Linguistic Correlates of Proficiency in Korean as a Second Language

[1]Sun-Young Lee, [2]Jihye Moon and [3]Michael H. Long
([1]Cyber Hankuk University of Foreign Studies,
[2&3]University of Maryland at College Park)

Lee, Sun-Young, Jihye Moon and Michael H. Long. (2009). Linguistic Correlates of Proficiency in Korean as a Second Language. ***Language Research*** **45.2, 319-348.**

This study investigates relationships between global oral proficiency ratings and measures of grammatical competence in the acquisition of Korean as a second language. Data were collected on the linguistic abilities of learners at 1+ to 4 on the ILR scale, focusing on perception in phonology, morphology, syntax, lexis, and collocation. The results show that (i) most of the tasks have high internal reliability, (ii) individual accuracy scores correlate strongly with levels on the ILR proficiency scale on most tasks, and (iii) heritage speakers outperform non-heritage speakers at the same high levels of oral proficiency on most tasks. The findings indicate that global proficiency scales like the OPI can be deconstructed using tasks that provide detailed measures of learners' control of linguistic features.

Keywords: linguistic correlates, Korean as a Second Language (KSL), acquisition of Korean, oral proficiency scales, grammatical competence, heritage speakers

1. Introduction

This study investigates the relationship between global oral proficiency ratings and measures of grammatical competence in the acquisition of Korean as a second language by English-speaking adults. The purpose of the study is to determine whether learners' overall proficiency can be specified in terms of their perception abilities in Korean. Global proficiency measures such as the Oral Proficiency Interview (OPI) and ratings on proficiency scales such as the ILR (Interagency Language Roundtable), ACTFL (American Council on the Teaching of Foreign Languages), CEFR (Common European Framework of Reference for Languages), and TOEFL (Test of English as a Foreign Language) do not provide learners or their teachers with sufficient information about their linguistic abilities. If told that a group of learners are "2s" on the ILR scale, for example, what is it that they already know, and what do they need to learn to reach level 3? It will be very helpful for learners to know their

strengths and weaknesses in terms of specific linguistic features in the language they are learning in the phonological, morphological, syntactic, lexical and collocational domains in Korean. If a diagnostic test could provide such information, it would help learners and their teachers know what to focus on in order to move to the next level of proficiency, e.g., from ILR 2+ to ILR 3. To that end, it is necessary to identify typical linguistic profiles of learners at each level on the scale of interest. However, no study has been conducted to identify the typical profiles of learners at each proficiency level in Korean as a second language or any other languages.

To provide such information, it is important to determine which linguistic features correspond to each level of the proficiency scale for Korean. For example, it is often noticed that Korean beginners have more difficulty using the honorific marker *-si-* than the subject case marker *-ka.* When do learners attain native-like competence with *-si-*, ILR 3+, or ILR 4? Such linguistic correlates need to be identified to create the typical profile of learners of Korean at each level of the ILR scale.

In this paper, we present the preliminary results on linguistic correlates of proficiency in perception of Korean by adult English-speaking learners. A total of 21 tasks focusing on a wide range of Korean linguistic features were designed, using eight types of perception tasks. Our discussion focuses on the validity of feature-specific tasks as diagnostic measures of the linguistic knowledge underlying L2 oral proficiency.

The organization of the paper is as follows. First, the underlying concept and the goals of the Linguistic Correlates of Proficiency project will be introduced. Second, the data-collection battery and the data-collection procedures are described, followed by the results presented by task-type. Finally, the results are summarized, focusing on the relationship between oral proficiency and grammatical competence in L2 learners' perception of Korean. Differences between heritage and non-heritage speakers are observed on many tasks. The implications of the study are proposed for the development of empirically-based syllabi, pedagogic materials, and diagnostic tests for foreign language education.

1.1. The Linguistic Correlates of Proficiency Project

The first major aim of the University of Maryland's Linguistic Correlates of Proficiency (LCP) project is to provide useful linguistic information for all those engaged in the learning, teaching or testing of advanced proficiency in less commonly taught languages, including Korean. The purpose is to deliver detailed linguistic profiles of the listening and speaking abilities of *typical* adult English-speaking students at levels 2, 3, and 4 on the ILR scale in those languages, as well as instrumentation usable to establish the linguistic profiles of

individual learners. The project covers perception and production in phonology, morphology, syntax, lexis, and collocation.

As indicated above, in the context of this project, Linguistic Correlates of Proficiency (LCPs), or linguistic profiles, are detailed inventories of the linguistic features (sounds, grammar, vocabulary, collocations, registers, etc.) and other abilities (dialect recognition, ability to comprehend speech in noise, etc.) typically mastered by, or still posing problems for, adult English-speaking learners of critical foreign languages at advanced proficiency levels, 2, 3 and 4, on the ILR scale. They spell out the linguistic items and abilities students need to master if their work requires them to move from one level on the ILR scale to the next. The information would be easy enough to obtain if diagnostic language tests (or for that matter, validated global proficiency measures) already existed in the 2 to 4+ range, but in most less commonly taught languages (LCTLs), almost all rarely taught languages (RTLs), and not a few more commonly taught ones (French, German, Spanish, etc.), that is simply not the case.

In the field of language testing, most language proficiency tests produce a global rating, often in the form of a label or number -- 'ACTFL intermediate low,' 'ILR 1+,' 'CEFR A2,' 'TOEFL 580,' etc. These ratings convey relevant information to those who developed or administered the test, or those who are at least familiar with the test and understand how such ratings are arrived at. To outsiders, such as professionals unfamiliar with the particular tests, many employers and most students, the results are less transparent. It would be very valuable for curriculum writers, teachers, and the learners themselves to know precisely what it is that students assigned such ratings can do, and what it is they still need to master in order to reach the next level, e.g., to move from ACTFL Intermediate Low to ACTFL Intermediate High, from ILR 2 to ILR 3, CEFR A2 to CEFR B1, TOEFL 580 to TOEFL 620, etc.

It is possible that transcripts of oral proficiency interviews (OPI) used to assess global proficiency can be analyzed for specific linguistic features, as well as for quantitative assessments, such as dependent clause ratio, active/passive ratio, error-free clause ratio, error frequency rate per clause, lexical type/token ratio, lexical sophistication ratio, and frequency of borrowings. In the field of SLA research, OPI interview recordings have been used to analyze interlanguage (IL) features in phonology, tense and aspect, and inflectional morphology. Similarly, data elicited via guided narratives, e.g., picture-cued descriptions of the September 11 terrorist attacks on New York employed by Kanno et al. (2008) and Y-G Lee et al. (2005), and from L1A work, the Frog Story (e.g., Polinsky 2008), lend themselves to those and other quantitative analyses, especially of verbal forms, including tense, aspect, and verbs of motion. Due to options in topic control and avoidance, however, such relatively open-ended procedures tend to elicit samples of what speakers can do, or in reality, a sub-

set of what they can do, i.e., what they happened to do on that occasion, and are less useful for finding out what they cannot do. This means that a satisfactory data-collection battery or a diagnostic test suitable for use with very advanced learners both need to include a number of closed tasks designed to probe the nooks and crannies of language proficiency, i.e., what learners still cannot do, with no wriggle-room allowed. Discrete-point tasks of several kinds are needed to identify gaps in learners' linguistic repertoires, and to measure accurate and appropriate usage in various phonological, morphological, syntactic, discourse, lexical, and collocational domains.

Diagnostic language testing in general is in its infancy (see Alderson 2005, Alderson & Huhta 2005, Kunnan & Jang 2009). The pioneering project in the field, DIALANG, involved development of computer-based, internet-delivered, open access tests of listening, reading, writing, vocabulary and grammar in 14 European languages linked to the six-level Common European Framework of Reference (CEFR) proficiency scale (Council of Europe 2001). Both the proficiency levels in the CEFR scale and assignment of items in the diagnostic tests to levels are based on expert judgments of difficulty, i.e., the intuitions of experienced teachers and testers of the languages concerned, not on data on the acquisition of those languages. DIALANG test items are also framed in terms of what might be thought of as proficiency sub-skills, e.g., for reading: 'ability to distinguish main idea from supporting detail,' 'ability to understand text literally,' 'ability to make appropriate inferences.' Alderson concludes his report on the project by stating that neither of those features of the tests are desirable or work in practice. In particular, proficiency-type items (e.g., reading sub-skills) do not pattern with CEFR level and do not get at what underlies proficiency:

> "The results of the piloting of the different DIALANG components showed fairly convincingly that the subskills tested do not vary by CEFR level. It is not the case that learners cannot inference at lower levels of proficiency, or that the ability to organize text is associated with high levels of proficiency only. This appeared to be true for all the subskills tested in DIALANG. If subskills are not distinguished by CEFR level, then one must ask whether the development of subskills is relevant to an understanding of how language proficiency develops and hence to diagnosis. . . . Such (factor) analyses failed to justify any belief that the various subskills contribute differentially to the macro skill. Either DIALANG has failed to identify relevant subskills or diagnosis in foreign language ability should proceed differently. The possibility is worth considering that what needs diagnosing is not language use and the underlying subskills, but linguistic features, of the sort measured by the Grammar and Vocabulary sections of DIALANG." (Alderson 2005: 261)

Empirically based acquisition data, in the form of inventories of objectively recognizable linguistic features, is the desirable approach. Since relationships between mastery of such features and proficiency levels are unknown, a priori assignment of features (or proficiency sub-skills) to levels is unjustified. Rather, charting mastery of sets of features relative to one another is what is needed, along with the relationships of feature mastery to proficiency scale levels, something to be determined by the data, not by intuition or a priori "standard" setting, even when, as in the DIALANG project, acceptably high inter-rater reliability was obtained among the judges of item difficulty. Standard setting on the basis of the judgment of item raters, however expert, means that one set of intuitions, the original proficiency scales, is being equated with another set of intuitions, when what is needed is an empirical basis to both.

