FAQ: Do Non-linguists Share the Same Intuition as Linguists?[*]

Sanghoun Song[**]
(Korea University)
Jae-Woong Choe
(Korea University)
Eunjeong Oh[***]
(Sangmyung University)

Song Sanghoun, Choe Jae-Woong, Oh Eunjeong. (2014). FAQ: Do Non-linguists Share the Same Intuition as Linguists?. *Language Research,* 50.2, 357-386.

When studying the nature of human language, we frequently ask ourselves the following question: Do native speakers agree with our judgments of the sentences in question? Many of us have encountered quite a few sentences which linguists report to be grammatical but which non-linguists find ungrammatical. Linguists try their best in their language analyses to accommodate the native speakers' intuitions in a systematic way, but these efforts are mostly confined to the so-called 'informal' method. A natural question that arises is if the naïve native speakers would agree to the introspective acceptability judgments. In order to properly answer this question, a rigorous and formal method that will ensure more systematic and fine-grained results is required. This paper aims to address questions relating to this issue, exclusively focusing on Korean. The present work intends to provide some substantive discussion on how similar or different linguists' intuitions are

[*] Acknowledgements: We are grateful to Chung-hye Han, Seok-Hoon You, Jungha Hong, Miran Kim, and Kyeong-min Kim for their help at the design stage of the experiment reported in this paper. An earlier version of this paper was presented at the conference hosted by the Linguistic Society of Korea (June 21, 2014), and we would like to thank Kiyong Choi and Jung Soo Mok for their helpful comments. In addition, we would like to express special thanks to Myung-Kwan Park and Hee-Don Ahn for their valuable comments on an earlier draft of this paper. Jon Sprouse kindly replied to our inquiries regarding a couple of statistical analyses in this paper and provided us with some useful ideas. Comments from three anonymous reviewers are also much appreciated. Joshua Crowgey proofread this paper and gave us valuable comments. Thanks are also due to Hye-Rin Hyeon, Yeon Woo Kim, and Hye Young Kim for their contribution. Needless to say, all remaining errors and infelicities are our own responsibility.

[**] first author

[***] corresponding author

to/from those of the general public estimating grammatical acceptability. Our main experiment was carried out with 138 subjects, using about one thousand sentences excerpted from two volumes of a linguistic journal. We calculated the convergence rate focusing on the pairwise sentences in the data, and the rate was computed to be 84.75%. This measure is somewhat lower than the convergence rate of 95% reported in Sprouse *et al.* (2013) for the English data.

Keywords: experimental syntax, acceptability judgment, intuition, PsychoPy, Likert scale, *z*-score, R

1. Introduction

This paper delves into one Frequently Asked Question in the syntactic study of the Korean language: Do Korean native speakers share the same intuitions as Korean linguists? Our experience of language studies tells us that linguists' judgments are not necessarily seamless. We have seen more than a few sentences whose acceptability judgment differs linguist by linguist. Moreover, there seems to be no evidence for believing that linguists have better intuitions about human language sentences than non-linguists, as was pointed out by some linguists.

(1) a. "As with any form of evidence, intuitive evidence will be most trustworthy when data from various sources-including linguists' intuitions, intuitions from nonlinguists, corpus evidence, and so on-all converge." (Clifton et al. 2006)

(2) b. "[W]e theoretical linguists had no privileged way of distinguishing the possible formal patterns of a language from the merely probable." (Bresnan 2011)

For this reason, disagreement in acceptability judgments on human language sentences is often controversial in language research (*inter alia*, Labov 1975; Greenbaum 1977). There have been several previous studies which address this question using quantitative methods, but the question still remains one of the fundamental issues in linguistic discussion.

"Unfortunately, the findings of the experimentalists in linguistics very rarely play a role in the work of generative grammarians. Rather, theory development tends to follow its own course, tested only by the unreliable and sometimes malleable intuitions of the theorists themselves. The theories are consequently of questionable relevance to the facts of language." (Wasow and Arnold 2005)

In a pioneering study, Spencer (1973) conducted an experiment to compare between linguists' and native speakers' intuition, using 150 sentences taken from six linguistic articles. The result indicates that the subjects (i.e., non-linguists) agree among themselves as to the acceptability or unacceptability of 80% of the sentences. Moreover, Spencer reports that the exemplars for which the subjects and the six authors share the same intuition account for only a half of the total. Gries (2013) claims that because linguists ponder over too many things about linguistic expressions, linguists' judgments may differ those of non-linguists. His experiment statistically indicates learning about meta-language has an effect on acceptability judgment of subjects. Recently, acceptability judgment testing has received much attention in the field of experimental syntax. In this line of research, Sprouse and Almeida (2012) calculated the reliability of acceptability judgment data created in a traditional fashion. They collected data provided in a syntax textbook (Adger 2003) and estimated the size of the discrepancy between these data and formally created (i.e., non-introspectively acquired) data. Their conclusion is that the discrepancy is maximally 2%. Sprouse *et al.* (2013) assessed acceptability judgments of the sentences exemplified in *Linguistic Inquiry* published over ten years (2001 to 2010). Their tests were conducted with 936 subjects using three experimental tasks and five statistical measurements, and they found a convergence rate of 95% between informal and formal methods, with a margin of error of 5.3-5.8%. These two recent experiments imply that non-linguists and linguists share almost the same intuition concerning languages.

For Korean, there have been several studies to conduct judgment testing for certain specific phenomena (C-h Han *et al.* 2011; H Ko and E Oh 2012; C-h Han 2013; B-S Park and S-R Oh 2013; Y-h Lee 2013; etc.). Unlike these studies, the current study looks at a variety of Korean sentences on a comprehensive scale. We tested the acceptability judgments of Korean speakers, using over 1000 sentences, rather than focusing on a single phenomenon. In addition, the experiment for the current study was conducted in a semi-automatic way, and the result acquired from the experiment was statistically analyzed in a fully automatic way using scripts in R (R Core Team, 2014) and in other programming languages.

This paper is structured as follows: Section 2 discusses the particulars of our methodology. Section 3 offers an overall explanation of how we carried out our experiment. Section 4 provides statistical analyses of the results obtained from the experiment and discusses what we learned. Section 5 summarizes the paper and presents further work that needs to be done based on our results here.

2. Methodology

The basic methods we employed for the current experiment are summarized in Table 1.

