒤에 영수의 선생님의 집을 그 들렀다
내가 수업 뒤에 보희의 이웃의 가게에 그 들렀다
내가 수업 뒤에 아버님의 친구의 사무실에 그 들렀다
그들이 오전에 영수의 선생님의 집을 그 도착했다
그들이 오전에 보희의 이웃의 가게에 그 도착했다
그들이 오전에 아버님의 친구의 사무실에 그 도착했다

2ARG–NP category: nominal phrase containing an external argument plus one subcategorized internal argument

vocabulary set:

DET	GEN	NOUN
그	일본, 서울	파괴
그	보희, 창호	비판
그	철수, 미자	사진

frames:

우리가 수업에서 ____ 토론했다
내가 사람들에게 ____ 기술했다[1]
그들이 편지에 ____ 이야기했다

example type: 10K* (GEN (ext. arg.) NOUN GEN (int. arg.))

우리가 수업에서 일본의 파괴를 서울의 토론했다
우리가 수업에서 보희의 비판을 창호의 토론했다
우리가 수업에서 철수의 사진을 미자의 토론했다
내가 사람들에게 일본의 파괴를 서울의 기술했다
내가 사람들에게 보희의 비판을 창호의 기술했다
내가 사람들에게 철수의 사진을 미자의 기술했다
그들이 편지에 일본의 파괴를 서울의 이야기했다
그들이 편지에 보희의 비판을 창호의 이야기했다
그들이 편지에 철수의 사진을 미자의 이야기했다

example type: 11K* (GEN (int. arg.) NOUN GEN (ext. arg.))

우리가 수업에서 서울의 파괴를 일본의 토론했다
우리가 수업에서 창호의비판을 보희의 토론했다
우리가 수업에서 미자의 사진을 철수의 토론했다
내가 사람들에게 서울의 파괴를 일본의 기술했다
내가 사람들에게 창호의 비판을 보희의 기술했다
내가 사람들에게 미자의 사진을 철수의 기술했다
그들이 편지에 서울의 파괴를 일본의 이야기했다
그들이 편지에 창호의비판을 보희의 이야기했다
그들이 편지에 미자의 사진을 철수의 이야기했다

---

1 Ho-min Sohn (personal communication) points out that the Korean verb *kiswulhata,* which I intended as a counterpart to the English verb *describe* actually has a narrower definition. *Kiswulhata* implies that the description is in writing, thus, the way that this verb is used here, e.g.:

wuli-ka salamtul-eykey chelswu-uy pohuy-eykey-uy kkoch-uy senmul-ul kiswulhayssta.
we-NOM group-DAT [name]-GEN [name]-DAT-GEN flower-GEN gift-ACC describe-PST-SE
'We described [in writing] Chelswu's gift of flowers to Pohuy to the group.'

may have been semantically weird, causing some skew in the reactions of the Korean native speakers.

example type: 12K (DET NOUN GEN (int. arg.) GEN (ext. arg.))

우리가 수업에서 그 파괴를 서울의 일본의 토론했다
우리가 수업에서 그 비판을 창호의 보희의 토론했다
우리가 수업에서 그 사진을 미자의 철수의 토론했다
내가 사람들에게 그 파괴를 서울의 일본의 기술했다
내가 사람들에게 그 비판을 창호의보희의 기술했다
내가 사람들에게 그 사진을 미자의 철수의 기술했다
그들이 편지에 그 파괴를 서울의 일본의 이야기했다
그들이 편지에 그 비판을 창호의 보희의 이야기했다
그들이 편지에 그 사진을 미자의 철수의 이야기했다

example type: 13K (GEN (ext. arg.) GEN (int. arg.) NOUN)

우리가 수업에서 일본의 서울의 파괴를 토론했다
우리가 수업에서 보희의 창호의 비판을 토론했다
우리가 수업에서 철수의 미자의 사진을 토론했다
내가 사람들에게 일본의 서울의 파괴를 기술했다
내가 사람들에게 보희의 창호의 비판을 기술했다
내가 사람들에게 철수의 미자의 사진을 기술했다
그들이 편지에 일본의 서울의 파괴를 이야기했다
그들이 편지에 보희의 창호의 비판을 이야기했다
그들이 편지에 철수의 미자의 사진을 이야기했다

example type: 14K (GEN (ext. arg.) GEN (int. arg.) DET NOUN)

우리가 수업에서 일본의 서울의 그 파괴를 토론했다
우리가 수업에서 보희의 창호의 그 비판을 토론했다
우리가 수업에서 철수의 미자의 그 사진을 토론했다
내가 사람들에게 일본의 서울의 그 파괴를 기술했다
내가 사람들에게 보희의 창호의 그 비판을 기술했다
내가 사람들에게 철수의 미자의 그 사진을 기술했다
그들이 편지에 일본의 서울의 그 파괴를 이야기했다
그들이 편지에 보희의 창호의 그 비판을 이야기했다
그들이 편지에 철수의 미자의 그 사진을 이야기했다

example type: 15K* (GEN (ext. arg.) GEN (int. arg.) NOUN DET)

우리가 수업에서 일본의 서울의 파괴를 그 토론했다
우리가 수업에서 보희의 창호의 비판을 그 토론했다
우리가 수업에서 철수의 미자의 사진을 그 토론했다
내가 사람들에게 일본의 서울의 파괴를 그 기술했다
내가 사람들에게 보희의 창호의 비판을 그 기술했다
내가 사람들에게 철수의 미자의 사진을 그 기술했다
그들이 편지에 일본의 서울의 파괴를 그 이야기했다
그들이 편지에 보희의 창호의 비판을 그 이야기했다
그들이 편지에 철수의 미자의 사진을 그 이야기했다

3ARG–NP category: nominal phrase containing an external argument plus two subcategorized internal arguments

vocabulary set:

DET	GEN	NOUN
그	철수, 보희에게, 꽃	선물
그	미자, 석기에게, 직장	제안
그	창호, 미자에게, 돈	기부

frames:

그들이 수업에서 ____ 토론했다
우리가 사람들에게 ____ 기술했다
내가 이 편지에 ____ 추천했다

example type: 16K* (GEN (ext. arg.) NOUN GEN (int. arg.) GEN (int. arg.))

그들이 수업에서 창호의 선물을 꽃의 보희에게의 토론했다
그들이 수업에서 영희의 제안을 직장의 석기에게의 토론했다
그들이 수업에서 미자의 기부를 돈의 철수에게의 토론했다
우리가 사람들에게 창호의 선물을 꽃의 보희에게의 기술했다
우리가 사람들에게 영희의 제안을 직장의 석기에게의 기술했다
우리가 사람들에게 미자의 기부를 돈의 철수에게의 기술했다
그들이 편지에 창호의 선물을 꽃의 보희에게의 추천했다
그들이 편지에 영희의 제안을 직장의 석기에게의 추천했다
그들이 편지에 미자의 기부를 돈의 철수에게의 추천했다

example type: 17K* (GEN (int. arg.) NOUN GEN (int. arg.) GEN (ext. arg.))

그들이 수업에서 보희에게의 선물을 꽃의 창호의 토론했다
그들이 수업에서 석기에게의 제안을 직장의 영희의 토론했다
그들이 수업에서 철수에게의 기부를 돈의 미자의 토론했다
우리가 사람들에게 보희에게의 선물을 꽃의 창호의 기술했다
우리가 사람들에게 석기에게의 제안을 직장의 영희의 기술했다
우리가 사람들에게 ㄱ 철수에게의 l부를 돈의 미자의 기술했다
그들이 편지에 보희에게의 선물을 꽃의 창호의추천했다
그들이 편지에 석기에게의 제안을 직장의 영희의 추천했다
그들이 편지에 철수에게의 기부를 돈의 미자의 추천했다

example type: 18K* (DET NOUN GEN (int. arg.) GEN (ext. arg.) GEN (int. arg.))

그들이 수업에서 그 선물을 꽃의 창호의 보희에게의 토론했다
그들이 수업에서 그 제안을 직장의 영희의 석기에게의 토론했다
그들이 수업에서 그 기부를 돈의 미자의 철수에게의 토론했다
우리가 사람들에게 그 선물을 꽃의 창호의 보희에게의 기술했다
우리가 사람들에게 그 제안을 직장의 영희의 석기에게의 기술했다
우리가 사람들에게 그 기부를 돈의 미자의 철수에게의 기술했다
그들이 편지에 그 선물을 꽃의 창호의 보희에게의 추천했다
그들이 편지에 그 제안을 직장의 영희의 석기에게의 추천했다
그들이 편지에 그 기부를 돈의 미자의 철수에게의 추천했다

example type: 19K (GEN (ext. arg.) GEN (int. arg.) GEN (int. arg.) NOUN)

그들이 수업에서 창호의 보희에게의 꽃의 선물을 토론했다
그들이 수업에서 영희의 석기에게의 직장의 제안을 토론했다
그들이 수업에서 미자의 철수에게의 돈의 기부를 토론했다
우리가 사람들에게 창호의 보희에게의 꽃의 선물을 기술했다
우리가 사람들에게 영희의 석기에게의 직장의 제안을 기술했다
우리가 사람들에게 미자의 철수에게의 돈의 기부를 기술했다
그들이 편지에 창호의 보희에게의 꽃의 선물을 추천했다
그들이 편지에 영희의 석기에게의 직장의 제안을 추천했다
그들이 편지에 미자의 철수에게의 돈의 기부를 추천했다

example type: 20K (GEN (ext. arg.) GEN (int. arg.) GEN (int. arg.) DET NOUN)

그들이 수업에서 창호의 보희에게의 꽃의 그 선물을 토론했다
그들이 수업에서 영희의 석기에게의 직장의 그 제안을 토론했다
그들이 수업에서 미자에게의 철수에게의 돈의 그 기부를 토론했다
우리가 사람들에게 창호의 보희에게의 꽃의 그 선물을 기술했다
우리가 사람들에게 영희의 석기에게의 직장의 그 제안을 기술했다
우리가 사람들에게 미자에게의 철수에게의 돈의 그 기부를 기술했다
그들이 편지에 창호의 보희에게의 꽃의 그 선물을 추천했다
그들이 편지에 영희의 석기에게의 직장의 그 제안을 추천했다
그들이 편지에 미자에게의 철수에게의 돈의 그 기부를 추천했다

example type: 21K* (GEN (ext. arg.) GEN (int. arg.) GEN (int. arg.) NOUN DET)