A secondary goal of the research is to investigate commonalities and the degree to which linguistic profiles at each proficiency level differ for heritage and non-heritage students, and by implication, the justification for separate tracks for heritage and non-heritage students beyond the early stages. Research on the typical linguistic profiles in the ILR 2 - 3 range of English-speaking learners of Japanese (Kanno et al. 2008) and Korean (Y-G Lee et al. 2005) identified commonalities across those two typologically closely related languages. The typicality of many of the problems and the relative saliency of the categories identified in that work is promising, and suggests that similar success may be had from research at higher proficiency levels, too. Importantly, Kanno et al. (2008) and Y-G Lee et al. (2005) found considerable differences in the linguistic profiles of heritage and non-heritage speakers, both within and between groups. Identifiable sub-groups of heritage speakers, for example, performed very differently in Japanese and Korean, depending on the kinds of language exposure they had experienced *subsequent to* (L1 or bilingual) exposure during the early home years. Echoing findings in ESL (Pica 1983), variation in prior language-learning experience was related to considerable differences in the profiles of non-heritage speakers, too, e.g., between those whose Japanese or Korean was mostly the product of classroom instruction, and those who had learned mostly through naturalistic exposure in-country (submersion). It is quite likely that such differences persist in the 3 to 4+ range, but whether or not they really do is addressed in this project via a Language Experience Questionnaire. Just as different medical diagnoses require different treatment programs, so should different linguistic profiles merit at least somewhat different instructional programs. At stake are efficiency (financial costs and time for employers and students) and success rates (the ultimate level of proficiency attained).

In this paper, we present preliminary findings on linguistic correlates of proficiency in Korean as a second language, based on data from perception tasks in the Korean data-collection battery.

2. Method

2.1. Participants

A total of 66 adult speakers of Korean participated in this first round of data collection: 32 heritage speakers, 20 English-speaking L2 learners, and 14 native speakers of Korean. There were 29 males and 37 females in the sample, ranging in age from 19 to 67, with a mean age of 28. The number of participants in each group is shown in Table 1, along with their mean age, age range, gender, and tested ILR proficiency level.

Table 1. Participants

Subject Category	*Number of Participants*	*Mean Age (age range)*	*Gender*		*ILR Levels*					
			M	F	1+	2	2+	3	3+	4
Native Speakers	14	34 (20-58)	5	9						
Heritage Speakers	32	23 (19-49)	9	23	2	6	8	14	1	1
L2 Learners	20	30 (20-67)	15	5	4	4	6	3	2	1
Total	66	28 (19-67)	29	37	6	10	14	17	3	2

32 participants were recruited from within the greater Washington, D.C., area, including Maryland and Virginia, and 34 participants from Seoul, South Korea. Most participants, including native controls, were college students, college graduates or graduate students. In particular, the majority of the advanced heritage speakers with an ILR 3 proficiency rating were graduate students in the federally funded Korean Flagship Program, starting the overseas portion of their program in Korea at the time of data collection. Most of the heritage speakers had more than one year of Korean language education at the university level and spoke Korean at home. The L2 learners at the ILR 3+ and 4 levels had lived in Korea for at least seven years (one of them for 35 years) working as university English teachers. The L2 learner with an ILR 4 rating was a graduate student and English teacher at a university in the US, had lived in Korea for about 17 years, and was married to a Korean. The heritage speaker with an ILR 4 rating worked as a translator for a broadcasting company in Korea, taught at a university there, and had spent her childhood in both England and Korea, moving back and forth between the two countries. She had also traveled to many different countries around world, where she was exposed to, and learned, different languages, including French, Arabic, and German.

Proficiency levels of the participants in this study is measured with Oral Proficiency Interview (OPI) with a certified OPI tester and labeled using the Interagency Language Roundtable (ILR) speaking scale used in the U.S. gov-

ernment. The scale ranges from 0-5 and each proficiency level is named as follows to roughly describe the proficiency at each level:

Speaking 0	No Proficiency
Speaking 0+	Memorized Proficiency
Speaking 1	Elementary Proficiency
Speaking 1+	Elementary Proficiency, Plus
Speaking 2	Limited Working Proficiency
Speaking 2+	Limited Working Proficiency, Plus
Speaking 3	General Professional Proficiency
Speaking 3+	General Professional Proficiency, Plus
Speaking 4	Advanced Professional Proficiency
Speaking 4+	Advanced Professional Proficiency, Plus
Speaking 5	Functionally Native Proficiency

The description of each level is found in the website, http://www.govtilr.org/Skills/ILRscale2.htm. For example, the lowest level ILR 1+ and the highest level ILR 4 found in our study are described as follows:

Speaking 1+ (Elementary Proficiency, Plus) *Can initiate and maintain predictable face-to-face conversations and satisfy limited social demands. He/she may, however, have little understanding of the social conventions of conversation. The interlocutor is generally required to strain and employ real-world knowledge to understand even some simple speech. The speaker at this level may hesitate and may have to change subjects due to lack of language resources. Range and control of the language are limited. Speech largely consists of a series of short, discrete utterances.*

Speaking 4 (Advanced Professional Proficiency) *Able to use the language fluently and accurately on all levels normally pertinent to professional needs. The individual's language usage and ability to function are fully successful. Organizes discourse well, using appropriate rhetorical speech devices, native cultural references and understanding. Language ability only rarely hinders him/her in performing any task requiring language; yet, the individual would seldom be perceived as a native. Speaks effortlessly and smoothly and is able to use the language with a high degree of effectiveness, reliability and precision for all representational purposes within the range of personal and professional experience and scope of responsibilities. Can serve as in informal interpreter in a range of unpredictable circumstances. Can perform extensive, sophisticated language tasks, encompassing most matters of interest to well-educated native speakers, including tasks which do not bear directly on a professional specialty.*
(http://www.govtilr.org/Skills/ILRscale2.htm)

As can be seen in Table 1, a majority of our participants belong to the levels

2-3 with very few that belong to the other levels on the ILR scale. However, there are still one heritage speaker and one L2 learner at ILR 4, which enables us to see any tendency in terms of differences in their linguistic abilities between the heritage learners and L2 learners in such a high level as native-like proficiency.

2.2. Materials

A total of eight types of tasks were used to test learners' perception of Korean in 21 Korean linguistic features in the areas of phonology, morphology, syntax, lexis, collocation, and accent detection. Some task-types (e.g., grammaticality judgment) have been widely used in mainstream language testing and/or SLA research for many years, especially in studies testing the Critical Period Hypothesis. Others (e.g., lexical decision, phoneme monitoring, and picture matching) originated in experimental psychology and L1 psycholinguistics. The same task-type (e.g., grammaticality judgment and lexical decision) was often used to test more than one linguistic feature.

A variety of ideas from various sources were used to choose the linguistic features to include in the tasks. They included the reflections of experienced teachers of Korean, introspections (obtained through interviews) of high-achieving learners of Korean concerning residual linguistic difficulties for them, research findings on late-acquired items in first language acquisition and on late-acquired or rarely acquired features in adult L2 acquisition, and the literature on critical periods in SLA (although, few of the last concerned the acquisition of Korean). Details of the task-types employed and the features tested using them are shown below. Written forms of the target language were avoided, given that some of the eventual L2 learners who will use the designed test will not necessarily have full command of the Korean writing system, *Hangul,* even though they can speak and understand Korean quite well.

Grammaticality Judgment

In this task, participants listen to sentences and determine if each is grammatical. There were a total of four warm-up items and 264 test items (132 grammatical and ungrammatical pairs). Linguistic features tested via this task are described below, with an example and the number of test items in parentheses.

- Tense dependency (40): The past tense marker *–a/ess-* is deleted in the subordinate clause of a past-tense sentence, depending on the conjunction (e.g., pika wa-se/*w**a-ss**-se, wusan-ul kace w**a-ss**-ta. 'I brought an umbrella because it was raining.'
- Past-Tense in Relative Clauses (20): The past tense marker, *–a/ess-* is deleted in the relative clauses with the past-tense adnominal marker *-(u)n*

(e.g., [nay-ka ecey **po-n/*po-ass-nun**] salam, 'the man that [I saw yesterday]').