Table 1. Overview

# of articles	29	**target scope**	GG 18, 23
# of initial items	1,125	**random shuffling**	Y
# of chosen items	955	**random sampling**	N
# of pairwise sets	118	**result filtering**	Y (filler)
# of filler sets	44	**time checking**	Y
# of pretest sets	6	**experimental method**	standalone software
# of subjects	138	**tool**	PsychoPy
# of judgments	41,135	**task**	Likert (1-5)
# of items/tokens	300	**descriptive statistics**	Pearson correlation
# of items/subjects	344	**inferential statistics**	t-test, Wilcoxon test

2.1. Terminology

2.1.1. Informal *vs.* Formal

Sprouse *et al.* (2013) make use of two different terms for the basic methods of data collection in the syntactic studies: namely, an informal method and a formal method. The former has been widely used for the last several decades. In this method, a linguist provides a set of evidence for his or her linguistic arguments, and the set mostly consists of pair-wise sentences in which every condition is the same except for only one item which is intended to show the linguistic property in question (i.e., minimal pairs). Notably, in this method, the linguist judges the particular sentences to be (un)grammatical based on their own intuition. In contrast, formal methods gaining popularity over the last ten years or so rely on statistical measurements found through language processing experiments to indicate which sentence sounds (un)grammatical to native speakers. In the latter method, the linguist is forced to validate their predicted judgments by going out and soliciting the judgments of others, not just relying on their own intuition. That is to say, the data that we used in the current experiment are the ones that were originally created in an informal fashion, and we tested such informally constructed data in a formal fashion. This method facilitates identifying how much those with little knowledge of linguistics (i.e., naïve native speakers) corroborate the acceptability or unacceptability judgments of the linguists.

2.1.2. Acceptable *vs*. Grammatical

These two terms have sometimes been used synonymously with each other, but the aforementioned recent studies using the formal method strictly differentiate between them. *Acceptability judgment* refers to a perceptual rating, thereby being concerned with how good a sentence sounds. In other words, acceptability is a property of sentences that native speakers have conscious access to. In contrast, Schütze and Sprouse (2013) argue that *grammaticality judgment* is a misleading term: “Since a grammar is a mental construct not accessible to conscious awareness, speakers cannot have any impressions about the status of a sentence with respect to that grammar.” Following this distinction, the current study makes use of the term *acceptability judgment*, rather than *grammaticality judgment*.[1]

[1] A different point of view is provided in E Cho (1996; 1998). He regards the symbol ‘*’ as a marker for an ill-formed expression that a theory of grammar does not allow. In other words, his argument is that so-called grammaticality associated with ‘*’ is tantamount to acceptability only within a specific grammatical framework.

2.1.3. Random Shuffling and Sampling

Random sampling and random shuffling are very important in data-oriented studies of human language (J Hong, 2014): Outcomes based upon interactions with data that is not randomly ordered are apt to be biased as well as imperfect. As the same goes for experimental syntax, only randomized data has the statistical power to allow experiments to make sense with respect to a variety of language phenomena (Sprouse *et al.* 2013).[2)]

Technically speaking, so-called randomization can be classified into random sampling and random shuffling. The former has more to do with extracting samples used for statistical testing, while the latter has more to do with reordering the samples. In other words, the random sampling method extracts some portion of items and discards the rest, and the random shuffling method works with the entire items though the order is always changed. Both methods are crucial in data processing, but not all experiments can necessarily employ both. It is our understanding that at least one method between them has to be used in order not to provide a hasty conclusion.

We made use of random shuffling for generating stimulus sets and presenting the stimulus items to subjects. On the other hand, we did not randomly sample the data when gathering a set of pairwise sentences to show a contrast in acceptability. The main reason for not using random sampling is that our basic data were not big enough. Because of the shortage, we had no choice but to use all of the minimal pairs evidentially and directly presented in two volumes of journals on Korean syntax.[3)]

2.2. Data Compilation

This subsection provides the entire workflow of how we compiled our dataset, which is divided into four steps: Collection, selection, con-

[2)] For example, Sprouse *et al.* (2013) collected their basic data from articles published over ten years, and the initial data consisted of 3,635 data points. Out of them, they randomly sampled 450 pairwise phenomena.

[3)] This is a limitation of the current study we are fully aware of, so in a sense it is a kind of pilot study that requires a statistically stronger approach in the future.

version, and inspection.[4]

2.2.1. Data Collection

The initial step was to download the PDF files of the journal articles contained in *Studies in Generative Grammar* from the official website of the *Korean Generative Grammar Circle* (*http://www.kggc.org*). The main reason for the selection of the journal as the data source is that it contains the largest number of the 'standard' (See Footnote 5) acceptability judgments appraised in the traditional and informal way.

The *Korean Generative Grammar Circle* provides the PDF version of all articles published since 1991. For our experiment, we selected two of these volumes, one published in 2008 (*vol.* 18) and the other in 2013 (*vol.* 23). At the beginning, only the latest volume (published in 2013) was chosen, but we soon found that a single volume did not have enough sentences to provide the breadth of examples we required to for this broad-coverage experiment. In order to have a time interval between two volumes for a more comprehensive and balanced analysis, we added *vol.* 18 (published in 2008).

There were 42 articles in *vol.* 18 and 33 articles in *vol.* 23 (75 papers, in total). These PDF files were converted to MS-Word files (*.docx*) using an OCR (Optical Character Recognition) tool for ease of coping and pasting in gathering examples. Then, because we are only concerned with Korean sentences, the papers without Korean examples were filtered out, leaving as our basic dataset 29 papers that contain 1,125 sentences.

2.2.2. Data Selection

After collecting the basic data, we classified examples into two groups-the ones we could use as stimuli in the current experiment, and the others we exclude.[5]

[4] Although some of the discussion in this section may not sound crucial for the main topic of this study, we report them anyway since they would show how the whole process can be done in a mostly automatic fashion.

[5] Sprouse *et al.* (2013) divide data points provided in *Linguistic Inquiry* into several subtypes: Standard acceptability judgments (48%), coreference judgments (15%), in-

The examples excluded belong to one of the following types. First, examples with co-indexation, such as Xi and Xj, were filtered out, because this co-indexation convention could cause some confusion to those that are not familiar with them, even with some extra instruction. Second, examples involving felicity conditions (marked as #) were filtered out, because determining felicity requires access to critical contextual information which would not be available in the experimental setup. Third, Q/A pairs and examples consisting of two or more sentences were also filtered out, because the Likert scale (also known as a category scale) we employed in this study as an experimental task assumes that each stimulus is independent. This type of scale has been widely used in marketing surveys, which normally provides a questionnaire as exemplified in (3). It is noted that a stimulus item such as (4) is not fully adequate, because two propositions related to each other show up simultaneously.

(3) How much salty does this cookie taste?
|-----|-----|-----|-----|
1 2 3 4 5

(4) This cookie tastes salty. How much does this cookie match the sourness?
|-----|-----|-----|-----|
1 2 3 4 5

That is, it is not clear whether the answerer would base his/her response on the first or the second sentence in (4). This means that there must be no internal relations between stimuli in measuring on the Likert scale. Fourth, given that our experimental environment is text-based (i.e., not using an acoustic system), we ruled out examples in which prosodic information (tone, stress, duration, intonation, etc.) was included. Fifth, because we are exclusively interested in sentence processing, examples lacking any finite verb (e.g., fragments composed of only NPs) were filtered out.

terpretation judgments (23%), judgments involving relatively few lexical items (12%), and judgments involving prosodic manipulations (2%). We made use of the same distinction.