그들이 수업에서 창호의 보희에게의 꽃의 선물을 그 토론했다
그들이 수업에서 영희의 석기에게의 직장의 제안을 그 토론했다
그들이 수업에서 미자의 철수에게의 돈의 기부를 그 토론했다
우리가 사람들에게 창호의 보희에게의 꽃의 선물을 그 기술했다
우리가 사람들에게 영희의 석기에게의 직장의 제안을 그 기술했다
우리가 사람들에게 미자의 철수에게의 돈의 기부를 그 기술했다
그들이 편지에 창호의 보희에게의 꽃의 선물을 그 추천했다
그들이 편지에 영희의 석기에게의 직장의 제안을 그 추천했다
그들이 편지에 미자의 철수에게의 돈의 기부를 그 추천했다

ENGLISH P&P instructions

ENGLISH NATIVE SPEAKER INSTRUCTIONS

name _____ age _____

Speakers of a language seem to develop a 'feel' for what is a possible sentence, even in many cases where they have never been taught a particular rule.
For example, you might feel sentence number 1 below sounds like a good English sentence, while sentence number 2 is bad, even though you can't state a particular rule. Sentence number 3 probably sounds bad, but not as bad as number 2.

1. John is likely to win the race.
2. John the race likely is to win.
3. John is probable to win the race.

We are going to present you with list of sentences. We would like you to rate these sentences on a scale of bad to good, 1 being "bad" and 5 being "good". Use the numbers 2 through 4 for those sentences you are not sure about, like sentence number 3 above.

After you read the sentence, rate it by circling a number to the right of it, for example:

	bad ← → good
1. John is likely to win the race.	1 2 3 4 5
2. John the race likely is to win .	1 2 3 4 5
3. John is probable to win the race.	1 2 3 4 5

You would probably rate sentence number 3 as a "2", "3", or "4". There are no right or wrong answers, we just want to know how you, as a native speaker of English, feel about these sentences.

Before you begin, please tell us about any other language you speak or have studied. How did you learn the language? Did you study it formally or pick it up naturally? How old were you? Did you live in a place where the language was spoken? Are you studying a language currently? Please include all languages you have knowledge of.

_____
_____
_____
_____
_____
_____
_____
_____
_____
_____
_____
_____
_____
_____

Are you principally left-handed? ❑ yes ❑ no

Do you use your left hand for some things? ❑ yes ❑ no
(such as using tools, utensils, writing, throwing, playing tennis)

Is anyone in your immediate family left-handed? ❑ yes ❑ no
(mother, father, sister, brother)

Please take you time and read each sentence on the following pages carefully before you answer. Concentrate on the structure of the sentence. Ignore any problems with spelling, punctuation, etc. Mark only one answer for each sentence. Make sure you have marked all 63 sentences.

Thank you for your cooperation.

KOREAN P&P instructions (translation follows)

한국어 모국어 화자용 교본

이름 _____ 나이 _____

사람들은 따로 문법 교육을 받지 않고서도 모국어에서 가능한 문장과 그렇지 않은 것을 구분하는 능력을 갖고 있습니다. 예를들어, 구체적인 규칙을 말할 수는 없어도, 아래 주어진 문장 1 번은 바른 한국말 이지만 2 번은 아니다고 느낄 것입니다. 그런가하면, 3 번은 이상할 지 모르지만 2 번만큼 나쁘지는 않을 것입니다.

1. 창수의 머리가 좋다.
2. 머리가 창수의 좋다.
3. 머리가 창수가 좋다.

여러분들에게 주어진 문장 리스트를 보고 나쁜 문장에서 좋은 문장 스케일 1 번에서 5 번 중 여러분의 직관과 일치하는 번호를 하나 선택 하시오. 스케일의 1 번은 나쁜 문장이고 5 번은 좋은 문장이며 2 번, 3 번, 4 번은 위의 3 번째예와 같이 확실치 않은 경우 입니다. 주어진 문장을 읽고, 오른쪽의 번호 중 여러분의 직관에 맞는 문법 스케일 번호를 하나 골라 동그라미를 치시오. 아래 예를 참고 하시오.

	나쁨 ←			→ 좋음
1. 창수의 머리가 좋다.	1	2	3	4	(5)
2. 머리가 창수의 좋다.	(1)	2	3	4	5
3. 머리가 창수가 좋다.	1	2	3	4	5

3 번 문장은 아마 문법 스케일 2 번, 3 번 아니면 4 번을 선택 했을 것입니다. 여기서 맞고 틀린 답은 없으며, 우리가 알고자 하는 것은 단지 여러분들이 한국어 모국어 사용자로써 주어진 문장들을 문법 스케일 상에서 어떻게판단 하는가 하는 것 입니다.

시작하기 전에, 여러분들이 아는 외국어에 대해서 아래에 주어진 여백에 묻는대로 대답하시오.

외국어를 배운 적이 있습니까? 아래에 해당하는 곳에 가위표를 하시오.

☐ 예 ☐ 아니오

외국를 배운적이 있으면, 그 외국어에 관한 정보가 필요합니다. 그 언어를 어떻게 배웠으며 얼마나배웠습니까? 그리고, 배우기 시작 했을 때 몇살이었으며 몇년 동안이나 배웠습니까? 지금도 외국어를 배우고 있습니까? 아는 외국어를 전부 기술하시오.

_____

_____

_____

_____

_____

_____

왼손 잡이 입니까? ☐ 예 ☐ 아니오

부분적으로, 왼손잡이입니까? ☐ 예 ☐ 아니오
(예를들어, 도구를 다루거나 공을 던질때)

가족중에 왼손잡이가 있습니까? ☐ 예 ☐ 아니오

시간을 충분히 갖고, 여러분이 답하기 전에 다음페이지에 있는 문장들을 잘읽어야 합니다. 문장의 맞춤법이나 문장부호가 틀린 것은 무시합니다. 대신에 여러 분들이 이 문장들을 들었을 때 얼마나 자연스러운 한국말로 느낄 것인 지에 집중해서 답합니다. 각 문장에 하나의 문법 스케일 번호에 만 동그라미 하시오. 주어진 63 개 문장 모두 빠짐없이 답하시오.

여러분의 협조에 감사합니다.

Translation of KOREAN P&P instructions

MANUAL FOR NATIVE SPEAKERS OF KOREAN

name _____ age _____

People have a linguistic ability in their native languages to distinguish grammatical sentences from ungrammatical ones without any grammar training. For example, without being able to articulate why, you might judge that sentence 1 below is a grammatical Korean sentence, but sentence 2 is not, and that sentence 3 is not as bad as sentence 2.

1. Changsoo's head is good. *[lit. 'Changsoo is smart]*
2. Head Changsoo's is good. *[ungrammatical version of sentence 1]*
3. Head Changsoo is good. *[ungrammatical version of sentence 1]*

Read the list of sentences given to you and select one of the numbers on the grammaticality scale from 1 to 5, in the order of worst to best, corresponding to your grammatical judgment. Refer to the following example.

	bad ← → good
1. Changsoo's head is good.	1 2 3 4 5
2. Head Changsoo's is good.	1 2 3 4 5
3. Head Changsoo is good.	1 2 3 4 5

Probably you chose the grammaticality value "2", "3", or "4" for sentence 3. There are no right or wrong answers; what we want to find out is how you, as a native speaker of Korean, judge the grammaticality of a given sentence on the grammaticality scale.

Before we start, please answer the questions about foreign languages you already know in the blank spaces provided below.

Have you learned any foreign language? Put an "X" in the appropriate box below.

❑ yes ❑ no

If you have learned any foreign language, we need information about your language experience. How did you learn the language and how long did you learn it? How old were you when you started learning the language? Are you still learning the language? Please list all foreign languages.

Are you left-handed? ❑ yes ❑ no

Are you partially left-handed? (for example, when you use tools or throw a ball)

❑ yes ❑ no

Do you have any family member who is left-handed? ❑ yes ❑ no

Take your time and read the each sentence on the next page carefully before you answer. Please disregard any misspellings or incorrect punctuation. Concentrate on how natural the given Korean sentence is when you read it and then answer. Please circle only one value on the grammaticality scale. There are 63 sentences in all. Please answer all of them.

We appreciate your help.

APPENDIX C: RECORDED INSTRUCTIONS FOR EI TASK

ENGLISH EI instructions transcript

This is a test to see how well you can remember and repeat EXACTLY what you have just heard. You will hear a sentence spoken. Sometimes the sentence will sound fine, other times it may sound a bit strange. You must repeat it, word for word, as accurately as possible. You may take as much time as you like. Listen now and hear an example of what we would like you to do:

[the following sentences are used to model the elicitation/imitation procedure — the instructor's voice is heard reading each elicitation in turn and a different voice is heard producing a correct imitation]

Mary went to the house of a friend after school.
Jane found a dollar behind the sofa last week.
Mary went to a friend's house the after school.
I showed Steven's of his mother drawing to a friend.

Notice that the person repeated the sentences exactly as I said them, even though some of the words may have seemed out of place. Even though you cannot remember all of the sentence, it is important to repeat everything you DO remember.

Now we are ready to begin. There are about 60 sentences in all. Please remember to repeat EXACTLY what you hear. Here is your first sentence:

KOREAN EI instructions transcript (translation follows)

이 테스트는 여러분들이 기억을 얼마나 잘할 수 있는가를 보고자 하는 것입니다. 들은 대로 정확하게 따라해야 합니다. 여러분들은 녹음된 문장을 들을 것입니다. 어떤때는 녹음된 문장이 이상한 문장일 수도 있습니다. 가능한 정확하게 들은 그대로 따라해야 합니다. 시간은 얼마든지 여러분이 원하는만큼 걸려도 좋습니다. 그러면 여러분이 해야할 것을 연습해 봅시다.