- Particle stacking (10): Multiple case markers and particles can be combined in a relatively fixed order (e.g., i mwuncey-nun younghee-**eykey-ka/*-ka-eykey** himtul-ta, 'This problem is difficult for Younghee.').
- Locative Verbs (36): Each locative verb takes a different type of direct object. (e.g., can-ey **mwul-ul/*can-ul** mwul-lo pwus-ta, 'Pour the water in the glass/* the glass with water.').
- Negation (20): Different types of negative words are used, depending on the type of sentence. For example, *-ci mal-* is used in imperative sentences (e.g., hakkyo-ey ka-ci **mal /*anh**-a-la, 'Don't go to school.').
- Conjunction (20): Different conjuncts are used, depending on the type of sentence. For example, *-ese* cannot be used in imperatives (e.g., pi-ka o-**nikka/*-ase**, wusan-ul kace ka-la, 'Take an umbrella with you since it is raining.').
- Numerals (30): Korean uses Sino Korean and Native Korean numerals, depending on the bound noun (classifier or numbering word). (e.g., cey tongsayng-un **sey/*sam** sal i-ta, 'My younger brother is three years old.').
- *Wear*-verbs (20): Different *wear*-type verbs are used, depending on the type of object or clothes (e.g., yangmal-ul **sin/*ip** –ela, 'Put the socks on').
- Apperceptive (20): Korean uses a different verb-ending apperceptive marker in the present tense, depending on the category of the verb: adjectival verb or verb (e.g., pi-ka o-**nun-kwuna/*kuna**!, 'It is raining.'/ younghee-ka yeppu-**kwuna/*-nun-kwuna!** 'Younghee is cute!').

Acceptability Judgment

The acceptability judgment (AJ) task is very similar to the grammaticality judgment task described above (Sorace 1996). In this task, participants hear a sentence and judge whether it is appropriate in Korean. Honorifics and light verb idioms were tested using this task.

- Honorifics (20): Honorific marker *–si-*and honorific case marker, *-kkeyse* are used, depending on the age and social status of the subject in a sentence related to those of the speaker and the hearer (e.g., halapeci-**kkeyse/*-ka** o-**si**/*o-n ta, 'Granpa is coming.').
- Light Verbs Idioms (40): Idiomatic expressions use typical types of verbs or light verbs (e.g., cenyek **ha**-ta, **do dinner/eat dinner*, meli-**ha**-ta, 'do hair/*make hair').

Lexical Decision

A lexical decision task is analogous to a word-level version of GJ. Participants hear a word in Korean and are asked to judge whether or not it is a real word. A total of eight warm-up items and 311 test items were prepared. The

linguistic features tested are described below:

- Syllable Structure (76): Consonant clusters are not allowed in Korean syllables (e.g., lema/*krema). 38 illegal and 38 legal syllable structures were constructed.
- Morphological rule (75): Pul-, *not*, Sino-Korean negative prefix can only be attached to Sino-Korean nominals, not to native Korean nominals (e.g., pul-kanung, 'not-possibility'/*pul-apum, 'not-pain') with exceptions (e.g., *pul-kongkay, 'not-publicization').
- Compound Nouns (60): There are existing compound nouns and possible, but non-existent, compound nouns (e.g., ingkko pwupwu, 'parrot couple'/*ceypi pwupwu, 'sparrow couple').
- Overall vocabulary (100): Various nouns, verbs, adjectives and adverbs were chosen with balanced numbers of items in each category, and with frequency controlled based on the Sejong corpus (e.g., frequent (more than 4,500/1.5M words): cayen 'nature,' cenhwa 'telephone,' wus-ta 'to laugh,' nop-ta 'high,' etc.; infrequent (less than 9/1.5M words): kyopyen 'teaching job,' cintuki 'tick,' kyenwu-ta 'to target,' pili-ta 'fishy,' etc.).

Picture-Word Discrimination

In this task, participants look at a picture and determine if it matches the sound (presented as a word) they are hearing. For example, they are presented with a picture of *tal, moon*, with sound of tal, in a matched condition, or the sound of ttal, *daughter* in a mismatched condition. Three kinds of voiceless alveolar stops were tested in this task: /t/ (ㄷ, plain), /t'/ (ㄸ, tensed), and /t^h/ (ㅌ, aspirated). A total of four practice items and 60 test items with corresponding pictures were prepared.

AXB

Participants listen to three consecutive sounds and determine which the second sound matches, the first sound or the third. Korean plain vs. tensed stop sound pairs, /p/ vs, /p'/, /t/ vs, /t'/, /c/ vs, /c'/, /k/ vs, /k'/, and /s/ vs, /s'/, were tested. Six pairs of practice items and 150 pairs of test items (30 pairs per each distinction) were included.

Phoneme Monitoring

Participants listen to words and determine whether or not each contains /p/. They are asked to find the first sound of *pul* 'fire,' with a picture of fire and the English script, 'fire' on the screen while they hear the sound of the word *pul* in Korean. A total of eight warm-up items and 80 items were prepared. About half the items included the target sound in the word-initial position, while the other half had the target sounds in intervocalic position.

Picture-Sentence Matching

Participants look at a picture, listen to a sentence, and judge if the picture matches the sentence they are hearing. Korean reflexive *caki* (self) was tested in this task in order to find out if L2 learners of Korean know the reference of *caki* (White 2003: 46). In English, a reflexive cannot refer to the subject of the main clause in complex sentences. Unlike English, Korean reflexive *caki* can refer to either the subject in the main clause or the subject in the subordinate clause. For example, in the sentence *younghee$_i$-nun [sunhee$_j$-ka* ***caki*** $_{i/j}$ *-lul coha-han-ta]-ko sayngkak-han-ta* 'Younghee$_i$ thinks [that Sunhee$_j$ likes herself$_{*i/j}$, caki' can refer to either Younghee or Sunhee. A total of four practice items and 20 test items were devised, with 10 matched and 10 mismatched pictures.

Accent Detection

Participants hear a thirty-second speech sample of Korean regional dialects and judge in which area of Korea the dialect is spoken. Four choices are provided as possible answers: Seoul, Kyoungsang, Chela, and North Korea. A total of two warm-up items and 12 test items were prepared, with three speech samples for each of the four dialects.

Programming

The test batteries were prepared in a computerized format using the DMDX software package.[1] All the test items were recorded by a male and a female speaker. The items were randomized within tasks and presented orally, sometimes along with picture stimuli and/or English scripts on the monitor, depending on the type of task. No Korean scripts were provided on the monitor. The participants' oral and button-pressed responses were automatically recorded by the program for later analysis. All the tasks were divided into two separate blocks. Each block took approximately one hour to complete. The two perception blocks were alternated with two production blocks in the whole test battery. Grammaticality judgment and lexical decision tasks were divided into two blocks because of the very large number of items, which, it was thought, might cause subject fatigue. All test items were randomized within each task. Clear instructions were given at the beginning of each task, with several practice items, in order to make sure participants knew how to respond.

2.3. Procedure

Data collection took place within the greater Washington, D.C., area, including Maryland and Virginia, and in Seoul, South Korea. It was usually

[1] A program developed and maintained at the University of Arizona by Jonathan Forster, http://www.u.arizona.edu/~kforster/dmdx/dmdx.htm

conducted in quiet offices on the University of Maryland campus, but also at similar locations at several other universities, and in a few cases, elsewhere, such as private homes. The entire test battery took about four hours to complete. As mentioned above, there were two blocks of production tasks and two blocks of perception tasks, alternating with each other. Each block took about 40 minutes to one hour, depending on the participants. Between blocks, participants were encouraged to take a break to prevent them from getting tired. In addition, within each block, they were able to take a short break after finishing one task and before they started a new task. Within tasks with 20 or more items, participants could take a break in every 10 items before they went on to next set of items. Most participants completed the whole test battery on the same day, with a roughly 10-minute break between the second and third blocks. Only a few participants had to return to complete the test. While wholly computer-delivered, a researcher was always present to solve any technical problems, provide the informed consent forms, administer the language experience questionnaire, and answer procedural questions participants might have.[2] About one week after each data collection, each participant's global proficiency in Korean was assessed by a certified OPI tester on the basis of a telephone interview, and their OPI scores were recorded in terms of a proficiency level on the ILR scale. Participants also completed a questionnaire concerning their language-learning history. Data collection took approximately four hours, and participants were paid $80 by way of compensation.