2.2.3. Data Conversion

The next step was to convert Yale Romanized strings (a sequence of alphabetic characters) into *Hangul* strings. For this purpose, a non-deterministic finite state transducer was implemented, whose basic skeleton is roughly sketched out in Figure 1.[6]

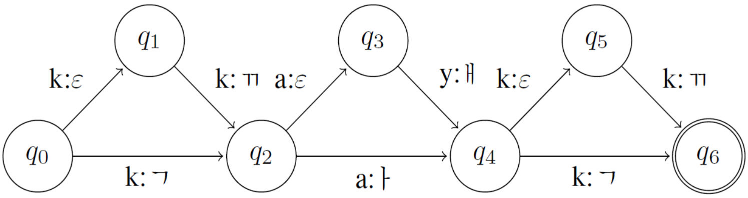

Figure 1. Yale2Hangul Automaton.

After each transduction was completed, the running script invokes an independent module to unify the vowel and the two types of consonants (i.e., *choseng*, *cwungseng*, and *congseng*) into a single syllable.[7] This process is carried out character by character, and each character is incrementally gathered to form a set of words. Note that conversion in the direction of Yale→Hangul does not necessarily have one-to-one mapping unlike that in the opposite direction, because there is no delimiter between syllables. For example, a Yale Romanized form *hakkyo* 'school' can be analyzed as either *ha.kkyo*, *hak.kyo*, or *hakk.yo*, and amongst them only the second one is registered in the dictionary. The unwanted ones, such as *ha.kkyo* and *hakk.yo*, are removed by two substeps. First, the transducer runs a postprocessor which filters out words that contain an uncanonical *Hangul* characters.[8] For instance, given that *hakk* is not a productively used character in Korean, *hakk.yo* is filtered out in this stage. Second, we made use of the spell checker built in MS-Word. For instance, a misspelled word *ha.kkyo* is underlined in MS-Word. One in-

[6] Notice that arrows via *q*1, *q*3, and *q*5 in Figure 1 serve to handle double consonants and double vowels. For more information, see Mohri (1997).

[7] This module was implemented by Hye-Shik Chang under the GNU Lesser General Public License (i.e., free software): *https://raw.githubusercontent.com/sublee/hangulize/master/hangulize/hangul.py*

[8] The set of canonical characters was acquired from the *Sejong* corpora (*http://www.sejong.or.kr*).

spector manually deleted such misspelled words and left only correct forms.

2.2.4. Data Inspection

The final step in data compilation was to proofread the converted data items with reference to the original papers. This step checked whether converted sentences written in Hangul were well-formed, whether word spaces were correctly inserted, and whether sentence items out of the scope of the current experiment were filtered out. One additional task we carried out in this step was to substitute uncommon proper names with names familiar to non-linguists. For instance, quite a few papers use Kim as a gender-neutral personal name, but the name may not sound familiar to ordinary Koreans. In order for this unfamiliarity not to interfere with the result of our experiment, such an uncommon name was replaced (e.g., Kim sensayng 'Sir/Madam. Kim'). This task was iterated four times by different proofreaders. The total number of stimulus items compiled in this way is 955.

2.3. Toolkit

In the current experimental study, we made use of an open-source toolkit for psychological experiments entitled PsychoPy (Peirce 2007; 2009).[9] From a viewpoint of designing linguistic experiments, we examined several options in terms of robustness, reusability, random shuffling, concentration degree of subjects, ease of data analysis, and so on. Additionally, we took several technical factors into consideration, including character encoding, format of input and output files, stability of platform, etc. This survey led us to select PsychoPy as the experiment toolkit for this study.

[9] This software package is implemented in the Python programming language (*ver.* 2.7 and 32-bit machine-based) and distributed under the GPL license. Of course, there are other available toolkits, including DMDX (Foster and Foster 2003), Linger (Rohde 2003), WebExp (Keller *et al.* 2009), and Amazon Mechanical Turk.

3. Experiment

3.1. Subjects

A total of 154 adult native speakers of Korean (male 53; female 104) participated in the present experiment. They ranged in age from 20 to 31 (mean age 21.19, median age 20). They were all university students in two universities in Seoul. Table 2 provides the summary of subjects. Note that the subjects were divided into four groups.

Table 2. Four Groups of Subjects

abbreviation	group	description	count
GEN	generalization	subjects majoring in English	126
REF	reference	subjects of other majors	12
TEST	test	unit-test in implementation	6
PILOT	pilot	construction of the filler sentences	10
		total	154

Out of 154, 138 subjects took part in the main task, which involved two subgroups of the subjects. The first subgroup of subjects (abbreviated as GEN), which formed the majority of the subjects (126 out of 138), were English majors. In order to ensure that responses from these subjects were not biased by their majors and to exclude the possibility that their responses were distinct from Koreans not majoring in English, the second batch of data was collected from 12 subjects whose major was other than English (abbreviated as REF). Their majors were diverse, such as computer science, political science, German, fine arts, etc. Responses from these two different groups revealed no significant differences and, thus, we decided to merge them into one. We turn to this point in detail in §4.2.

The second group of subjects was included in the experiment for different purposes. Out of the 16 subjects, 6 subjects (abbreviated as TEST) were involved in the main task but their results were separately stored to test the source code of data munging and data analyses. The remaining 10 subjects (abbreviated as PILOT) took part in a pilot study, whose purpose was to select filler sentences.

Prior to the main acceptability judgment task, the subjects were asked to fill out a short questionnaire, which collected information such as age, gender, major, and place of birth (to check the type of dialect each subject uses). Such information was collected in order to assess any potential bias driven by these factors (§4.3).

Response-based outlier removal was performed. Out of 138 participants, 16 were excluded from the data analysis because they failed to meet a 68% accuracy criterion on the 44 filler sentences. The rationale was to set up a criterion stringent enough to exclude outliers, whose performance was at chance, but relaxed enough to include as many participants as possible (§4.1).

3.2. Materials and Presentation

As the main task, a five-point Likert scale task was administered to the subjects. This scale was meant to capture a five-way distinction in acceptability, with 1 being labeled *most acceptable*, 5 being labeled *least acceptable* (marked as '*' in syntactic literature), and the midpoint 3 being labeled *so-so* (in between two opposite values of acceptability, normally marked as '??'). In this task, the subjects were asked to rate the acceptability of the sentences provided on the screen using the five-point scale.