[the following sentences are used to model the elicitation/imitation procedure —the instructor's voice is heard reading each elicitation in turn and a different voice is heard producing a correct imitation]

영희가 수업후 친구의 집에 갔다.
한국이 작년에 많은 차를 유럽에 수출했다.
영희가 수업후 집에 친구의 갔다.
앞집의 우는 그 개가 아기를 미자의 놀라게 했다.

문장들이 자연스럽지 못하더라도, 연습에서 그대로 따라하는 것에 주목하기 바랍니다. 그리고, 비록 문장 전부를 기억하지 못할지라도, 기억하는 모든 것을 반복하는 것이 중요합니다.

자, 지작할 준비가 되었습니까? 모두 60 여개 문장이 있습니다. 여러분이 들은 것을 정확하게 따라 하시오. 그러면 지금부터 시작하겠습니다.

Translation of KOREAN EI instructions

This test is to check how well you remember. You must repeat what you hear as exactly as possible. You will hear recorded sentences. Sometimes the sentence you hear may sound awkward. Repeat what you hear as is. You can take as much time as you want. Well, then, let's practice what you are to do.

Younghuy went to a friend's house after class.
Last year many cars?? were imported in Korea.
Younghuy went to house a friend's after class.
The front house's barking the dog frightened child Mica's.

Even though some sentences are not natural, please note that the person repeated what she heard as it was. Even if you do not remember all that you hear, it is important to repeated everything that you do remember.

Are you ready to start? There are 60 sentences all together. Repeat what you hear as exactly as possible. Let's start now.

APPENDIX D: INTERACTIVE INSTRUCTIONS FOR RT TASK

ENGLISH RT instructions screens

The following Figures 1 though 12 show the instruction screens for the RT instrument. With the supervision of a researcher, the participant guides him- or herself through the instructions for the RT task. After being welcomed to program (Figure 1), the participant is shown what to expect: first, the priming sentence appears at the bottom of the screen (Figure 2), then the pointing finger indicates in which of four random locations the pair sentence will appear (Figure 3). Following, the use of the red and blue keys are explained (Figure 4) and the participant practices pressing these keys. When the blue key is pressed the pop-up message in Figure 5 is displayed; when red is pressed, the message in Figure 6 is displayed; and when a key other than red or blue is pressed the message in Figure 7 is displayed. Next, the participant tries pressing the space bar and an encouraging message is displayed (Figure 8). Before the practice session begins (Figure 9), the participant has a chance to ask questions of the researcher. The following practice session contains five pairs of filler sentences and may be repeated if desired. Following the practice session is another opportunity for the participant to ask questions; finally, the participant enters his or her name and initiates the actual task by pressing the space bar.