2.4. Results

First, item analyses were conducted to assess the value of each item individually. If three or more out of 14 native controls missed a particular item, the item was removed from further analysis. Data from all participants were then used to calculate the p-value (the probability of observing a correct score on that item) and the item's discriminatory power. Following standard procedures, item discrimination was calculated as the difference in p-value between the upper 27 percent of the sample and the lower 27% of the sample. The upper and lower 27 percent were determined using examinees' proportion correct scores on the set of items under investigation (usually the items within a single task). Any items with negative discrimination values (indicating that the lower ability examinees were more likely to obtain a correct score on that particular item than the higher ability examinees) or zero discrimination were removed from further analysis. After item analysis was complete, participants serving as NS controls were removed from further analyses, so that the functioning of the

[2] Such an arrangement will not be required with the eventual diagnostic tests, which will be deliverable in person or at a distance with no-one present but the test-taker.

task as a whole could be assessed for learners only.

The remaining examinees' proportion correct scores were also recalculated, so that items with poor discrimination would not be included in the score. These participants were separated into subgroups based on their ILR levels, and several different analyses were conducted to assess the extent to which proportion correct scores on a task aligned with ILR level. First, the mean and standard deviation of the proportion correct scores were calculated for participants at each ILR level. Next, Spearman correlations were calculated to assess the strength of relationships between ILR levels and proportion correct scores for individual participants for each task.[3] Spearman correlation coefficients between ILR levels and mean proportion correct scores were also computed using the mean proportion correct score for examinees at each ILR level. Alpha for all analyses was set at .05. Finally, the internal consistency of items within each task was assessed by computing a Cronbach's alpha statistic. Cronbach's alpha uses item variances and total score variance to estimate the proportion of observed score variability that is "true," or due to differences in examinees' abilities, rather than to random error.

The overall perception task statistics are summarized in Table 2, with number of test items before eliminating those with inadequate discrimination indices, the number of eliminated items, internal reliability, and correlations between test scores (individual scores and mean scores) and ILR levels for each test and for the overall task, as well.

All tasks show high internal reliability, with most Cronbach alpha coefficients close to, or substantially higher than, .70, except for Conjunction (r = .54), Particle Stacking (r = .55), Honorifics (r = .56), Phoneme Monitoring (r = .54) and Accent Detection (r = .45). One of the main reasons for the lower internal reliability of those particular tasks may be that the test items for them were sub-grouped into different conditions, each with smaller numbers of items, as a result, some of which were expected to be more difficult than the others within a task. Some items were also thought more difficult for L2 learners than for heritage speakers. Furthermore, certain test items may be judged differently when presented out of context. For example, learners' acceptability judgments of items in the honorifics test could be affected by their control of individual honorific systems. One such example could relate to the possible difference between the honorific case marker, *-kkeyse* and the honorific verbal suffix, *-si.* There seems to be a tendency in the use of Korean honorifics such

[3] For these calculations, the different ILR categories needed to be ranked. This was accomplished in three different ways: (i) giving each category its own unique ranking (ILR10), potentially resulting in 10 categories (0, 1, 1+, etc.), (ii) grouping "+" scores with the next *lower* category, e.g., collapsing 2+ and 2 (ILR5), and (iii) grouping "+" scores with the next *higher* category, e.g., collapsing 2+ and 3 (ILR5ALT). Since the results are almost the same, we present the data only with 10 different categories (ILR10).

Table 2. Task Statistics for Korean Perception

Task (Subtask)	Number of Items	Cronbach alpha	Spearman Correlation	
			Individual Scores (n=52)	Mean Scores (n= 6)
GJT Task	254 (-57)[4]	0.96	0.52*	0.85*
Conjunctions	20 (-7)	0.54	0.40*	0.97*
Tense Dependency	40 (-11)	0.82	0.52*	0.85*
Past Tense in Relative Clauses	20 (-1)	0.83	0.44*	0.97*
Particle Stacking	10 (-2)	0.55	0.34*	0.85*
Negation	30 (-7)	0.70	0.54*	0.85*
Apperceptives	20 (-5)	0.73	0.50*	0.97*
Locative Verbs	36 (-11)	0.82	0.25	0.74
Numerals	30 (-4)	0.81	0.45*	0.88*
Wear-verbs	20 (-3)	0.74	0.39*	0.85*
Acceptability Judgment Task_ALL	65 (-18)	0.84	0.54*	0.97*
Honorifics	20 (-7)	0.56	0.51*	0.97*
Light Verbs	45 (-11)	0.83	0.47*	0.97*
Lexical Decision Task_ALL	311 (-117)	0.92	0.52*	0.85*
Syllable Structure	76 (-21)	0.87	0.47*	0.68
Morphological Rule	75 (-28)	0.83	0.50*	0.85*
Compound	60 (-17)	0.71	0.42*	0.64
Overall Vocabulary	100 (-51)	0.70	0.13	0.35
Picture-Word Discrimination (PWD)	60 (-11)	0.91	0.33*	0.74
AXB	150 (-21)	0.94	0.33*	0.74
Phoneme Monitoring	80 (-54)	0.51	0.23	0.76
Picture Sentence Matching_Reflexives	20 (-5)	0.72	0.48*	0.97*
Accent Detection	12 (-8)	0.45	0.31*	0.59

*$p < .05$

that *-kkeyse* is less strictly used than *-si.* For instance, that is to say, *kyoswu-nim*-i *ka*-si-*ess-ta* 'the professor is gone' with *-si* but lacking *-kkeyse* (*-i* is used, instead) sounds less impolite than *kyoswu-nim-kkeyse ka-ss-ta* with *-kkeyse* but lacking *-si*. The reason for including those items was to identify any acceptability differences among the test items in different conditions, as well as among heritage speakers and L2 learners.

The following sections report the proportion correct mean scores at each level of the ILR scale divided into sub-tests targeting different linguistic features. The data are presented by task-type.

[4] The numbers include additional 38 filler items with five dropped items.

Table 3. Grammaticality Judgment

ILR	N.	Conjunction		Tense Dependency		Past Tense in RC		Particle Stacking		Negation		Apper-ceptive	
		M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
1+	5	0.55	0.13	0.55	0.11	0.46	0.70	0.40	0.11	0.62	0.18	0.60	0.19
2	10	0.74	0.16	0.79	0.10	0.64	0.25	0.60	0.21	0.74	0.12	0.87	0.15
2+	12	0.73	0.16	0.78	0.17	0.74	0.18	0.67	0.15	0.83	0.11	0.83	0.13
3	17	0.84	0.14	0.85	0.13	0.78	0.21	0.63	0.28	0.91	0.11	0.93	0.08
3+	2	0.60	0.28	0.92	0.11	0.82	0.18	0.61	0.39	0.82	0.02	0.87	0.00
4	2	0.78	0.11	0.96	0.05	0.90	0.15	0.52	0.21	0.98	0.04	0.97	0.05

ILR	N	Locative Verbs		Numerals		*Wear*-verbs	
		M	*SD*	*M*	*SD*	*M*	*SD*
1+	5	0.51	0.22	0.52	0.11	0.49	0.20
2	10	0.75	0.20	0.70	0.18	0.73	0.23
2+	12	0.76	0.13	0.73	0.19	0.73	0.17
3	17	0.79	0.09	0.84	0.13	0.81	0.15
3+	2	0.72	0.30	0.83	0.13	0.91	0.04
4	2	0.89	0.16	0.78	0.31	0.81	0.09

NOTE: Shaded areas indicate native-like accuracy (i.e., higher than .85).

Grammaticality Judgment

The results of the Grammaticality Judgment task are summarized in Table 3.

Scores on most items in the Grammaticality Judgment task showed a general tendency to increase with means from ILR 1+ to ILR 4, except for Particle Stacking and Conjunction. This tendency indicates that learners' grammatical knowledge increases as their oral proficiency level increases. The tendency is shown better with some features than others. For example, a sharp increase in judgment accuracy from 50% to nearly 90-100% from ILR 1+ to ILR 4 is found for Tense Dependency, Past Tense in Relative Clauses, Negation, Apperceptive, Locative verbs, and *Wear*-verbs. In particular, a steady increase from level to level, without the mean score falling, is found for Past Tense in Relative Clauses (.49 > .64 > .74 > .78 > .82 > .90), as shown in Figure 1.

Conversely, accuracy for the other three features does not reach a native-like ability (85%) even at higher proficiency levels on the ILR scale: Conjunction (55-84%), Particle Stacking (40-67%), and Numerals (52-84%). This suggests that even very advanced learners with native-like oral proficiency (ILR 4) cannot achieve native-like grammatical competence with these grammatical features compared to other features. The development of learners' knowledge of those linguistic features, and their relative difficulty, is depicted in Figure 2.