Prior to administration of the main acceptability judgment task, a training session was provided in order to familiarize the subjects with using the five-point scale. Six pretest sentences were employed during the session. In order to ensure that the subjects would be exposed to a wide range of acceptability, these pretest sentences included two each of most acceptable (1), least acceptable (5), and so-so (3). These items were taken from S-J Chang (1995). In selecting pretest items, sentence length was taken into consideration and sentences with comparable length were selected. The pretest items included points of the grammar whose acceptability is rather straightforward. Since the purpose of the pretest sentences was to ensure that all the subjects were on the same page with respect to acceptability ratings, the practice sentences were identical, and were presented in identical order across subjects. These items were excluded from the data analysis.

In the following main task, 300 test sentences were presented along with 44 filler sentences. The selection of the filler sentences was determined by the responses from a pilot study. Out of 214 sentences used for the pilot study (excerpted from J-i Kwon (1995)), we selected 44 sentences showing the least variation in terms of acceptability ratings, thereby being considered as a representative example of each point of acceptability. Both test and filler sentences were counterbalanced in terms of acceptability, including either an equal or a comparable number of grammatical and ungrammatical sentences. In such a way, the effect of either "yes" or "no" response bias was minimized. The filler sentences served to determine whether subjects were outliers. As aforementioned, in order to be included in the data analysis, the subjects needed to meet a 68% criterion on the 44 filler sentences.

A total of 344 sentences were presented in random order to control for ordering effects. Moreover, the order in which these test and filler sentences were presented was different across subjects, so that each subject rated his or her own set of data with its unique distribution.

Testing took place in a classroom environment (i.e., a computer laboratory) with the subjects tested individually or in small groups. The subjects completed the task at their own pace, without a time constraint. Nevertheless, for the vast majority of the subjects, the task lasted about 30 minutes.

After gathering the experiment results, we examined the whole datasets once again. In this step, we discovered two errors and got rid of them them from the data points.[10]

[10] First, one experiment log file was corrupted. This corruption cropped up when the subject responded to stimulus items of the reference set (REF). This was a technical error caused by PsychoPy. The computer program requires a substantial size of resource (a memory size, a capacity of the video card chipset, etc.), and if the computer that the program runs on gives relatively low performance, this error may occur. The broken log file was abandoned. Second, we found that one article which was supposed to be discarded in the process of data selection (§2.2.2) was included. This mistake happened while transferring the data from one computer to the other one. As a consequence, five unwanted stimulus items on average were included into each test item consisting of 300 sentences. These problematic entries

4. Statistical Results and Discussion

First of all, we calculated the Pearson's product-moment correlation for all of the data. This measure indicates the degree of correlation between two variables, or two columns in a data table. The data table in this step is taken from the generalization set (i.e., GEN) produced by 126 subjects. The first data column consists of the linguists' judgment made in an informal way (headed as *ling*), and the second data column is filled with responses that subjects provided (headed as *response*). There are three types of correlation values. The first value is computed using only the filler set. Recall that this set consists of 24 sentences perfectly acceptable (indicated as 1) and 20 sentences assumed to be highly unacceptable (indicated as 5). The second value is computed using only the experiment set (abbreviated as *exp*), and the third value is computed bringing the first and second sets together. These values are 92.43%, 42.5%, and 49.98%, respectively. Notably, it has been reported that "the Pearson correlation is robust with respect to skewness and non-normality" (Norman, 2010).

These initial measures have a profound significance with respect to the research question of this paper. When subjects are tested with the filler set which is presumably composed of sentences with less controversial acceptability status, subjects' responses are highly correlated to the values given in the *ling* column. In contrast, the correlation coefficient between *ling* and *response* for the experiment set sharply drops to a value below 50. These measures indicate that difference between non-linguists' intuition and linguists' intuition shows up in the result of the present experiment. The following subsections delve into whether the difference is indeed significant by means of more elaborate statistics.

4.1. Filtering

The first step in data analysis was to filter out outliers in the data table. Schütze and Spouse (2013) state that there seems to be no clear

were all eliminated prior to data analysis. This is the reason why the measure of data points per subject in Table 3 (§4.3) is approximately 295, not exactly 300.

consensus as yet about how to eliminate outliers for acceptability judgments though they agree with the necessity of using filler data for language processing experiments. In other words, removing the outliers presumably ends up with an ad-hoc approach to data. In the present study, we ruled out the outliers with reference to the response accuracy on respect to the filler sentences. The accuracy ratio increases if and only if a subject responded correctly (i.e., 1 for a grammatical sentence and 4-5 for ungrammatical sentences). We calculated the accuracy ratio (μ) and the standard deviation (σ) of each subject's result. The average accuracy ratio with respect to all 126 subjects who participated in the generalization experiment is 0.8551, and the standard deviation is 0.0274.

The next step was to set up a threshold. If the accuracy ratio of a subject is over the threshold, we can assume that the subject conscientiously responded to the stimuli given in the experiment. We grouped the accuracy ratios of the subjects, and then calculated the Pearson's correlation for each group. For instance, if the threshold is chosen as 70%, the correlation was measured using the data provided by only subjects whose accuracy ratio is over 0.7. There were 28 ratio groups, whose plot charts are presented in Figures 2 and 3.

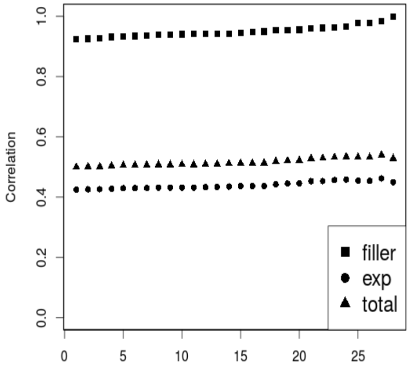

Figure 2. Distribution of correlation values.

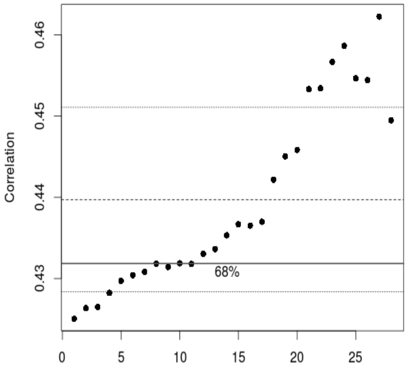

Figure 3. Distribution of correlation of [*exp*].

Given that the correlation was computed with reference to three types of sets as mentioned above, Figure 2 also contains three types of plots: namely, *filler*, *exp*, and *total*. All the distribution patterns in Figure 2 look fairly flat, indicating that there seems to be no significant increase by the ratio groups. Focusing on only *exp*, we created Figure 3, in which the difference between the lowest plot and the highest plot is less than 4%. The dashed line in the middle stands for μ, and the two dotted lines up and underneath stand for $\mu+\sigma$ and $\mu-\sigma$, respectively. All of them look unnotable, but we took notice of the first flatland (represented as the solid line) along the ridge. The flatland starts with the eighth group, whose accuracy ratio is approximately 68%. We decided to use this ratio as the threshold for filtering out the outliers. As a consequence, we filtered the responses of 14 subjects out of the dataset.