Figure 5

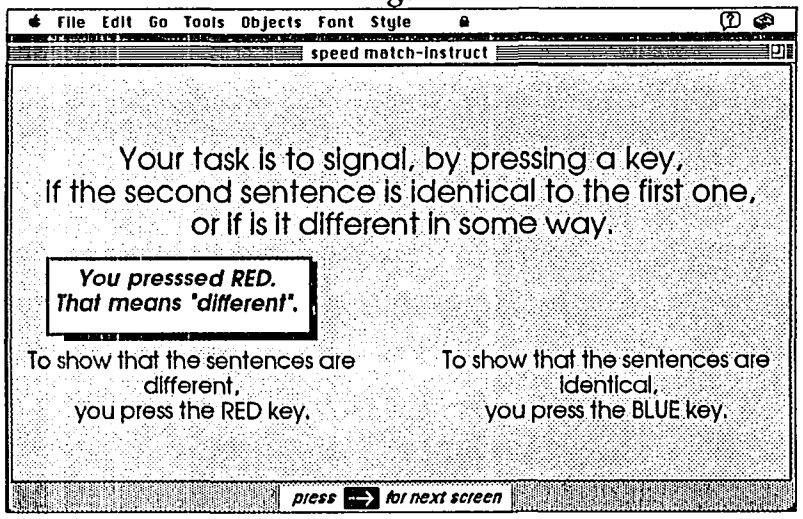


Figure 6

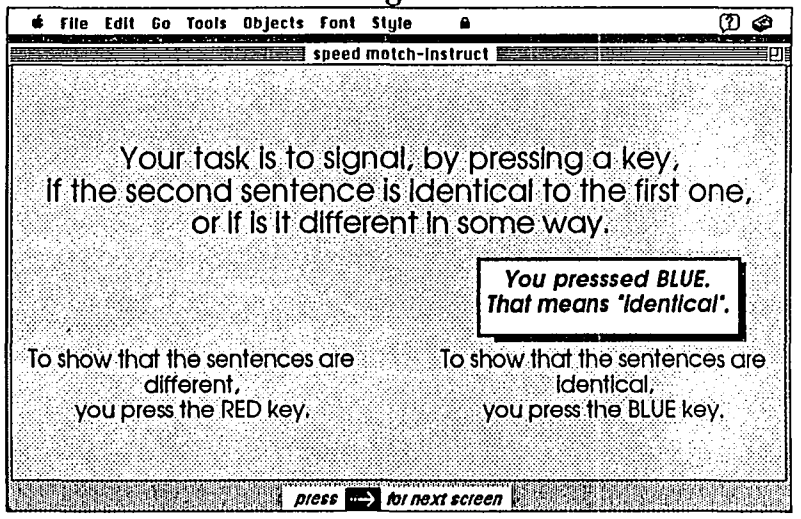


Figure 7

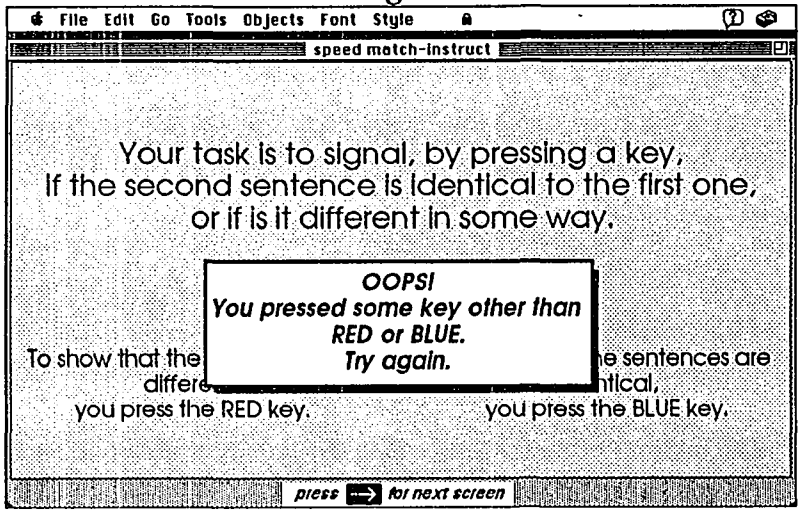


Figure 8

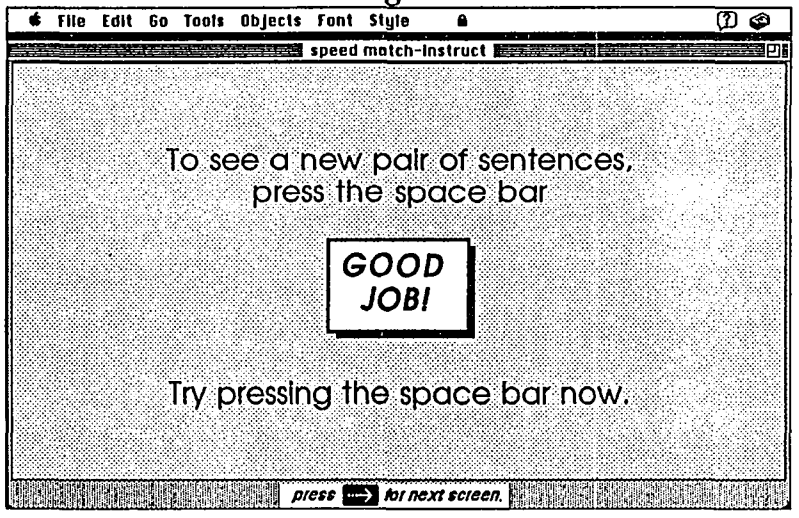


Figure 1

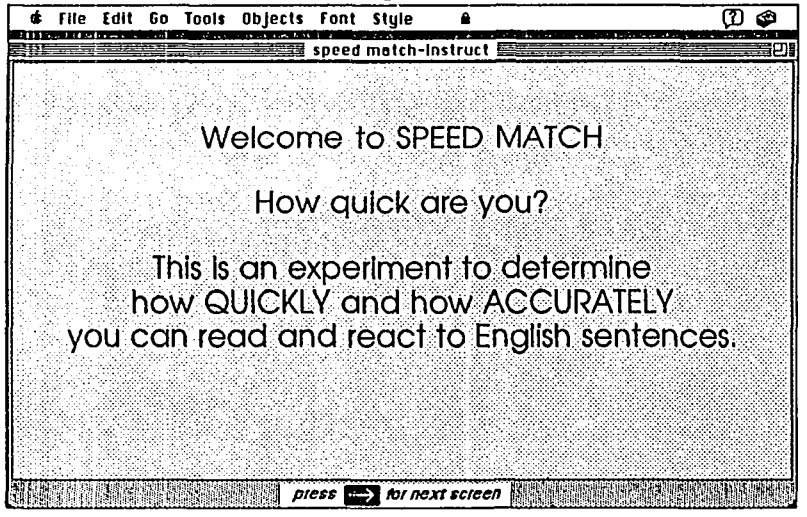


Figure 2

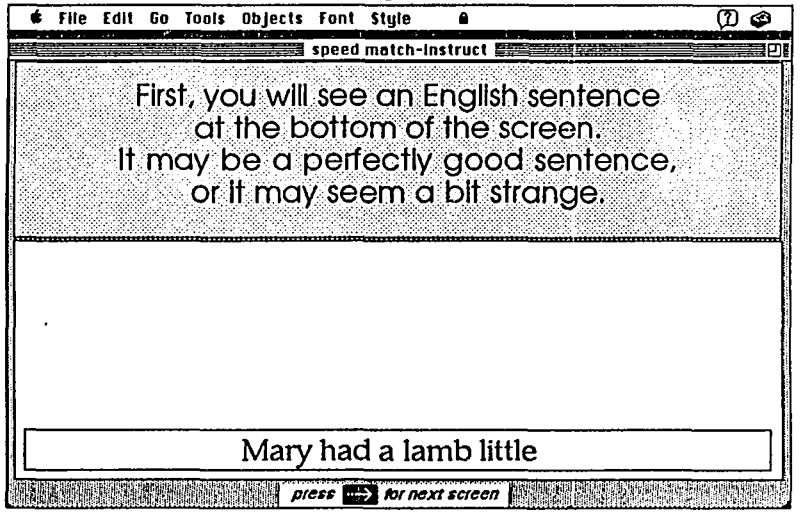


Figure 3

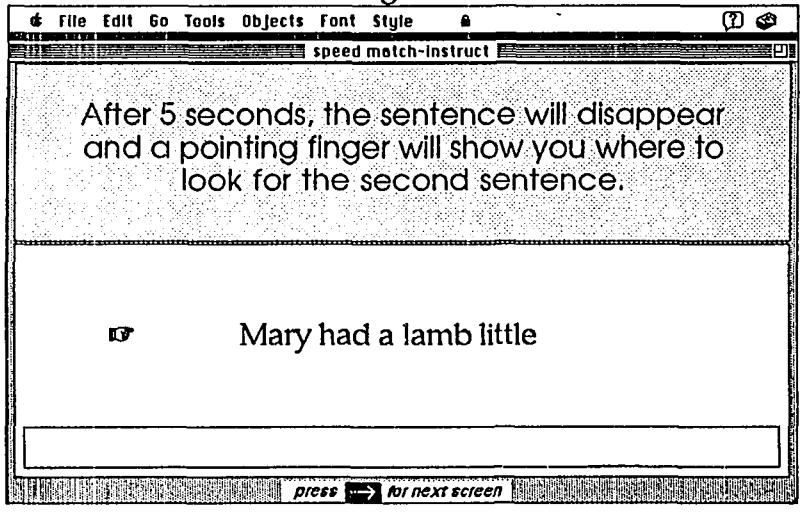


Figure 4

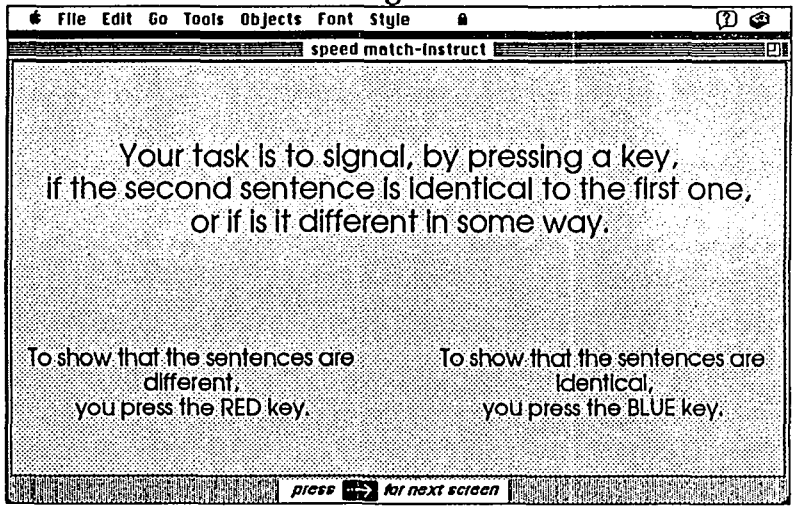


Figure 9

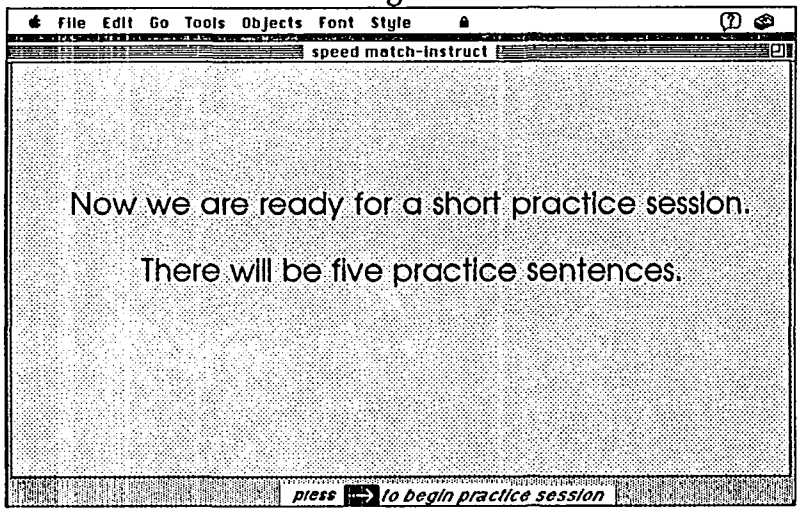


Figure 10

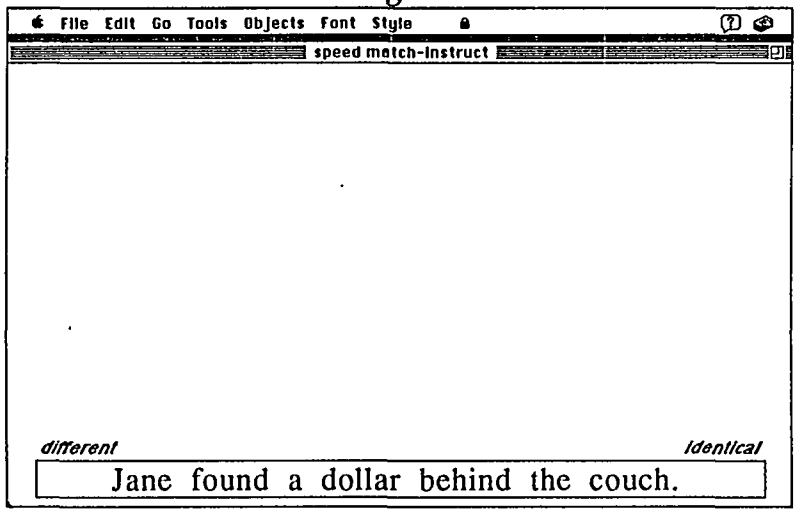


Figure 11

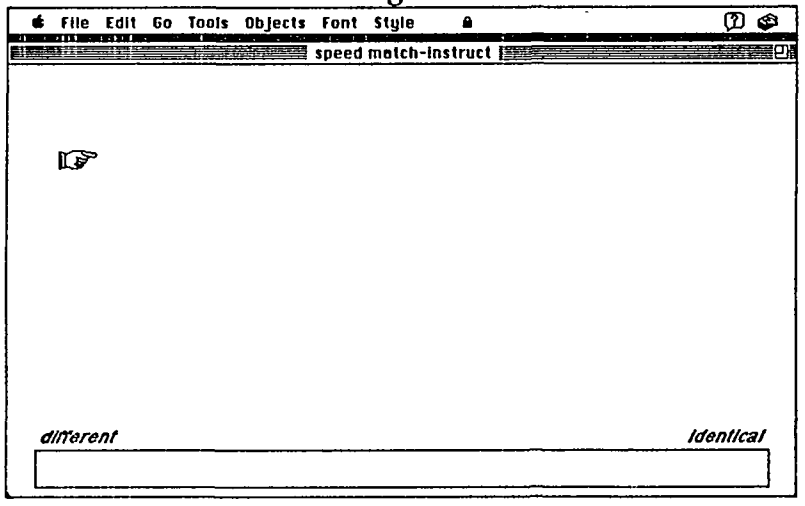
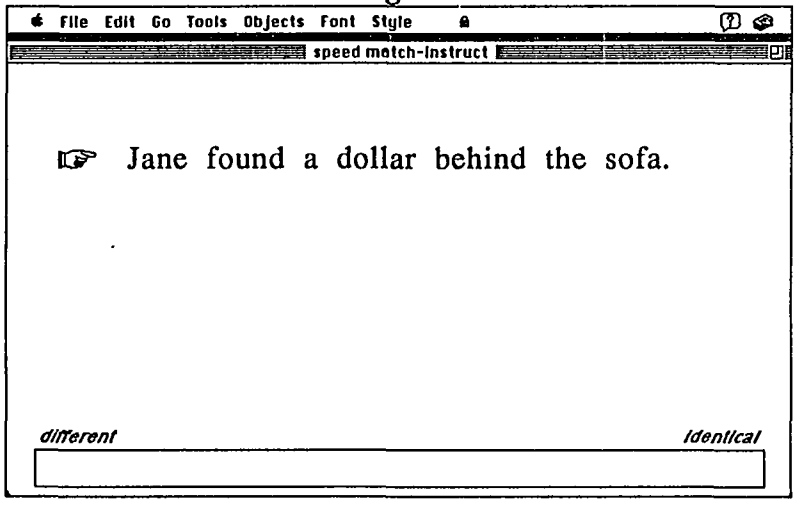


Figure 12



KOREAN RT instructions screens

The instruction screens for the Korean the RT instrument looked like the English instructions, but were written in hangul script. The text of the instructions plus a transcription appears below:

속도 비교 게임이 여러분의 참여를 환영합니다
The Speed Match game welcomes your participation

얼마나 빨리 반응할 수 있습니까?
How fast can you respond?

이실험은 즈어진 한국어 문장을 여러분이 읽고, 얼마나 정확하고 빨리 비교 반응을 할 소 있는 가를 보는 것입니다 .
This experiment is for seeing how fast and correctly you can compare two sentences and then react as to whether or not the two sentences are identical when you read the given Korean sentences.