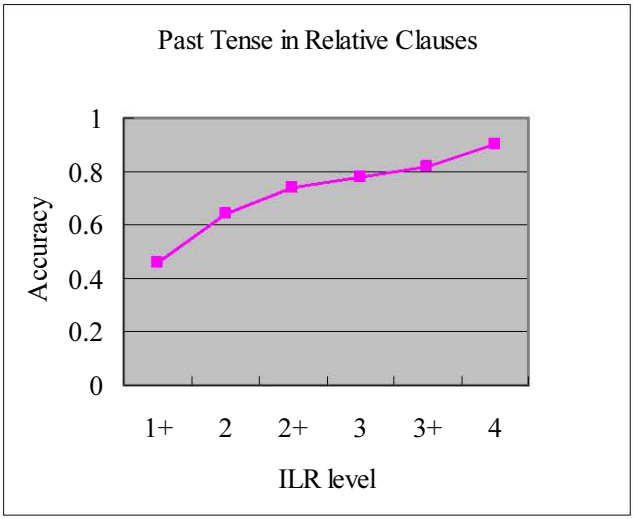

Figure 1. Past Tense in Relative Clauses

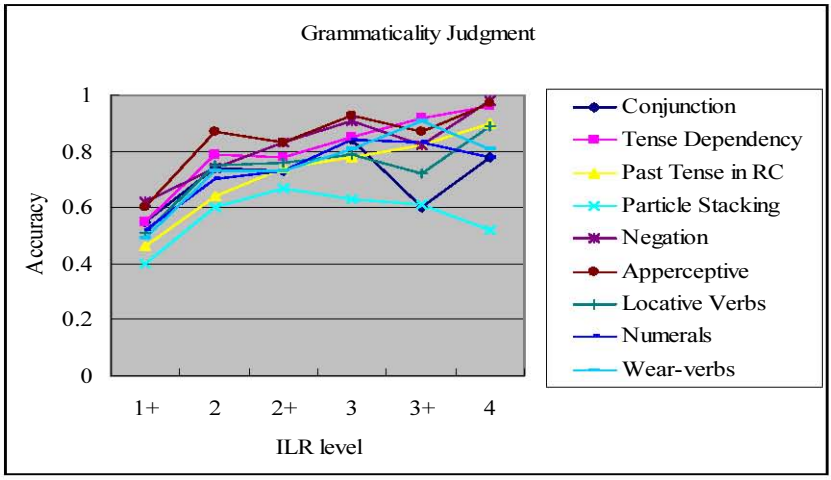

Figure 2. Grammaticality Judgment

Notice that the mean scores at ILR level 3 are higher than those at ILR 3+ in Conjunction (.84 > .6), Particle Stacking (.63 > .61), Negation (.91 > .82), Apperceptive (.93 > .98), Locative (.79 > .72) and Numerals (.84 > .83), and even those at ILR 4 in Conjunction (.84 > .78), Particle Stacking (.63 > .52) and Numerals (.84 >.78). Statistical analysis was not performed due to the small and unbalanced sample size. However, if any differences are identified subsequently, with more subjects added, these results could come about for different reasons. First, the oral proficiency interview (OPI) results for individual participants could be questioned, since there was no second rating that

might have improved the accuracy of OPI test scores. Another reason may concern the different properties of heritage speakers and L2 learners at each ILR level. Even though the difference does not appear in the OPI test score reflecting learners' oral proficiency in this study, it is possible that the difference is shown in their grammaticality judgments targeting knowledge of specific grammatical features. Comparing the population of our subjects at ILR 3, ILR 3+ and ILR 4 in Table 1 in terms of proportions of heritage speakers and L2 learners, we find that 82% of the participants in the ILR 3 group are heritage speakers whereas 18% are L2 learners of Korean. In contrast, at ILR 3+, 33% are heritage speakers and 67% L2 learners. The proportions are equal at ILR 4. If the findings of previous studies are correct, in that heritage speakers outperform L2 learners at comparable levels of overall proficiency, the decrease in accuracy from ILR 3 to ILR 3+ (even to 4) found in this study may be due to the higher proportion of heritage speakers at ILR 3. If this is true, a difference between the heritage and L2 learners is to be expected at both ILR 3+ and ILR level 4, with the mean scores at ILR 3 higher than those at higher ILR levels. The poor performance of L2 learners at ILR 3+ and ILR 4 would have lowered the mean scores shown in Table 3. In order to find out if this was true, individual data on the learners at those ILR levels were compared in Table 4. (There were only two L2 learners and one heritage learner at ILR 3+, and one heritage learner and one L2 learner at ILR 4.) One of the two L2 learners at ILR 3+ was randomly chosen to produce an equal number of subjects in each cell. The data from these subjects used in all the heritage and L2 learner comparisons included all those items before non-discriminating ones were eliminated).

Overall, large differences were found in heritage and non-heritage speakers at both ILR levels. Heritage speakers did much better than L2 learners (.90 vs. .67 at ILR 3+ and .87 vs. .75 at ILR 4). Both the heritage speakers at ILR 3+ and ILR 4 exhibited a very high level of competence in overall grammatical judgment ability (.90 at ILR 3+ and .87 at ILR 4), whereas neither of the L2 learners had achieved the same level of ability. However, even though the heritage speakers' overall grammatical judgments were very accurate, they still showed less accurate performance with certain linguistic features, such as negation (.75 in ILR 3+, .73 in ILR 4), conjunction (.75 in ILR 4) and apperceptive (.70 in ILR 4). Large differences between heritage speakers and L2 learners were found in past tense in relative clauses, (.95 vs. .65), locative verbs (.89 vs. .50), particle stacking (.90 vs. .60), numerals (.93 vs..73), and *wear*-verbs (.95 vs. .65) and conjunction (.90 vs. .50) at Level 3+, and in locative verbs (.89 vs. .72), particle stacking (.90 vs. 50), numerals (1.00 vs. .63), and *wear*-verbs (.90 vs. .75) at Level 4. The poor performance of L2 learners at ILR 3+ and high proportion of L2 learners in the ILR 3+ group compared to the ILR 3 group seem to be responsible for the ILR 3+ means being lower than the

Table 4. Grammaticality Judgments in Advanced Heritage Speakers and L2 Learners

Linguistic Features	ILR 3+		ILR 4	
	Heritage Speaker (26, F)	L2 Learner (35, M)	Heritage Speaker (36, F)	L2 Learner (45, M)
Tense Dependency	0.95	0.80	0.98	0.83
Past Tense in Relative Clauses	0.95	0.65	1.00	0.80
Locative Verbs	0.89	0.50	0.89	0.72
Particle Stacking	0.90	0.60	0.90	0.50
Negation	0.75	0.78	0.73	0.84
Numerals	0.93	0.73	1.00	0.63
Wear-verbs	0.95	0.65	0.90	0.75
Conjunctions	0.90	0.50	0.75	0.85
Apperceptives	0.85	0.80	0.70	0.80
GJT All	0.90	0.67	0.87	0.75

means at the lower level, ILR3, where 82% of the population consisted of heritage speakers. In order to achieve a better picture of the relationship between OPI level and grammaticality judgment ability with statistical analysis, the sample size must be increased and balanced, or else the linguistic background of participants must be controlled for, with a sample of either heritage or non-heritage learners.

To summarize, overall, learners' accuracy scores increased with increasing proficiency on the ILR scale. In particular, tense dependency, past tense in relative clauses, locative verbs, and negation seemed to be good indicators of learners' proficiency level. Better performance of heritage speakers than L2 learners was observed at the very high levels on most features in this grammaticality judgment task. More importantly, analysis of individual learner data at advanced levels revealed that heritage speakers were able to achieve a very high level of competence in their grammatical judgments on most of the linguistic features tested in this study, whereas L2 learners did not reach the same level of competence in most cases. Sample size must be increased to obtain a clearer picture of what may be a difference between the two groups across proficiency levels on the ILR scales.

Acceptability Judgment

The results for the acceptability judgment task are presented in Table 5.

Similarly steady developmental progress is also found for both honorifics (.72 > 1.00) and light verb idioms (.57 > .97). Native-like performance was achieved at ILR level 3 on both tests (.90 in honorifics, .89 in light verb idioms). A high correlation was found between the ILR scale and individual scores, as well as mean scores (see Table 2 for details). Similar steady development of

Table 5. Acceptability Judgment

ILR	N	Honorifics		Light Verb Idioms	
		M	*SD*	*M*	*SD*
1+	5	0.72	0.10	0.57	0.20
2	10	0.75	0.12	0.80	0.14
2+	12	0.82	0.18	0.83	0.12
3	17	0.90	0.12	0.89	0.09
3+	2	0.87	0.12	0.87	0.14
4	2	1.00	0.00	0.97	0.05

NOTE: Shaded areas indicate native-like accuracy (i.e., higher than .85).

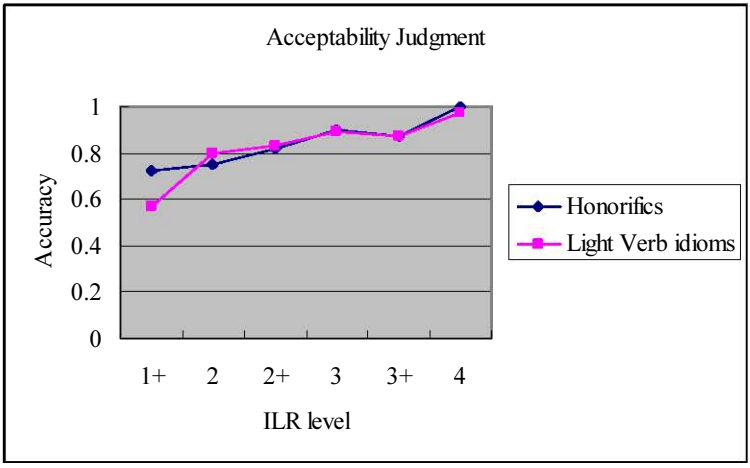

Figure 3. Acceptability Judgment

these two linguistic features is well depicted in Figure 3.