4.2. Comparison

Next, we compared two experiment results: The generalization set (GEN) and the reference set (REF). Recall that the former consists of data mostly created by those whose major is English, whereas the latter consists of data created by subjects majoring in various disciplines. The average accuracy ratio of the REF set is 0.8017, and the standard deviation is 0.0959. This means that the values of $\mu+\sigma$ and $\mu-\sigma$ of these two sets overlap with each other. What is notably different between these two sets is the response time. The average response time of the GEN set is 3.82 seconds with 69 milliseconds of standard deviation. Those values of the REF set are 3.16 seconds and 200 milliseconds. This difference seems to be caused by reading skill rather than any difference in acceptability judgments. As we did with the GEN set, we calculated Pearson's product-moment correlation coefficient for the REF set measuring the responses of the students against the judgments of the linguists. The values were 0.9102 for *filler*, 0.4396 for *exp*, and 0.512 for *total*. These measures are not much different from those of the GEN set provided in the previous subsection. For a more statistically rigorous comparison, we figured out boxplots for each set. Figure 4 indicates that the two sets do not show such a statistically different distribution with respect to responses.

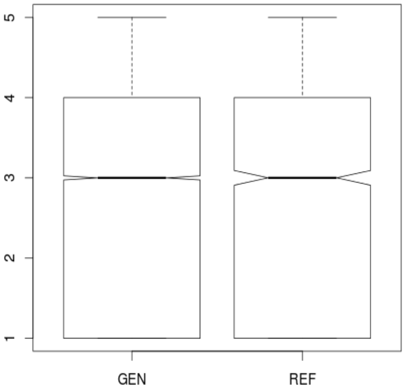

Figure 4. Response (GEN *vs.* REF).

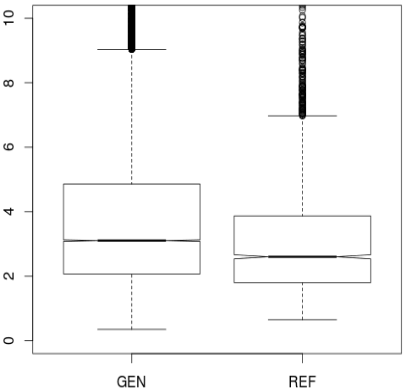

Figure 5. Time (GEN *vs.* REF).

Likewise, Figure 5 shows a small but not significant difference in response time.[11] In Figure 5, the GEN set involves more outliers than the REF set, because the GEN set contains sets more than ten times larger than the REF set. In sum, no significant difference between the GEN set and the REF set was found. Applying the same threshold (68%), we filtered out two datasets out of the 11 datasets of the REF set, and then added the remaining nine datasets into the original generalization set. The newly created generalization set consisted of 121 datasets: 112 taken from the original set (GEN) and 9 taken from the reference set.

[11] One reviewer commented there were too many outliers in Figure 5, Figure 10, and Figure 12. They are all related to response time. Because our experiment was not carried out under a tightly controlled environment, there could be quite a few reasons for the deviation in response time. Nonetheless, since the response time is not the main part that this study examines, these outliers do not have an influence on the present analysis.

4.3. Descriptive Analysis

Table 3. Basic Measures

# of data points	35,811	**gender: F/M**	53.14%/46.86%
data points per subject	295.96	**length: μ (σ)**	46.98 (19.23) bytes
correlation	43.22%	**time: μ (σ)**	4.00 (3.44) sec.
ling: 1	17,661	**response: 1**	10,795
ling: 2	846	**response: 2**	5,038
ling: 3	1,183	**response: 3**	4,067
ling: 4	1,013	**response: 4**	7,962
ling: 5	15,108	**response: 5**	7,949

Now we have the new dataset created in the previous subsection, and the correlation between *ling* and *response* was computed again. The values are 0.9357 for *filler*, 0.4322 for *exp*, and 0.5077 for *total*. That is, there still exists a significant gap between *filler* and *exp*. The average of the values in the *ling* column is 2.86, whose standard deviation is 1.92. That in the *response* column is 2.92, and the standard deviation is 1.57. These measures may not look close to each other *prima facie*, but their distribution patterns are quite different as indicated in Figures 6 and 7. These two histograms show that linguists' judgments are predominantly either 1 (with no mark) or 5 (marked as '*'), while non-linguists' responses in the current experiment are spread across the scale. Note that 2, 3, and 4 in Figure 6 refer to the data points marked as '?', '??', and '?*', respectively.

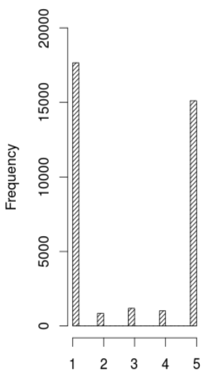

Figure 6. Histogram (*ling*).

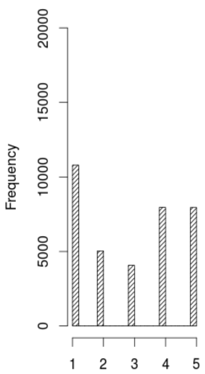

Figure 7. Histogram (*response*).

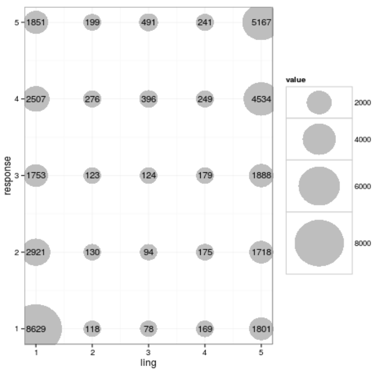

Figure 8. Bubble matrix between *ling* and *response.*

Figure 8 represents the matrix table between *ling* and *response*. For instance, in the whole data table, the rows in which the value of *ling* is 1 and that of *response* is also 1 (the leftmost and the bottom, viz. linguists' and subjects' judgments converge) number 8,629 and account for 48.86% out of the leftmost bubbles. The antipode which represents the rows in which the value of *ling* is 5 and that of *response* is also 5 (i.e., the rightmost and the top, viz. linguists' and subjects' judgments converge). The number of this bubble is 5,167 (34.2% out of the rightmost bubbles). The diagonal from lower left to upper right contains the numbers of responses that agree exactly with the linguists' judgment (14,299 responses, accounting for 39.93% of the total). The most intriguing cases are the bubble at the leftmost and the top (1,851, 5.17% out of the whole bubbles) and the bubble at the rightmost and the bottom (1,801, 5.03%). The former represents the rows in which linguists' judgments are 'most acceptable', but non-linguists' judgments are 'least acceptable'. The latter represents the rows in the opposite case. Particularly, these two bubbles have an adverse influence on the correlation between linguists' judgments (i.e., *ling*) and non-linguists' judgments (i.e., *response*).