처음에, 스크린 밑부분에 한국어 문장이 나타남니다. 그 문장은 좋은 한국어 일 수도, 이상한 한국어 일 수도 있입니다.
First, a Korean sentence will appear at the bottom of the screen. The sentence may be grammatically correct Korean or awkward Korean.

머리가 장수의 좋다
Head Changsoo's is good.

5초후어, 첫 문장은 사라지고, 손 가락 표시가 나타나면서, 두번째 문장이 나타날곳을 가르킵니다.
After 5 seconds, the first sentence will disappear and a hand with a pointing finger will appear to indicate where the second sentence will appear.

머리가 장수의 좋다
Head Changsoo's is good.

두번째 문장이 첫문장과 동일한지를 여러분은 키를 눌러 표시합니다.
You may express whether or not the two sentences are identical by pressing a key.

두 문장이 다르면,
빨강키를 누릅니다.
If the two sentences differ,
press the red key.

두 문장이 동일하면,
파랑키를 누릅니다.
If the two sentences are identical,
press the blue key.

새로운 두문장을 보기 위해 스페이스 바를 누릅니다.
Press the space bar to see two new sentences.

자, 지금 스페이스 바를 눌러봅시다.
OK, let's press the space bar now.

자, 이제 짧은 연습을 해볼 준비가 되었습니다.
OK, we are ready to do a short practice.

APPENDIX E: PRACTICE AND DISTRACTOR ITEMS FOR RT TASK

ENGLISH

practice items

Jane found a dollar behind the sofa.
Jane found a dollar behind the couch.

We went to the dog races last Saturday afternoon.
We went to the dog races last Saturday afternoon.

They ordered the of cockroaches extermination last week.
They ordered the of cockroaches extermination last week.

He admired James' portrait of the Queen very much.
He admired James' portrait of the Queen very much.

I read already interesting that book.
I wrote already interesting that book.

distractor items

Robert seems to be climbing a mountain
David seems to be climbing a mountain

Steven was discharged from the hospital yesterday
Steven was discharged from the hospital tomorrow

There is no reason why John should come here
There is no reason why John should enter here

I don't know whether Jenny is a doctor or a nurse.
I don't care whether Jenny is a doctor or a nurse.

I don't know whether it is spring or summer
I don't know whether it is spring or autumn

The price differs according to the barbershop
The price differs according to the barbershop

The fact is that Linda didn't get an injection yesterday
The fact is that Linda didn't get an injection yesterday

I have no time, moreover, I have no desire to go
I have no time, moreover, I have no desire to go

Two years ago John passed the higher civil service examination
Two years ago John passed the higher civil service examination

According to the weather forecast, it will clear up in the afternoon
According to the weather forecast, it will clear up in the afternoon

Susan was going to read a book, but Jenny dropped by
Susan was going to read a book, but Jenny dropped by

Because Linda is such a good teacher, she is well liked
Because Linda is such a good teacher, she is well liked

As long as David is working in this company, there is no problem
As long as David is working in this company, there is no problem

It doesn't matter to me if Richard has love affairs or not
It doesn't matter to me if Richard has love affairs or not

How wonderful it would be if we had a lot of time
How wonderful it would be if we had a lot of time

KOREAN

practice items

앞집의 우는 그 개가 미자의 아기를 놀라게 했다.
앞집의 우는 그 개가 미자의 아기를 놀라게 했다.

한국이 작년에 많은 차를 유럽에 수출했다.
한국이 작년에 많은 차를 유럽에 수출했다.

영희가 친구들에게 헌 책을 팔았다.
영희가 선생님에게 헌 책을 팔았다.

영희가 친구들에게 헌 책을 팔았다.
영희가 친구들에게 헌 책을 팔았다.

영희가 할머니의 새 집을 어제 방문했다.
영희가 할머니의 새 집을 내일 방문했다.

distractor items

석기가 등산하는가 보다
창호가 등산하는가 보다

창호는 어제 병원에서 퇴원했다
창호는 내일 병원에서 퇴원했다

그 사람이 여기에 올 리가 있읍니까?
그 사람이 여기에 갈 리가 있읍니까?

영희가 의사인지 간호원인지 모르겠다
영희가 의사인지 간호원인지 알겠다

봄인지 여름인지 모르겠다
봄인지 가을인지 모르겠다

이발소에 따라서 값이 다르다
이발소에 따라서 값이 다르다

미자가 어제 주사를 맞지않았다는 것이다
미자가 어제 주사를 맞지않았다는 것이다

시간도 없거니와 갈 마음이 없다
시간도 없거니와 갈 마음이 없다

철수가 이년전에 고등고시에 합격 했다
철수가 이년전에 고등고시에 합격 했다

일기 예보에 의하면 오후엔 개인다고 한다
일기 예보에 의하면 오후엔 개인다고 한다

미자 가 이 책을 읽으려니가, 보희가 찾아왔다
미자 가 이 책을 읽으려니가, 보희가 찾아왔다

영희가 어찌나 좋은 선생님지 모두 좋아한다
영희가 어찌나 좋은 선생님지 모두 좋아한다

철수가 이 회사에서 일하고 있는한 문제 없다
철수가 이 회사에서 일하고 있는한 문제 없다

영수가 연애를 하거나 말거나 저한텐 상관 없다
영수가 연애를 하거나 말거나 저한텐 상관 없다

시간이 많이 있었으면 얼마나 좋겠니
시간이 많이 있었으면 얼마나 좋겠니

REFERENCES

Abney, S. (1987). *The English noun phrase and its sentential aspect.* Doctoral dissertation. (Distributed by MIT Working Paper in Linguistics, Room 20D–219, MIT, Cambridge, MA).

Abney, S. (1991). Parsing by chunks. In R. C. Berwick, et al. (Eds.), *Principle-based parsing: Computation and psycholinguistics* (pp. 257–278). Netherlands:Kluwer Academic Publishers.

Adjémian, C. & Liceras, J. (1984). Accounting for adult acquisition of relative clauses: Universal Grammar, L1, and structuring the intake. In F. Eckman, L. Bell, & D. Nelson (Eds.), *Universals of second language acquisition* (pp. 101–118). Rowley, MA: Newbury House.

Ahn, H. D. (1988). Preliminary remarks on Korean NP. In *Proceedings of the Sixth International Conference of Korean linguists*. Seoul: Hanshin Publishing.

Ahn, H. D. & H.J. Yoon (1989). Functional categories in Korean. In S. Kuno, et al. (Eds.), *Harvard Studies in Korean Linguistics III* (pp. 79–88). Hanshin Publishing.

Amtserdam, L., Ammon, P., & Simons, H. (1990). Children's elicited imitations of controlled and rewritten reading texts. *Journal of Educational Psychology 82*(3), pp. 486–490.

Arthur, Bradford (1980). Gauging the boundaries of second language competence: a study of learner judgements. *Language Learning 30* (1), 177–194.

Barwise, J. & Cooper, R. (1981). Generalized quantifiers and natural language. *Linguistics and Philosophy 4*. 159–219.

Bauer, P. & Thal, D. (1990). Scripts or scraps: Reconsidering the development of sequential understanding. *Journal of Experimental Child Psychology 50*(2), pp. 287–384.

Beretta. A. (1991). Determinants of individual differences during skill acquisition: Cognitive abilities and information processing. *Journal of Experimental Psychology, 17* (3), 288–318.

Bever, T. (1970). The cognitive basis for linguistic structures. In J. R. Hayes (Ed.), *Cognition and the development of language* (pp. 341–438). New York: John Wiley and Sons.

Bever, T. G. & McElree, B. (1988). Empty categories access their antecedents during comprehension. *Linguistic Inquiry, 19* (1), 35–43.

Bialystock, E. (1981). Explicit and implicit judgments of L2 grammaticality. *Language Learning, 29* (1), 81–102.

Bialystock, E. (1986). Factors in the growth of linguistic awareness. *Child Development 57*, 498–510.

Bills, A. G. (1931). Blocking: A new principle in mental fatigue. *American Journal of Psychology 43*, 230–245.

Birdsong, D. (1989). *Metalinguistic performance and interlinguistic competence.* New York: Springer-Verlag.

Birdsong, D., Johnson, C., McMinn, J. & Ingmundson, P. (1985). *Universals versus transfer revisited.* Paper presented at the North Boston University Conference on Language Development. (ERIC document Reproduction Service No. ED 255 061) Boston University, Boston. MA.

Birdsong, D. (1984). *Variables in variation.* Paper presented to the Foreign Language Education Center Student Association, University of Texas at Austin.

Bley-Vroman, R. (1988). *MATCH.* [Computer program]. Department of ESL, University of Hawai'i.

Bley-Vroman, R. & Chaudron, C. (1990). Second language processing of subordinate clauses and anaphora — First language and universal influences: A review of Flynn's research. *Language Learning 40*, 2, . 245–285.

Bley-Vroman, R. & Chaudron, C. (in press). *Is elicited imitation an adequate measure of L2 competence?* In S. Gass. E. Tarone, & A. Cohen (Eds.), *Research methods in second language acquisition.* Hillsdale, NJ: Erlbaum.

Bley-Vroman, R. & Houng, C. (1988). Why do Chinese produce few relative clauses in English? In *University of Hawai'i Working Papers in ESL 7* (2), 93–98.

Bley-Vroman, R., Felix, S. & Ioup, G. (1988). The accessibility of universal grammar in adult language learning. *Second Language Research 4* (1), 1–32.

Bley-Vroman, R. & Masterson, D. (1989). Reaction time as a supplement to grammaticality judgements in the investigation of second language learners' competence. *University of Hawai'i Working Papers in ESL 8* (2), 207–237.

Botha, R. (1981). *The conduct of linguistic inquiry: A systematic introduction to the methodology of generative grammar*. The Hague: Mouton.

Bowers, J. (1987). Extended X' theory, the ECP and the Left Branch Condition. WCCFL.