However, again, a slight drop in mean scores between ILR 3 and ILR 3+ is found on both tests: .90 vs. .87 in honorifics and .89 vs. .87 in light verb idioms. A comparison between the same heritage speakers and L2 learners in ILR 3+ and ILR 4 was conducted, in order to identify possible differences between the two groups on these tests. The results are shown in Table 6.

Table 6 reveals differences between heritage speakers and L2 learners in their grammatical competence in honorifics and light verb idioms. Heritage speakers achieved a very high level of competence with both linguistic features (85-100%), whereas neither of the L2 learners achieved the same level of competence as heritage speakers (65-80% for honorifics, and 79-88% for light verb idioms). A potential indication of fundamental differences between heritage speakers and L2 learners in the acquisition of Korean honorifics and light verb idioms is also visible in their level of attainment on this task.

Table 6. Acceptability Judgment in Advanced Heritage Speakers and L2 Learners

Linguistic Features	ILR 3+		ILR 4	
	Heritage Speaker (26, F)	L2 Learner (35, M)	Heritage Speaker (36, F)	L2 Learner (45, M)
Honorifics	0.85	0.65	0.85	0.80
Light Verbs (Idioms)	0.90	0.79	1.00	0.88
AJ ALL	0.88	0.72	0.93	0.84

Lexical Decision

The results for the Lexical Decision task are summarized in Table 7.

Table 7. Lexical Decision

ILR	N	Syllable Structure		Morpholo-gical Rule		Compound Nouns		Overall Vocabulary	
		M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
1+	5	0.76	0.04	0.56	0.05	0.57	0.05	0.79	0.10
2	10	0.71	0.16	0.59	0.15	0.60	0.10	0.74	0.11
2+	12	0.81	0.12	0.73	0.13	0.70	0.14	0.76	0.11
3	17	0.91	0.09	0.78	0.14	0.75	0.13	0.82	0.11
3+	2	0.81	0.27	0.52	0.25	0.54	0.06	0.69	0.19
4	2	0.87	0.06	0.84	0.15	0.85	0.16	0.91	0.12

NOTE: Shaded areas indicate native-like accuracy (i.e., higher than .85).

A steady increase in mean scores is found for morphological rule (.56 > .84) and compound nouns (.57 > .85). It seems that syllable structure and overall vocabulary were easier than morphological rule and compound nouns even for the beginners. Learners at ILR 1+ had already achieved more than 75% of accuracy in syllable structure (76%) and in overall vocabulary (79%). The development of learners' knowledge of linguistic features tested in this task and their relative difficulties are shown in Figure 4.

Notice, again, that the ILR 3 group mean is higher than the mean of the higher proficiency level ILR 3+ group for all of the lexical decision tasks, i.e., .91 > .81 for syllable structure, .78 > .52 for morphological rule, .75 > .54 for compound nouns, and .82 > .69 for overall vocabulary. Again, it is possible that the unexpected results are due to the imbalance in number of participants at levels ILR 3, 3+ and 4, and also in the number of heritage speakers and L2 learners at each level. The individual data for the same subjects at each proficiency level were drawn for comparison and are presented in Table 8.

Contrary to the findings for grammaticality judgments and acceptability judgments, the lexical decision data do not show an advantage for heritage speakers, except for overall vocabulary. In the cases of morphological rule and

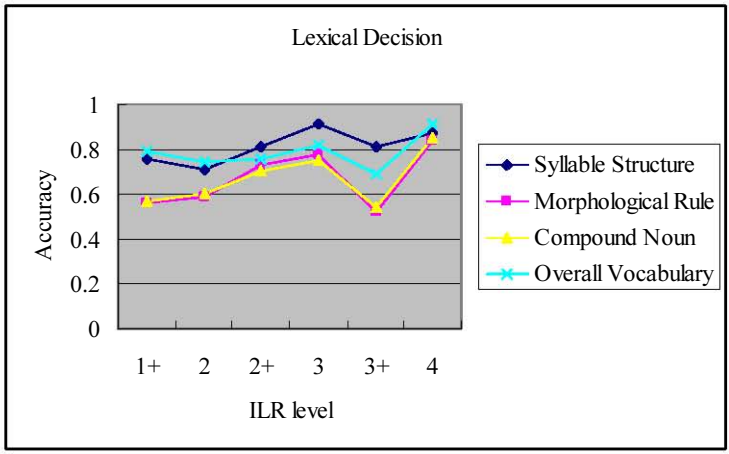

Figure 4. Lexical Decision

Table 8. Lexical Decision in Advanced Heritage Speakers and L2 Learners

Linguistic Features	ILR 3+		ILR 4	
	Heritage Speaker (26, F)	L2 Learner (35, M)	Heritage Speaker (36, F)	L2 learner (45, M)
Syllable Structure	0.62	0.89	0.92	0.84
Morphological Rule	0.35	0.68	0.50	0.66
Compound	0.53	0.54	0.51	0.66
Overall Vocabulary	0.79	0.57	0.95	0.69
LD All	0.57	0.67	0.72	0.71

compound, in particular, heritage speakers did not perform better than L2 learners at either proficiency level: .53 vs. 54 at ILR 3+ and .51 vs. .66 at ILR 4. The heritage learner at ILR 3+ did very poorly with syllable structure, with 62% accuracy, and with morphological rule, with 35% accuracy, compared to other linguistic features. Also, her score was the lowest of all participants on both the morphological rule and compound nouns tests. A close examination of her data on both tests revealed that most of the missed items concerned illegal and pseudo words (i.e., 'incorrect' words). That is to say, she had a tendency to give a 'positive response' (i.e., pressing the 'correct' button) to words she was not sure about, which resulted in wrong responses to all 'incorrect' items. This can lead to a very low score on this particular morphological rule test because it contained two thirds illegal and pseudo words and one third legal words (balancing the number of items across the three conditions, which resulted in two thirds incorrect and one third correct items). Conversely, the compound nouns test included the same number of true and non-existent words. This can explain the heritage learner's asymmetrical accuracy scores on

the same type of tests (i.e., morphological rule and compound nouns): .35 vs. .53. The same tendency was found in her syllable structure test responses. This individual tendency seems to have affected the correlation between the ILR scales and mean scores, showing no statistically significant results on most of the lexical decision tasks: syllable structure, compound nouns, and overall vocabulary. In order to prevent results from being affected by such individual tendencies, it is, of course, crucial to increase sample size (i.e., number of participants) and to balance legal and illegal items within each test rather than within a whole lexical decision task.

To summarize, the data from the lexical decision task showed that learners' knowledge of Korean morphological rules and compound nouns increased with their increasing overall proficiency level, as measured on the ILR scale. On the other hand, such development was not clearly found with syllable structure and overall vocabulary. More data are needed to confirm these findings.

Picture-Word Discrimination (PWD), AXB Distinction, Phoneme Monitoring

The results of the three tasks, picture-word discrimination, AXB distinction, and phoneme monitoring, are presented together for comparison because all these involve the same linguistic feature: Korean stop sounds. The tasks are designed to test L2 learners' Korean phoneme distinction ability. The results are summarized in Table 9.

Despite the fact that the phonemes tested in the three tasks are stops, regarded as difficult features for English-speaking learners of Korean, the results from the three different tests show that advanced learners of Korean have little problem distinguishing different stop sounds in their listening, plain vs. tensed sounds, in particular. However, scores vary depending on the type of task. Phoneme monitoring and AXB were easier than PWD, with even learners at ILR 1+ achieving 76% and 84% accuracy, respectively, on those two tests. Statistically significant Spearman correlation coefficients were found between ILR proficiency level and individual scores on the tasks, AXB and PWD, but not

Table 9. Picture-Word Discrimination, AXB Distinction, Phoneme Monitoring

ILR	PWD			AXB			Phoneme Monitoring		
	M	*SD*	*N*	*M*	*SD*	*N*	*M*	*SD*	*N*
1+	0.52	0.15	5	0.76	0.11	5	0.84	0.10	5
2	0.75	0.18	10	0.87	0.13	10	0.83	0.12	10
2+	0.79	0.15	13	0.89	0.10	12	0.81	0.22	12
3	0.84	0.11	17	0.94	0.07	17	0.91	0.13	17
3+	0.71	0.15	3	0.82	0.14	3	1.00	0.00	2
4	0.89	0.07	2	0.95	0.05	2	0.94	0.09	2

NOTE: Shaded areas indicate native-like accuracy (i.e., higher than .85).

on Phoneme Monitoring. No correlation was found between ILR level and mean scores on all three phoneme discrimination tasks (see Table 2), apparently due to a ceiling effect and the unbalanced number of subjects in each proficiency group. The lower mean scores on the PWD task, ranging from .52 at ILR 1+ to .89 at ILR 4, probably indicate that participants' ability to distinguish phonemes can be interfered with by top-down processing of sounds, as distinct from the processing employed by Korean native speakers. For example, when a subject is looking at a picture of *tal* 'moon' and hears *ttal* 'daughter,' instead of *tal*, she or he can easily take *ttal* with a tensed t, /t'/ as *tal* with a plain t, /t/, which could cause wrong responses, lowering the accuracy score. However, this type of negative effect of top-down processing was not found in native speakers' responses.