Turning to variants related to the subjects, there are two factors: namely, dialect and gender. Figures 9 and 10 show how much the dialect variant has an influence on the response value and the response time, respectively.

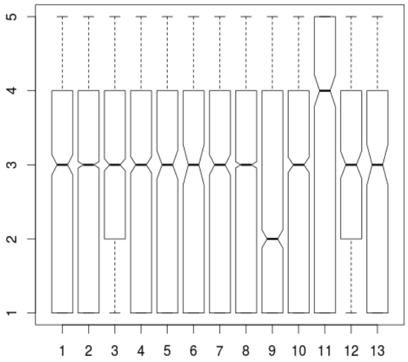

Figure 9. Response by dialects.

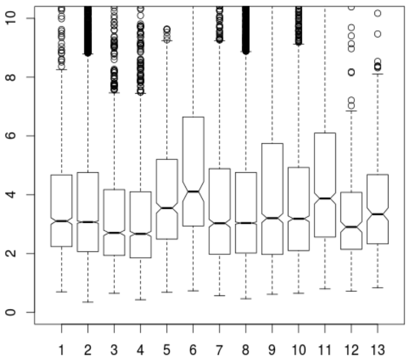

Figure 10. Time by dialects.

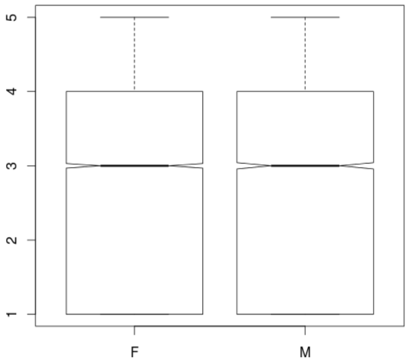

Figure 11. Response by gender.

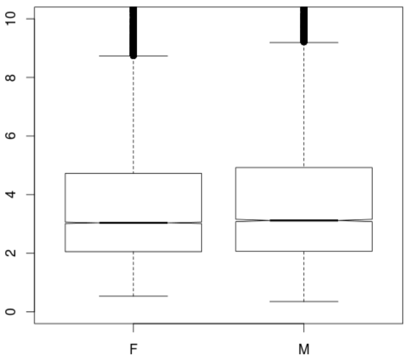

Figure 12. Time by gender.

Two boxplots on the ninth and the eleventh turn off the mean line, but this seems to be because the sample sizes of these cases are smaller than those of the others. Figure 10 indicates that the dialect variant makes a difference in the response time just as with Figure 5. Figures 11 and 12 are concerned with the gender variant. Likewise, no significant difference is found at least with respect to responses as shown in Figure 11. In sum, no variant has a significant effect on the distribution of the responses of the subjects.

4.4. Inferential Analysis

The whole dataset has one potential problem in terms of statistical analysis: The 1-5 scale in this experiment is not genuinely interval. For instance, it is not certain that the perceptual difference between 1 and 2 is the same as that between 4 and 5. Moreover, there may be individual variation in scaling: Subjects can respond more or less parsimoniously to the same sentence, even though all of them consider the acceptability of the sentence dubious. In order to overcome these potential flaws, the inferential analysis of the present study was made after converting the values into *z*-scores. A *z*-score transformation is a common statistical way of standardizing data on one scale. That is to say, *z*-scores that function like a common yard stick for all types of data tell us how far a particular score is away from the mean. For the current study, this *z*-score transformation is required to get rid of such a scale bias that may happen with rating tasks (Schütze and Sprouse, 2013).

Figures 13 and 14 show that linguists' judgments and non-linguists' judgments are significantly different from each other even when the values are transformed into *z*-scores. Figures 15 and 16 indicate that the distribution patterns of the two sets of judgments are still distinguishable from each other. The bars in Figure 15 (i.e., linguists' judgments) congregate to either 1 or 5, whereas those in Figure 16 are widely spread.

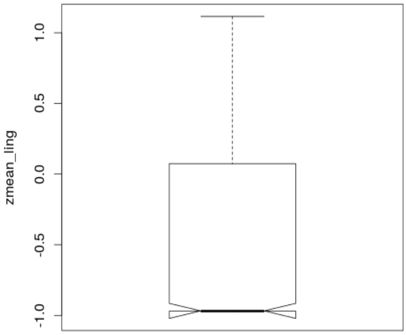

Figure 13. Boxplot (*ling*: *z*-score).

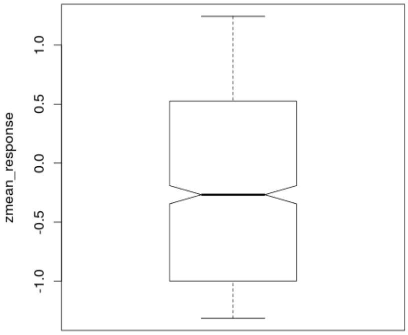

Figure 14. Boxplot (*response*: *z*-score).

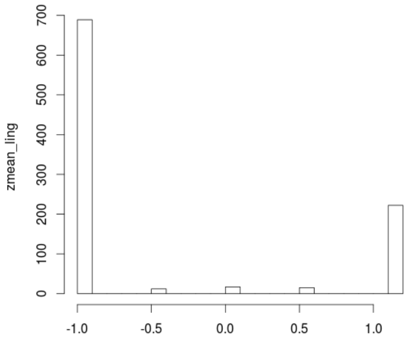

Figure 15. Histogram (*ling*: *z*-score).

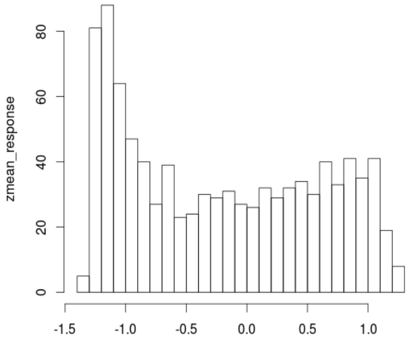

Figure 16. Histogram (*response*: *z*-score).

Given such differences, we set up the following hypothesis for inferential analysis. This hypothesis asks the main question of the present study: Do Korean native speakers share the same intuition as Korean linguists?

(5) H0: Non-linguists' judgments are the same as the linguists' judgments.
H1: Non-linguists' judgments are not the same as the linguists' judgments.

The distribution of the difference between *ling*'s *z*-score and *response*'s *z*-score is sketched out in Figure 17.