Boyd, L. (1992, March 15). Grab Bag. *San Francisco Chronicle,* Sunday Punch section, p. 5.

Braine, M. (1971). On two types of models in the internalization of grammars. In D. Slobin (Ed.), *The ontogenesis of grammar* pp. (153–186). New York: Academic Press.

Brame, M. 1982. The head-selector theory of lexical specifications and the nonexistence of coarse categories. Linguistic Analysis 10, vol. 4.

Byrne, B. & Davidson, E. (1985). On putting the horse before the cart: Exploring conceptual bases of word order via acquisition of a miniature artificial language. *Journal of Memory and Language 24*, 377–389.

Carden, G. & Dieterich, T. (1981).Introspection, observation, and experiment: An example where experiment pays off. In D. Asquith & R. N. Giere, (Eds.),*Proceedings of the 1980 Biennial Meeting of the Philosophy of Science Association,* vol. 2 (pp. 583–597). East Lansing, MI: Philosophy of Science Association

Carrithers, C. (1985). *The comprehension of canonical and non-canonical sentence constructions.* Unpublished doctoral dissertation, Columbia University, New York.

Carroll, J., Bever, T., & Pollack, C. (1981). The non-uniqueness of linguistic intuitions. *Language, 57* (2), 368–382.

Carlson,, G. & Tanenhaus, M. K. (Eds.). (1989). *Linguistic structure in language processing.* Amsterdam: Kluwer.

Chaudron, C. & Parker, K. (1990). Discourse markedness and structural markedness. *SSLA 12*, 43–64.

Chaudron, C. & Russell, G. (1991). *The validity of elicited imitation as a measure of second language competence*. Unpublished manuscript, University of Hawai'i, Honolulu, HI.

Chaudron, C. (1983). Research on metalinguistic judgments: A review of theory, methods and results. *Language Learning, 33* (3), 343–77.

Chomsky, N. (1957). *Syntactic structures*. The Hague: Mouton.

Chomsky, N. (1964). Current issues in linguistic theory. In A. Fodor & J. Katz (Eds.), *The Structure of language: Readings in the philosophy of language*, (pp. 50–118). Englewood Cliffs, N.J.: Prentiss Hall.

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge: MIT Press.

Chomsky, N. (1980). On binding. *Linguistic Inquiry 11*, 1–46.

Chomsky, N. (1981). *Lectures on Government and Binding*. Dordrecht: Foris.

Clahsen, H. & Hong, U. (1993). *Grammatical constraints in syntactic processing: Sentence matching experiments in German*. (Theorie des Lexikons: Arbeiten des Sonderforschungsbereichs 282, Nr. 43). Düsseldorf: Heinrich-Heine-Universität Düsseldorf, Seminar für Allgemeine Sprachwissenschaft.

Cloitre, M. (1984). The effects of kind of anaphor on accessibility to antecedent information. Unpublished doctoral dissertation, Columbia University, New York.

Cohen, A. D., & Robbins, M. (1976). Towards assessing interlanguage performance: The relationship between selected errors, learners' characteristics, and learners' explanations. *Language Learning, 26*, 45–62.

Coleman, E.B. (1965). Responses to a scale of grammaticalness. *Journal of Verbal Learning and Verbal Behavior 4*, 521–527.

Cowan, J. R. & Abe, Y. (1991). Investigating the reliability of native speaker and L2 learner judgments about sentences. Unpublished manuscript, University of Illinois/Purdue University.

Cowart, W. (1985). Notes of the biology of syntactic processing. *Journal of Psycholinguistic Research, 18* (1), 89-103.

Crain, S. & Fodor, J. D. (1987). Sentence matching and overgeneration. *Cognition 26*, 123–169.

Crookes. G. (1992). Theory formats in SLA. *Studies in Second Language Acquisition, 14*, 425–450.

Denzen, N. (1989). *The Research Act.* (3rd ed.). (pp. 234–247). Englewood Cliffs, NJ: Prentice-Hall.

DeVellis, R. (1991). *Scale development: Theories and applications.* Newbury Park, CA.: Sage.

deVilliers, J. G., & deVilliers, P. A. (1974). Competence and performance in child language: Are children really competent to judge? *Journal of Child Language, 1,* 11-22.

d'Olwer, L. N. (1987). *Fray Bernardino de Sahagun (1499-1590)* (M. J. Mixco, Trans.). Salt Lake City: University of Utah Press. (Original work published 1952).

Dowty, D., Kartunnen, L., & Zwicky. A. (1985). *Natural language processing: Psychological, computational, and theoretical perspectives.* Cambridge: Cambridge University Press.

duPlessis, J., Solin, D., Travis, L., & White, L. (1987). UG or not UG, that is the question: A reply to Clahsen and Muysken. *Second Language Research , 3* (1), 56–75.

Ehri, L. C. & Wilce, L. C. (1980). The influence of orthography on readers' conceptualization of the phonemic structure of words. *Applied Psycholinguistics, 1,* 371–385,

Ellis, R. (1991). Grammaticality judgments and second language acquisition. *SSLA 13,* 161–181.

Eubank, L. (1991). *Is sentence matching sensitive to constraints?.* Unpublished manuscript, University of North Texas, Denton.

Eubank, L. (in press). Sentence matching and processing in L2 development. *Second Language Research.*

Felix S. (1988). UG-generated knowledge in adult second language acquisition. In S. Flynn & W. O'Neil (Eds.), *Linguistic theory in second language acquisition* (pp. 277–294). Dordrecht: Kluwer Academic.

Finer, D. & Broselow, E. (1986). Second language acquisition of reflexive binding. *Proceedings of the Sixteenth Annual Meeting of the North East Linguistic Society.* (pp. 86–96). Amherst: University of Massachusetts.

Fischer, S. D., Albertini, J. A., & Samar, V. J. (1987). *Story or no story: The effects of discourse context on elicited imitation.* Monograph: National Technical Institute for the Deaf.

Flynn, S. (1987). *A parameter setting model of L2 acquisition: Experimental studies in anaphora.* Dordrecht: Reidel.

Flynn, S. & Espinal, I. (1985). Head-initial/head-final parameter in adult Chinese L2 acquisition of English. *Second Language Research 1* (2), 93–117.

Forster, K. I. (1979). Levels of processing and the structure of the language processor. In W. E. Cooper & E. C. T. Walker (Eds.), *Sentence proccessing: Psycholinguistics studies presented to Merrill Garrett*. Hillsdale, NJ: Erlbaum.

Forster, K. I. (1987). Binding, plausibility, and modularity. In J. Garfield (Ed.), *Modularity in knowledge representation and natural language processing.* Cambridge, MA: MIT Press.

Freedman, S. (1982). *Behavioral reflexes of constraints on transformations.* Unpublished doctoral dissertation. Monash University, Clayton, Victoria, Australia.

Freedman, S. & Forster, K.I. (1985). The psychological status of overgenerated sentences. *Cognition 19*, 101–31.

Fujiki, M. & Willbrand, M. L. (1982). A comparison of our informal methods of language evaluation. *Language, Speech and Hearing Services in Schools, 13,* 42–52.

Fukui, N. & Speas, M. (1986). Specifiers and projection. *MIT Working Papers in Linguistics, (8)*, 128–172.

Gallimore, R. & Tharp, R. G. (1981). The interpretation of elicited sentence imitation in a standardized context. *Language Learning, 31* (2), 369–392.

Gass, S. (1983). The development of L2 intuitions. *TESOL Quarterly 17* (2), 273–291.

Gass, S. (1983). The development of L2 intuitions. *TESOL Quarterly 17*, 271–291.

Gee, J. & Grosjean, F. (1983). Performance structures: A psycholinguistic and linguistic appraisal. *Cognitive Psychology 15*, 291–325.

Gerken, L & Bever, T. (1986). Linguistic intuitions are the result of interactions between perceptual processes and linguistic universals. *Cognitive Science, 100*, 457–476.

Goldstien, H., Angelo, D. & Wetherby, B. (1987). Effects of training method and word order on adult's acquisition of miniature linguistic systems. *The Psychological Record*, January, 89–107.

Greenbaum, S. (Ed.). (1973). Informant elicitation of data on syntactic variation. *Lingua 31*, 201–212.

Greenbaum, S. (Ed.). (1977). *Acceptability in language*. The Hague: Mouton.

Grigg, T. (1986). The effects of task, time, and rule knowledge on grammar performance for three English structures. *University of Hawai'i Working Papers in English as a Second Language, 5* (1), 37–60.

Hagen, L. Kirk (1991, October). *Constructs and measurement in parameter-based models of second language acquisition*. Paper presented at the conference on theory construction and methodology in second language acquisition research, Michigan State University, East Lansing, MI.

Hale, K. (1985). On nonconfigurational structures. In S. Kuno, et al. (Eds.), *Harvard studies in Korean linguistics*, (pp. 62–70). Harvard University, Cambridge.

Hale, K. (1982). Preliminary remarks on configurationality. In J. Pestejovsky & P. Sells (Eds.), *Proceedings of the Twelfth Annual Meeting of the North East Linguistic Society* (pp. 86–96). Amherst: University of Massachusetts.

Hellan, L. (1986). The headedness of NPs in Norwegian. In P. Muysken and H. Van Reimsdijk (Eds.), Features and Projections, Foris, Dordrecht, pp. 89–122. [I have]

Henning, G. (1983). Oral proficiency testing: Comparative validities of interview, imitation, and completion methods. *Language Learning, 33* (3), 315-332.

Heringer, J. (1970). Research on quantifier-negative idiolects. In *Papers from the sixth regional meeting of the Chicago Linguistic Society*, (pp. 287–295). Chicago: CLS.

Heuer, H, (1981). Binary choice reaction time as a criterion of motor equivalence: Further evidence. *Acta Psychologica 50*, 49–60.

Higginbotham, J. & May, R. (1980). Questions, quantifiers, and crossing. *The Linguistic Review 1* (1). 41–80.

Hill, A. (1961). Grammaticality. *Word 17*, 1–10.