Learners' ability to distinguish Korean stop sounds on these tasks is well depicted in Figure 5. In particular, the steady development of ability is shown well with PWD.

Notice, again, that the mean scores for learners at ILR 3 are higher than for those at ILR 3+ on PWD and AXB. This, again, may be related to the imbalance between heritage speakers and L2 learners at the two levels, as discussed in the previous sections. The differences between the two types of learners on these two tasks are shown in the individual data presented in Table 10.

Table 10, again, shows better performance of heritage speakers than L2 learners on both tasks (85-98% accuracy for heritage speakers vs. 77-91% accuracy for L2 learners).

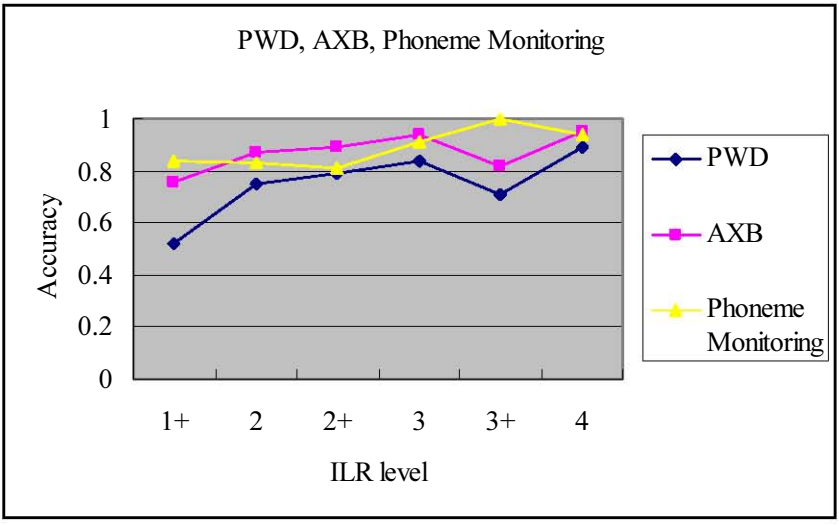

Figure 5. PWD, AXB, Phoneme Monitoring

Table 10. Picture-Word Discrimination and AXB Discrimination in Advanced Heritage Speakers and L2 Learners

	ILR 3+		ILR 4	
	Heritage Speaker (26, F)	L2 Learner (35, M)	Heritage Speaker (36, F)	L2 Learner (45, M)
Picture-Word Discrimination (PWD)	0.85	0.77	0.92	0.83
AXB	0.97	0.83	0.98	0.91

To summarize, tasks tapping L2 learners' knowledge of Korean phonemes showed that their control of Korean phonemic distinctions develops as their ILR proficiency level increases. Possible differences between heritage speakers and L2 learners were also found for picture-word discrimination and AXB discrimination. A ceiling effect was found in Phoneme Monitoring.

Picture-Sentence Matching: Reflexive caki

The results of picture-sentence matching: reflexive *caki* task are presented in Table 11.

Table 11. Picture Sentence Matching: Reflexive *caki*

ILR	Reflexive-*caki*		
	M	*SD*	*N*
1+	0.73	0.20	5
2	0.85	0.14	10
2+	0.86	0.16	13
3	0.94	0.11	17
3+	0.87	0.18	3
4	1.00	0.00	2

NOTE: Shaded areas indicate native-like accuracy (i.e., higher than .85).

The task seems easy for learners, compared to other tasks in this study. Native-like competence was found throughout the ILR 2 (85% accuracy) – ILR 4 (100% accuracy) range. However, statistically significant correlations were still obtained between proficiency scores measured on the ILR scale and individual, as well as mean accuracy, scores (see Table 2 for details). The increase in mean scores from ILR 1+ to ILR 4 is shown in Figure 6.

Once again, the lower mean score at ILR 3+ (.87) than that at ILR 3 (.94) seemed to relate to the different populations in these groups (that is, 82% heritage vs. 18% L2 learners at ILR 3, and 33% heritage vs. 67% L2 learners at ILR 3+). The comparison between the two individual mean scores at ILR 3+ revealed that heritage speakers did better than L2 learners (.90 vs. .85). No

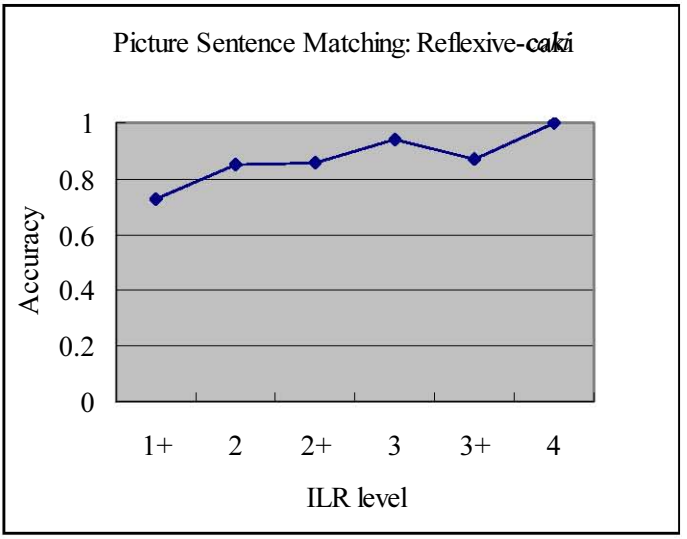

Figure 6. Picture Sentence Matching: Reflexive *caki*

such difference was found at ILR 4 (Note: *SD* = 0). Again, this indicates possible differences between the two learner types on this task, which can be confirmed with increased and balanced numbers of subjects in a subsequent study.

Accent Detection

The results of the accent-detection task are presented in Table 12. The overall low accuracy mean scores indicate the difficulty of the task. A statistically significant correlation was still found between proficiency measured on the ILR scale and individual mean scores, but not with the group mean scores (See Table 2 for details). The development of learners' ability to recognize regional accents in Korean is pictured in

Table 12. Accent Detection

ILR	Accent Detection		
	M	*SD*	*N*
1+	0.40	0.34	6
2	0.48	0.22	10
2+	0.69	0.30	13
3	0.72	0.29	16
3+	0.75	0.25	3
4	0.63	0.18	2

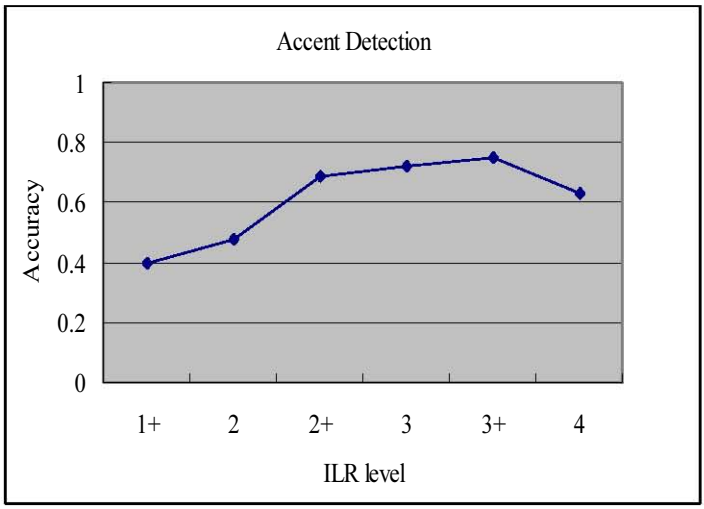

Figure 7. Accent Detection

Notice that the mean accuracy scores could be divided into two groups; one for ILR 1+ to 2, with an accuracy range of 40-48%, and the other for ILR 2+ to 4, with an accuracy range of 69-75%, indicating a sizeable improvement from ILR 2 to ILR 2+ in learners' ability to detect regional accents in Korean. Note, also, that the mean score for learners at the highest proficiency level, ILR 4 (.64), is lower than that for learners at lower proficiency levels, ILR 2+, 3 and 3+ (.69, .72, .75, respectively). Examination of the individual data revealed very poor performance of the L2 learner at ILR 4, compared to heritage speakers; 25% vs. 50% accuracy for all 12 test items. Such a difference was not found for learners at ILR 3+. Still, the overall mean score for learners at ILR 4 was not higher than that for learners at ILR 3+: 50% vs. 58% accuracy for all 12 test items. Since this is based on only one learner in each cell, more data with balanced numbers of subjects are needed to obtain more reliable data for comparison.