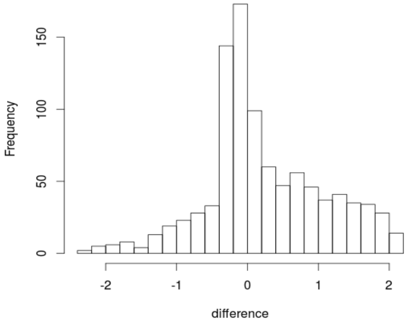

Figure 17. Histogram of difference.

The distribution of the difference between *ling*'s *z*-score and *response*'s *z*-score looks normal as shown in Figure 17, but this distribution does not pass the normality tests as shown in (6).

(6) a. **Shapiro-Wilk normality test**
data: index.resp.zmean$diff
W = 0.966, p-value = 4.008e-14
b. **One-sample Kolmogorov-Smirnov test**
D = 0.2067, p-value < 2.2e-16
alternative hypothesis: two-sided

Thus we tried a non-parametric test as well as the t-test: The two-tailed paired t-test as parametric testing and the Wilcoxon signed-rank test as non-parametric testing.[12] The statistical results are provided in (7),

[12] See Norman (2010) for applying a parametric test to data that do not strictly meet the so-called pre-conditions for parametric tests. Gries (2003) argues that in such cases it is preferable to use both parametric testing and non-parametric testing from a conservative standpoint.

respectively. The p-values in (7a-b) indicate that the H0 in (5) is rejected irrespective of the parametric type of testing. Therefore, it can be concluded that its logical alternative H1 in (5) is supported by the experimental data.

(7) a. **Paired t-test**
t = 8.3798, df = 954, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
    0.1741968 0.2807378
sample estimates:
mean of the differences
    0.2274673
b. **Wilcoxon signed rank test with continuity correction**
V = 281535, p-value = 4.102e-10
alternative hypothesis: true location is not equal to 0
95 percent confidence interval:
    0.1304200 0.2516536
sample estimates:
(pseudo)median
    0.1924921

4.5. The Convergence Rate

There is another statistical dimension to be explored regarding the comparison between linguists' and non-linguists' acceptability judgments. Linguists typically include in their discussion minimal pairs of sentences that presumably show a sharp contrast in acceptability, which in turn support some theoretical points being discussed by the linguists. Finding a "convergence rate" is supposed to sharpen the contrast in the native speaker's responses in order to make them more comparable to the binary contrasts represented in the linguists' judgments.

As a basis to find the convergence rate, a total of 118 minimal pairs of sentences from the 955 stimulus items were collected which comprise all the available pairs we could find from the given data set.[13] In order to calculate the convergence rate, adopting the same method as given in Sprouse *et al.* (2013), the response values (*z*-scores) of the 'acceptable' sentence in each minimal pair were compared with those of the 'unacceptable' sentence using the t-test and the Wilcoxon test. If there was a significant difference between the two in the predicted direction, as would be assumed by the linguists, then the intuitions of the subjects and the linguists would 'converge' with each other. Note that it is possible that the subjects' intuitions go in the opposite direction to the linguists', responding that the 'acceptable' sentence sounds worse than the 'unacceptable' sentence. Table 4 shows the results regarding the directionality of the response.

Table 4. Descriptive Analysis of the Directionality of the Responses

* based on the difference between means for each phenomenon.

predicted direction	opposite direction
113	5

Table 5 provides a fuller categorization of the results of statistical tests, namely, the t-test and the Wilcoxon test for each pair.[14]

Table 5. Categorized Results of Statistical Tests

* Significant p-values are defined at $p < .05$ in each direction; marginal p-values are defined at $p \leq .1$ in each direction.

	two-tailed t-test	Wilcoxon
significant in the opposite direction	1	1
marginal in the opposite direction	0	0
non-significant in the opposite direction	4	4
non-significant in the predicted direction	11	11
marginal in the predicted direction	3	3
significant in the predicted direction	99	99
total	118	118

[13] Again, no random sampling was administered due to the small size of the data set, which could weaken the statistical power of the following analyses.

[14] Though the numbers between the 'two-tailed t-test' column and the 'Wilcoxon' column in Table 5 are identical on the surface, there are two cases that are significant in one but not in the other, and vice versa, but the difference is still very slight.

Based on the numbers in Table 5, we can now calculate the convergence rates, and they are given in Table 6.

Table 6. Convergence Rates

only the significant results	including the marginal results
83.90% (99/118)	86.44% (102/118)

5. Summary and Outlook

The present study conducts acceptability judgment testing to substantiate to what degree naïve Korean native speakers share the intuition with Korean linguists. The data of linguists' acceptability judgments were collected from two volumes of articles contained in *Studies in Generative Grammar* (published in 2008 and 2013), and these data are assumed to be constructed in the traditional and informal way used for the last few decades. We chose 955 examples from the original data source, randomly shuffled in order for each subject to have different sets of stimulus items. The toolkit used for the current experiment is PsychoPy, which provides a standalone experimental environment. The experimental task in the present study was a five-point Likert scale, in which 1 stands for *most acceptable* while 5 stands for *least acceptable*. There were 154 subjects in total, and they were divided into four subgroups: namely, (i) pilot (10 subjects), (ii) test (6 subjects), (iii) reference (12 subjects), and (iv) generalization (126 subjects). Each subject of the last three groups responded to 350 stimulus items consisting of 6 pretest items, 44 filler items, and 300 experiment items.

According to our data analysis, the correlation between linguists' judgments and non-linguists' judgments is less than 0.5. This number implies that there exists a difference between the two types of acceptability judgments. For more sophisticated analyses, we filtered out outliers from the data table, using a 68% accuracy rate as the threshold and comparing the distribution pattern of two groups of subjects (i.e., generalization *vs.* reference). As a result, we analyzed the responses of 121 subjects, which include a dataset consisting of 35,811 data points.

The basic distribution of linguists' judgments looked different from that of non-linguists' judgments: The former congregates to either 1 or 5, whereas the latter is spread all across the scale. Since the five-point scale does not refer to an interval scale in a pure sense, we converted the raw data into *z*-scores. Using this converted measure, we examined the question of whether the acceptability judgments of these two groups are significantly different. All hypothesis tests we conducted rejected the null hypotheses, indicating there is a statistical reason to believe that linguists' acceptability judgment on the data in question is not supported by the general public. Finally, we calculated the convergence rate focusing on the pairwise sentences in the data, and the rate was computed to be 84.75%. This measure is somewhat lower than the convergence rate 95% reported in Sprouse *et al.* (2013) for the English data.[15]

The results reported in this paper should be considered with the caveat that they are valid only within certain limitations. For one thing, the size of the acquired data was not large enough for us to exploit the full-scale random sampling method. For another, the number of subjects could be increased for a more solid and stable result. These limitations in the number of test sentences and subjects would make the discussion of specific examples and linguistic phenomena somewhat premature in this paper, and that is why we decided not to include such discussion in the current study. There are still other aspects of the experiment that can be explored for more diverse and statistically powerful results. We leave these and others for future studies.