Hilles, Sharon. (1986). Interaction and the pro-drop parameter. *Second Language Research* 2. 33–52.

Hudson, R. A. (1984). *Word Grammar*. Oxford: Basil Blackwell.

Horrocks, G. & Stavrou, M. (1985). *Bounding theory and Greek syntax:Evidence for WH-movement in NP*. Ms., St. John's College, Cambridge, MA.

Ioup, G. & Kruse, M. (1977). Interference vs. structural complexity as a predictor of second language relative clause acquisition. In H. D. Brown, C. A. Yurio, & R. H. Crymes (Eds.), *On TESOL 1977* (pp. 150–171). Washington, D. C.: TESOL.

Issidorides, D. (1988). The discovery of a miniature linguistic system: Function words and comprehension of an unfamiliar language. *Journal of Psycholinguistic Research 17* (4),317–339.

Itkonen, E. (1981). The concept of linguistic introspection. In F. Coulmas (Ed.), *A Festschrift for native speakers* (pp. 127–140). The Hague: Mouton.

Iziwa, C. (Ed.). (1989). *Current issues in cognitive processes.* Hillsdale, NJ: Lawrence Erlbaum.

Jordens. P. (1977). Production rules in interlanguage: Evidence from case errors in L2 German. In Elica Kellerman & M. Sharwood-Smith (Eds.), *Crosslinguistic influence in second language acquisition* (pp. 91–109). Oxford: Pergamon.

Karmiloff-Smith, A. (1979). Micro- and macrodevelopmental changes in language acquisition and other representational systems. *Cognitive Science, 3,* 91–118.

Keller- Cohen, D. (1981). Elicited imitation in lexical development: Evidence from a study in temporal reference. *Journal of Psycholinguistic Research, 10* (3), 273–288.

Kellerman, E. (1986). An eye for an eye: Crosslinguistic constraints on the development of the L2 lexicon. In E. Kellerman & M. Sharwood-Smith (Eds.), *Crosslinguistic influence in second language acquisition* (pp. 61–73). Oxford: Pergamon.

Kellerman, E. (1983). Now you see it, now you don't. In S. Gass & L. Selinker (Eds.), *Language transfer and language learning* (pp. 112–134). Rowley, MA: Newbury House.

Kim, J. R., & Masterson, D. (1992). *SpeedMatch* [Computer program]. University of Hawai'i, Honolulu, HI.

Kornfilt, J. (1991). A case for emerging functional categories. *Syntax and Semantics, 25,* 11–36.

Kornfilt, Jaklin (1984). Case marking, empty categories, and agreement in Turkish. Unpublished doctoral dissertation, Harvard, Cambridge, MA.

Kramer, A. F. & Stayer, D. L. (1989). Assessing the development of automatic processing: Application of dual task and event brain potential methodology. *Applied Psycholinguistics, 12*, 299–311.

Kui, L. (1988). *Transfer or parameter setting in SLA*. Unpublished manuscript, University of Hawai'i, Honolulu, HI.

Kuno, S. (1976). Subject, theme, and the speaker's empathy — Reexamination of relativization phenomenon. In C. N. Li (Ed.), *Subject and topic* (pp. 417–444). New York: Academic Press.

Kuroda, S. (1988). Whether we agree or not: a comparative syntax of English and Japanese. In W Poser (Ed.), Papers from the second international workshop on Japanese syntax. Stanford: CSLI.

Lahey, M., Launer, P., and Schiff-Myers, N. (1983). Elicited imitation and spontaneous speech productions of language-disordered children. *Applied Psycholinguistics 4* (4), 317–344.

Lantolf, J., Goss, N. & Zhang, Y.-H. (1991, October). *Two heads may be better than one: Assessing mental activity in L2 grammaticality judgments*. Paper presented at the conference on theory construction and methodology in second language acquisition research, Michigan State University, East Lansing, MI.

Levelt, W. J. M. (1989). *Speaking: From intention to articulation*.Cambridge, MA: MIT Press.

Levelt, W. J. M. (1972). Some psychological aspects of linguistic data. *Linguistische Berichte 17*, 18–30.

Lewandowsky, S. & Murdock, B. B., Jr. (1989). Memory for serial order. *Psychological Review, 96* (1), 25–57.

Li, C . N. & Thompson, S. A. (1976). Subject and topic: A new typology of languages. In C. N. Li (Ed.), *Subject and topic*. New York: Academic Press.

Liberman, I. Y., Shankweiler, D., Liberman, A. M., Fowler, C., & Fisher, F. W. (1977). Phonetic segmentation and recoding in the beginning reader. In A. S. Reber & A. L. Scarborough (Eds.), Toward a psychology of reading. New York: Wiley.

Liceras. J. M. (1989). On some properties of the "pro-drop" parameter: Looking for missing subjects in non-native Spanish. In S. M. Gass & J. Schachter (Eds.), *Linguistic perspectives on second language acquisition* (pp. 109–133). Cambridge, MA: Cambridge University Press.

Lodge, M. (1981). *Magnitude scaling: Quantitative measurement of options.* (Sage University Paper Series on Quantitative Applications in the Social Sciences, Series No. 07-025). Beverly Hills: Sage Publications.

Luce, R. (1986). Response times: Their role in inferring elementary mental organization. *Oxford Psychology Series No. 8.* New York: Oxford University Press.

Lust, B., Chien, Y. C., & Flynn, S. (1987) What children know: Methods for the study of first language acquisition. In B. Lust (Ed.), *Studies in the acquisition of anaphora, Vol. II* (pp. 271–356). Dordrecht: Reidel.

Makino, S. (1968). *Some Aspects of Japanese Nominalizations.* Tokyo: Tokai University Press.

Manicas, P. (1982). The human sciences: A radical separation of psychology and the social sciences. In Secord, P.F. (Ed.), *Explaining human behaviour.* Beverly Hills, CA: Sage Press.

Marquer, J. & Periera, M. (1990). Reaction times in the study of sentence-picture verification: A reconsideration. *Quarterly Journal of Experimental Psychology — Human Experimental Psychology, 42* (1A), 147–168.

Marshall, J. & Morton, J. (1978). On the mechanics of EMMA. In A. Sinclair, R. J. Jarvella, & W. J. M. Levelt (Eds.), *The child's conception of language* (pp. 225–239). Berlin: Springer-Verlag.

Masterson, D. (1989). Functional Categories and parameter setting in L2 acquisition. *Korean Linguistics 6,* 177–192.

Masterson, D. (1990). *An examination of grammaticality judgements and alternative methodologies.* Unpublished manuscript, University of Hawai'i, Honolulu, HI.

Masterson, D. (1992). Instruments for measuring grammaticality: The status of modified NPs in English and Korean. *University of Hawai'i Working Papers in English as a Second Language , 11* (1), 125–156.

Mazurkewich, I. (1988). The acquisition of infinitive and gerund complements by second language learners. In S. Flynn & W. O'Neil (Eds.), *Linguistic theory in second language acquisition* (pp. 127–143). Dordrecht: Kluwer Academic.

McDade, H. L., Simpson, M. S., & Lamb, D. E. (1982).The use of elicited imitation as a measure of expressive grammar. *Journal of Speech and Hearing Disorders, 47,* 19–24.

McLaughlin, B (1980). On the use of miniature artificial languages in second language research. *Applied Psycholinguistics 1*, 357–369.

McLay, H. & Sleator, M. (1960). Responses to language: Judgements of grammaticalness. *International Journal of American Linguistics 26*, 275–282.

McNeill, D. (1966). Developmental psycholinguistics. In F. Smith

Mecham, M. J. (1974). *Motivation and learning-centered training programs for language-delayed children.* Salt Lake City, UT: Word Making Productions.

Meier, R. & Bower, G. (1986). Semantic reference and phrasal grouping in the acquisition of a miniature phrase structure language. *Journal of Memory and Language 25*, 492–505.

Merzenich, A. (1989). Relativity in research and the acquisition of reflexive binding. Unpublished manuscript, University of Hawai'i, Honolulu, HI.

Miller, J. (1988). A warning about median reaction time. *Journal of Experimental Psychology: Human Perception and Performance, 14* (3), 539–543.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review 63*, 81–97.

Mistler-Lachman, J. L. (1972). Levels of comprehension in processing of normal and ambiguous sentences. *Journal of Verbal Learning and Verbal Behavior 11*, (5), 614–623.

Moeser, Shannon D. (1975). Memory for language organization in concrete and abstract sentences. *Memory and Cognition 3* (5), 506–568.

Moeser, Shannon D. (1977). Semantics and miniature artificial languages. In J. MacNamara (Ed.), *Language learning and thought* (pp. 227–250). New York: Academic Press.

Mohan, B. (1977). Acceptability testing and fuzzy grammar. In S. Greenbaum (Ed.), *Acceptability in language.* The Hague: Mouton.

Montague, R. H. (1974). *Formal philosophy: Selected papers of Richard Montague.* New Haven, CT: Yale University Press.

Moore, G. A. B. (1972). The use of closed-circuit television for oral tests in the language laboratory. *NALLD Journal, 47* (1), 27–30.

Morias, J., Cary, L., Algeria, J., & Bertelson, P. (1979). Does the awareness of speech as a sequence of phones arise spontaneously? *Cognition, 7*, 323–331.

Munnich, E., Flynn, S., and Martojardjono, G., (1990, October). *Elicited imitation and grammaticality judgments tasks: What they measure and how they relate to each other.* Paper presented at the conference on theory construction and methodology in second language acquisition research, Michigan State University, East Lansing, MI.

Murray, W. S. (1982). Sentences matching: The influence of meaning and structure. Unpublished doctoral dissertation, Monash University, Clayton, Victoria, Australia.

Nagata, H. (1985). Extraction of linguistically relevant referential properties through language learning. *Journal of Psycholinguistic Research 14* (5), 447–464.

Nagata, H. (1988). The relativity of linguistic intuition: The effect of repetition on grammaticality judgements. *Journal of Psycholinguistic Research, 16* (1), 1–15.