3. Discussion

Results of analyses of perception data on 21 tests tapping knowledge of various linguistic features in Korean and obtained from 52 learners of Korean (32 heritage speakers and 20 L2 learners) can be summarized as follows:

1. Most of the 21 tests showed significant correlations between learners' ILR proficiency scores and mean accuracy scores. Such correlations were found on a grammaticality judgment task for all features studied: Tense

Dependency, Past Tense in Relative Clauses, Particle Stacking, Negation, Numerals, Wear verbs, Conjunctions and Apperceptives, except for Locative Verbs. Both acceptability judgment tasks for Honorifics and Light Verb Idioms also showed statistically significant correlations between oral proficiency scores on the ILR scale and accuracy scores (both individual and group mean scores). The lexical decision task also showed high correlations between ILR proficiency scores and individual mean scores for three out of four features: Syllable Structure, Morphological Rule, and Compound Nouns, but not for the Overall Vocabulary test. Two of three tasks tapping learners' knowledge of Korean phoneme distinction, Picture-Word Discrimination and AXB also found statistically significant correlations between ILR proficiency scores and individual mean scores. However, the phoneme-monitoring task was not a good indicator of proficiency development. The picture-sentence matching: reflexive, *caki* task, as well as the accent-detection task, found statistically significant correlations between ILR proficiency scores and individual mean scores, even though learners' overall performance was better on the former task than the latter. Those linguistic features showing a high correlation between their mean accuracy scores and proficiency scores measured on the ILR speaking scale (OPI test) seem to be good indicators of learners' grammatical knowledge at certain oral proficiency levels. The tests that did not show significant correlations between proficiency level and accuracy scores were either too easy (e.g., Overall Vocabulary, and Phoneme Monitoring) or too difficult for all levels of learners (e.g., Locative Verbs).

2. The data also revealed relative difficulties among different linguistic features on the same type of task. For example, comparing mean accuracy scores for grammaticality judgments showed that even advanced level learners at ILR 4 could not reach native-like performance, that is, higher than 85% accuracy in this study, with Conjunction (55-84%), Particle Stacking (40-67%) and Numerals (52-84%). On the lexical decision task, difficulty was found with Morphological Rule (56-84%) and Compound Nouns (57-85%). The Accent-Detection task was also very difficult for the learners, producing a low mean accuracy range of 40% to 75%.
3. Mean accuracy differences between heritage speakers and L2 learners were found on many tasks at ILR 3+ and ILR 4. Superior performance by heritage speakers was observed with most grammaticality judgment tasks, except Negation for learners at ILR3+, and Negation, Conjunction and Apperceptives for learners at ILR 4. The same tendency was found on such tasks as acceptability judgment with Honorifics and Light Verb Idioms, lexical decision with Overall Vocabulary, AXB discrimination, picture-word discrimination, and Picture-Sentence Matching: Reflexive, *caki,* and Accent Detection. These results add more detailed information

to previous findings on differences in the language profiles of heritage and non-heritage speakers (e.g., Kanno et al. 2008, Y-G Lee et al. 2005).

Additional data from more participants, especially at very advanced proficiency levels (ILR 3+, 4 and 4+), are needed to test the robustness of the preliminary findings from this study (we call the findings preliminary due to the limited number of participants in each level). In addition, comparison of the findings from the perception data with those (upcoming) from the production data will reveal any correlations between the two distinct abilities within the same individual participants, each proficiency level, and each type of participant, i.e., heritage speakers and L2 learners.

4. Conclusions

The purpose of this study was to identify linguistic correlates of proficiency, that is, a set of linguistic features that correlate with levels of proficiency defined by a global proficiency measure, such as the OPI, and the ILR scale. Preliminary analyses of data presented in this paper show that most tasks have promise, some clearly so, despite the study having been conducted with truncated samples (the majority of participants were in the ILR 2-3 range). The concept underlying the project is clearly proven. It is, indeed, possible to identify detailed linguistic abilities and remaining linguistic needs for typical heritage and non-heritage learners at advanced levels on the ILR scale, and by implication, other proficiency scales. Observed relationships between ILR proficiency scale levels and scores on a wide range of perception tasks are robust at the level of individual means, as well as group means representing typical learners. Analyses of these tasks and features have revealed differences in the profiles of heritage and non-heritage speakers visible at the advanced levels under study, supporting findings in earlier studies, such as Kanno et al. (2008) and Y-G Lee et al. (2005), but much more remains to be done. Using information from the Language Experience Questionnaire, the data can also be analyzed to investigate the influence of such factors as age of first exposure (AO), length of residence in a target-language-speaking environment (LOR), and amount of formal instruction.

Implications for Language Testing

The finding of this study has implications for syllabus design, pedagogic materials development, and language testing, and for diagnostic tests in particular. On the basis of the findings to date, the current data-collection batteries will clearly yield an ample supply of tasks and task-types with which to pro-

duce reliable diagnostic measures for Korean (and by implication, for any language of interest). In a second phase of the study, tasks with proven internal reliability and effectiveness in differentiating among learners at various levels of the ILR scale will be improved by replacing poor items and increasing total numbers of items. On the basis of the results of this first major trial, unsuccessful tasks and task-types will be discarded from the data-collection battery.

Once improved versions of the batteries are obtained, a next step will be to apply the same measures and methodology to learners in the ILR 0 to 2 range. More attention will be focused on differences between heritage and non-heritage learners, with an eye towards modifying instructional materials and recommendations appropriately for each group.

Finally, attention must be devoted to identifying optimal ways of providing feedback on performance for differing end-user constituencies. The communication of information on individual learners' abilities and needs must obviously be transparent and accessible if it is to serve its diagnostic purpose. However, both the level of information and the way it is communicated should obviously vary for linguistically sophisticated language professionals and, e.g., learners with little or no use for meta-linguistic knowledge. Various types and forms of feedback should be used for different audiences, and trials of several are currently under way.

References

Alderson, J. Charles. (2005). *Diagnosing Foreign Language Proficiency: The Interface between Learning and Assessment*. London: Continuum.

Alderson, J. Charles, and Ari Huhta. (2005). The development of a suite of computer-based diagnostic tests based on the Common European Framework. *Language Testing* 22.1, 301-320.

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge: Cambridge University Press.

European Science Foundation. (2006). *Scientific Report of ESF Sponsored Exploratory Workshop: Bridging the Gap between Research on Second-language Acquisition and Research on Language Testing*. Ref.: EW05-208(SCH)

Kanno, Kazue, Tomomi Hasegawa, Keiko Ikeda, Yasuko Ito, and Michael H. Long. (2008). Relationships between prior language-learning experience and variation in the linguistic profiles of advanced English-speaking learners of Japanese. In: Donna Brinton, Olga Kagan, and Susan Bauckus, eds., *Heritage Language Education: A New Field Emerging,* 165-180. New York: Routledge..

Kunnan, Antony J. and Eunice E. Jang. (2009). Diagnostic feedback in language assessment. In Long, Michael H. and Doughty, Catherine, eds., *Handbook of Second and Foreign Language Teaching*, 610-627. Oxford: Blackwell.

Lee, Young-Geun, Hi-Sun H. Kim, Dong-Kwan Kong, Jong-Myung Hong, and Michael H. Long. (2005). Variation in the linguistic profiles of advanced English-speaking learners of Korean. *Language Research* 41.2, 437-56.

Long, Michael. H., Scott Jackson, Rajaa Aquil, Ilhan Cagri, Kira Gor and Sun-Young Lee. (2006). *Linguistic Correlates of Proficiency: Rationale, Methodology, and Content*. Technical Report, College Park: University of Maryland.

Pica, Teresa. (1983). Adult acquisition of English as a second language under different conditions of exposure. *Language Learning* 33.4, 465-497.

Polinsky, Maria. (1998). Heritage language narratives. In Donna Brinton, Olga Kagan, and Susan Bauckus, eds., *Heritage languages: A New Field Emerging,* 149-164. New York: Routledge.

White, Lydia. (2003). *Second Language Acquistition and Universal Grammar*. Cambridge: Cambridge University Press.

Sun-Young Lee
English Department
Cyber Hankuk University of Foreign Studies
270 Imun-dong, Dongdaemun-gu
Seoul, 130-791, Korea
E-mail: alohasylee@cufs.ac.kr

Jihye Moon
School of Language, Literatures, and Cultures
3215 Jiménez Hall
College Park, MD 20742, U.S.A.
E-mail: jimoon@umd.edu

Michael H. Long
School of Language, Literatures, and Cultures
3124 Jiménez Hall
College Park, MD 20742, U.S.A.
E-mail: mlong5@umd.edu

Received: January 29, 2009
Revised version received: October 26, 2009
Accepted: November 20, 2009