Nayak, N., Hansen, N., Krueger, N. & McLaughlin, B. (1990). Language-learning strategies in monolingual and multilingual adults. *Language Learning 40* (2), 221–244.

O'Connell, B., and Gerard, A. (1985). Scripts or scraps: The development of sequential understanding. *Child Development 56*(3), 671–681.

Odlin, T. (1986). Another look at passage correction tests. *TESOL Quarterly 20,* 123–130.

Oh, J. (1992). The effects of L2 reading assessment methods on anxiety level. *TESOL Quarterly, 26* (1), 172–176.

Otsu, Y. (1981). Universal grammar and syntactic development in children: Toward a theory of syntactic development. Unpublished doctoral dissertation, MIT, Cambridge.

Parker, K. & Chaudron, C. (1990). Discourse markedness and structural markedness. *SSLA 12*, 43–64.

Perkins, K., Brutten, S. R., & Angelis, P. J. (1986). Derivational complexity and item difficulty in a sentence repetition task. *Language Learning, 36* (2), 125–141.

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

Pinker, S. (1989). *Learnability and cognition*. Cambridge: MIT Press.

Poulisse, N., Bongaerts, T. & Kellerman, E. (1987). The use of retrospective verbal reports in the analysis of compensatory strategies. In C.Faerch and G. Kasper (Eds.), *Introspection in second language research* (pp. 213–229). Clevedon: Multilingual Matters.

Pritchett, B. L. (1992). *Grammatical competence and parsing performance.* Chicago: University of Chicago Press.

Pritchett, B. L. (1988). The grammatical basis of language learning. *Language Learning, 64* (3), 539–576.

Quirk, R. & Svartvik, J. (1966). *Investigating linguistic acceptability.* The Hague: Mouton.

Radford, A. (1990). *Syntactic theory and the acquisition of English syntax.* Oxford: Basil Blackwell.

Read, C. & Schreiber, P. (1982). Why short subjects are harder to find than long ones. In (E. Wanner & L. R. Gleitmen, (Eds.), *Language acquisition: State of the art* (pp. 78–101).

Read, C., Yun-Fei, Z., Hong-Yin, N., & Bao-Qing, D. (1986). The ability to manipulate speech sounds depends on knowing alphabetic writing. *Cognition, 24* (special issue), 313–44.

Renandya, W. (1989). *An investigation of the pro-drop parameter in foreign language learning.* Unpublished manuscript, University of Hawai'i, Honolulu, HI.

Reuland, E. (1985). *A feature system for the set of categorial heads.* Unpublished manuscript, Rijksuniversiteit Gröningen.

Reyle, U. & Rohrer, C. (Eds.). (1988). *Natural language parsing and linguistic theories.* Dordrecht: Reidel.

Ritchie, W. (1978). The right-roof constraint in adult language acquisition. In W. Ritchie (Ed.), *Second language research: Issues and implications* (pp. 33–63). New York: Academic Press.

Robinson, Peter (1992). *Reaction-time data and second language acquisition research.* Unpublished manuscript, University of Hawai'i, Honolulu, HI.

Rosenthal, R. (1966). *Experimenter effects in behavioral research.* New York: Appleton-Century-Crofts.

Ross, J. (1979). Where's English? In individual differences in language ability and language behavior. In W. C. Ritchie (Ed.), *Second language acquisition research : Issues and implications* (pp. 33–63). New York: Academic Press.

Schachter, J. (1989a). Testing a proposed universal. In S. Flynn & W. O'Neil (Eds.), *Linguistic theory in secound language acquisition* (pp. 73–88). Dordrecht: Kluwer Academic.

Schachter, J. (1989b). On the issue of completeness in second language acquisition. Unpublished manuscript, University of Southern California.

Schmidtke, H. (1976). Disturbance of processing of information. In E. Simonson & P. C. Weiser (Eds.), *Psychological aspects and physiological correlates of work and fatigue*. Springfield, IL: C. C. Thomas.

Scholes, R. &Willis, B. (1987a). Age and education in oral language skills. *Developmental Neuropsychology 3* (3&4), 239–248.

Scholes, R. &Willis, B. (1987b). Language and literacy. *Journal of Literary Semantics 16*, 3–11.

Scholes, R. &Willis, B. (1987c). The illiterate native speaker of English: Oral language and intentionality. In J. Klesius and Radencich (Eds.), *Links to literacy* (pp. 33–42). Miami: Florida Reading Association.

Segalowitz, S., & Graves, R. (1990). Suitability of the IBM XT, AT, and PS/2 keyboard, mouse, and game port as response devices in reaction time paradigms. *Behavior Research Methods, Instruments, and Computers*, 22 (3), 823–289.

Sharwood Smith (1988). Modularity in second language research. *Polyglot 9*, (Fiche 2, D1).

Shimura, A, & Yoshino, J. (1988). *Himself or zibun: That is the question*. Unpublished manuscript, University of Hawai'i, Honolulu, HI.

Shimura, A. (1990). *Testing the availability of Universal grammar in second language acquisition: The governing category parameter and proper antecedent parameter in Japanese as a second language* (Occasional Paper #17). Honolulu: Department of English as a Second Language, University of Hawai'i.

Silverman,W. P. & Goldberg, S. L. (1975). Further confirmation of same vs. different processing differences. *Perception & Psychophysics, 17*,189–193.

Singh, R., d'Anglejan, A. & Carroll, S. (1982). Elicitation of inter-English. *Language Learning 32*, 271–288.

Singleton, W. T. (1953). Deterioration of performance on short term perceptual-motor task. In W. F. Floyd & A. T. Welford (Eds.), *Symposium on fatigue*. (pp. 163–172). H. K. Lewis & Co. for Ergonomics Research Society, London.

Slobin, D. I. & Welsh, C. A. (1973). Elicited imitation as a research tool in developmental psycholinguistics. In C. Ferguson & D. Slobin (Eds.), *Studies of child language development* (pp. 485–497). New York: Holt, Reinhart, & Winston.

Snodgrass, J. G., Levy-Berger, G. & Hayden, H. (1985). *Human experimental psychology*. New York: Oxford University Press.

Snow, C, & Meijer, G. (1977). On the nature of syntactic intuitions. In S. Greenbaum (Ed.), *Acceptability in language.* The Hague: Mouton.

Sorace A. & Bard, E. G. (1991, October). *Magnitude estimation techniques for the elicitation of non-native acceptability judgments.* Paper presented at the conference on theory construction and methodology in second language acquisition research, Michigan State University, East Lansing, MI.

Spencer, N. (1972). Differences between linguists and non-linguists in intuitions of grammaticality-acceptability. *Journal of Psycholinguistic Research 2,* 83–98.

Suciadi, E. (1989). *Universal grammar in second language learning: The case of Indonesians learning English as a second language.* Unpublished manuscript, University of Hawai'i, Honolulu, HI.

Szabolcsi, A. (1981). The possessive construction in Hungarian: a configurational category in a non-configurational language. *Acta Linguistica Academiae Scientiarum Hungaricae 31,* 261–289.

Szabolcsi, A. (1984). The possessor that ran away from home. In A. Kenesei (Ed.), Approaches to Hungarian, vol. 2, Szeged.

Takahashi, S. & Roitblat, H. L. (1992). Comprehension process of L2 indirect requests. Unpublished manuscript, University of Hawai'i, Honolulu, HI.

Thom, K. (1988). *Testing a theory of Universal Grammar in adult second language learning.* Unpublished manuscript, University of Hawai'i, Honolulu, HI.

Tucker, G. R., & Sarofim, M. (1978). Investigation of linguistic acceptability with Egyptian ESL students. *TESOL Quarterly, 13,* 29–39.

Vetter, H., Volovecky, J., and Howell, R. (1979). Judgments of grammaticalness: A partial replication and extension. *Journal of Psycholinguistic Research 8,* 576–583.

Weber-Olsen, M., Putnam-Sims, P., & Gannon, J. (1983). Elicited imitation and the Oral Language Sentence Imitation Screening Test (OLSIST): Content or context?. *Journal of Speech and Hearing Disorders 48* (4), 368–378.

Welford, A. T. (1980). Relationships between reaction time and fatigue, stress, age, and sex. In A. T. Welford (Ed.), *Reaction times*, (pp. 321–354). London: Academic Press.

Welford, A. T. (1977). Motor performance. In J. E. Birren & K. W. Shaie (Eds.), *Handbook of the psychology of aging*. (pp. 450–496). New York: Van Nostrand Reinhold.

Welford, A. T. (1971). What is the basis of choice reaction time. *Ergonomics 14*, 679–693.

White, L. (1989a). The adjacency principle of second language acquisition: Do L2 learners observe the subset principle? McGill University, MS.

White, L. (1989b, February). *Processing constraints: Are they sufficient for explaining adult SLA?* Paper presented at Second Language Research Forum, University of California at Los Angeles.

White, L. (1988). Island effects in second language acquisition. In S. Flynn & W. O'Neil (Eds.), *Linguistic theory in second language acquisition* (pp. 277–294). Dordrecht: Kluwer Academic.

White, L. (1987 , February). *Universal grammar: is it just a new name for an old problem?* Paper presented at Second Language Research Forum 1987. Los Angeles, California.

White, L. (1985). The pro-drop parameter in adult second language acquisition. *Language Learning, 35* (1), 47–62.

Wolfe-Quintero, K. (1992). *The representation and acquisition of the lexical structure of English dative verbs: Experimental studies of native English speakers and Japanese and Chinese adult learners of English.* Unpublished doctoral dissertation, University of Hawai'i, Honolulu, HI.

Yang, I. S. (1972). *Korean syntax: the nature of phrasal and clausal case.* Unpublished doctoral dissertation, University of Washington, Seattle, WA.

Yoon, J.–R. (1990). Korean syntax and generalized X–bar theory. Seoul, South Korea: Hanshin Publishing.

Yoon, J.–R. (1992). Functional categories in Korean clausal and nominal structures. *Generative Grammar Circle of Linguistic Society of Korea, 2* (2), 427–464.