[15] All materials and scripts used in this study are available upon request, following the spirit of data and code sharing for academic research (Pedersen, 2008; Halchenko and Hanke, 2012). They include:
a. Yale2Hangul Finite-State Transducer: implemented in Python (*ver.* 2.7), cp949-based
b. pretest set: taken from Chang (1995)
c. filler set: taken from Kwon (1992)
d. stimulus items: taken from *vol.* 18 and *vol.* 23 of *Studies in Generative Grammar*
e. stimulus set generator: implemented in Python (*ver.* 2.7), creating CSV-formatted files, random shuffling, requiring two parameters (one for the number of stimuli, one for the number of sets)
f. PsychoPy source code: implemented in Python (*ver.* 2.7), 32-bit, UTF8-based
g. source code for computing *z*-score: implemented in Perl (*ver.* 5.18)
h. R script for statistical analysis: R ver. 3.0.2 (2013-09-25), or the later version
i. running script for data munging: a shell script for Linux / a batch profile for Windows

References

Adger, David. (2003). *Core Syntax: A Minimalist Approach*. Oxford: Oxford University Press.

Bresnan, Joan. (2011). A Voyage into Uncertainty. In a volume of essays by Reed College alumni on the occasion of the college's centenary, edited by Roger Porter and Robert Reynolds, Reed College.

Clifton, Charles Jr., Gisbert Fanselow, and Lyn Frazier. (2006). Amnestying Superiority Violations: Processing Multiple Questions. *Linguistic Inquiry* 37.1: 51-68.

Chang, Suk-Jin. (1995). *Information-based Korean Grammar* [in Korean]. Hanshin Publishing, Seoul, Korea.

Cho, Euiyon. (1996). Functional Grammar and Grammaticality. *Language Research* 32.3: 477-489.

Cho, Euiyon. (1998). Theoretical Commitments and the Use of the Asterisk '*' in Functional Linguistics. *Discourse and Cognition* 5.2: 285-292.

Forster, Kenneth I. and Jonathan C. Forster. (2003). DMDX: A Window Display Program with Millisecond Accuracy. *Behavior Research Methods, Instruments and Computers* 35: 116-124.

Greenbaum, Sidney. (1977). Contextual Influence on Acceptability Judgements. *Linguistics* 15.187: 5-12.

Gries, Stefan Th. (2013). *Statistics for Linguistics with R: A Practical Introduction (2nd Edition)*. Berlin: Walter De Gruyter.

Halchenko, Yaroslav O., and Michael Hanke. (2012). Open is not enough. Let's take the next step: an integrated, community-driven computing platform for neuroscience. *Frontiers in Neuroinformatics* 6.

Han, Chung-hye. (2013). On the Syntax of Relative Clauses in Korean. *Canadian Journal of Linguistics* 58.2: 319-347.

Han, Chung-hye, Dennis Ryan Storoshenko, and R. Calen Walshe. (2011). An Experimental Study of the Grammatical Status of *caki* in Korean. In Ho-min Sohn, Haruko Cook, William O'Grady, Leon A. Serafim, and Sang Yee Cheon (eds.), *Japanese/Korean Linguistics* 19, pages 81-94, Stanford, CA: CSLI Publications.

Hong, Jungha. (2014). A Corpus-linguistic Approach to Random Samples [in Korean]. *Language Information* 18: 137-162.

Keller, Frank, Subahshini Gunasekharan, Neil Mayo, and Martin Corley. (2009). Timing Accuracy of Web Experiments: A Case Study using the Webexp Software Package. *Behavior Research Methods* 41: 1-12.

Ko, Heejeong and Eunjeong Oh. 2012. A Hybrid Approach to Floating Quantifiers: Some Experimental Evidence. *Linguistic Research* 29.1: 69-106.

Kwon, Jae-il. (1992). Korean Syntax [in Korean]. Minumsa, Seoul, Korea.

Labov, William. (1975). *What Is a Linguistic Fact?* Lisse: Peter De Ridder

Press.

Lee, Yong-hun. (2013). An Experimental Approach to Multiple Case Constructions in Korean. *Language and Information* 17.2: 29-50.

Mohri, Mehryar. (1997). Finite-State Transducers in Language and Speech Processing. *Computational Linguistics* 23.2: 269-311.

Norman, Geoff. (2010). Likert Scales, Levels of Measurement and the "laws" of Statistics. *Advances in Health Sciences Education* 15.5: 625-632.

Park, Bum-Sik and Sei-Rang Oh. (2013). Re-identifying Null Arguments with Ellipsis: A Reply to Ahn and Cho (2013). *Studies in Generative Grammar* 23.4: 797-822.

Pedersen, Ted. (2008). Empiricism is Not a Matter of Faith. *Computational Linguistics* 34.3: 465-470.

Peirce, Jonathan W. (2007). PsychoPy – Psychophysics Software in Python. *Journal of Neuroscience Methods* 162.1: 8-13.

Peirce, Jonathan W. (2009). Generating Stimuli for Neuroscience Using PsychoPy. *Frontiers in Neuroinformatics* 2.

R Core Team. (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rohde, Doug. (2003). *Linger: a Flexible Platform for Language Processing Experiments*. (*http://tedlab.mit.edu/~dr/Linger*)

Schütze, Carson T and Jon Sprouse. (2013). Judgement Data. In Robert J. Podesva, Robert Podesva, and Devyani Sharma (eds.), *Research Methods in Linguistics*, pages 27-50, Cambridge: Cambridge University Press.

Spencer, Nancy J. (1973). Differences between Linguists and Nonlinguists in Intuitions of Grammaticality-Acceptability. *Journal of Psycholinguistic Research* 2.2: 83-98.

Sprouse, Jon and Diogo Almeida. (2012). Assessing the Reliability of Textbook Data in Syntax: Adger's Core Syntax. *Journal of Linguistics.* 48.3: 609-652.

Sprouse, Jon, Carson T. Schütze, and Diogo Almeida. (2013). A Comparison of Informal and Formal Acceptability Judgments Using a Random Sample from *Linguistic Inquiry* 2001-2010. *Lingua* 134: 219-248.

Wasow, Thomas and Jennifer Arnold. (2005). Intuitions in Linguistic Argumentation. *Lingua* 115: 1481-1496.

Sanghoun Song
Department of Linguistics, Korea University
Anam-dong Seongbuk-gu, Seoul, 136-701 Korea
E-mail: sanghoun@uw.edu

Jae-Woong Choe
Department of Linguistics, Korea University
Anam-dong Seongbuk-gu, Seoul, 136-701 Korea
E-mail: jchoe@korea.ac.kr

Eunjeong Oh
Department of English Education, Sangmyung University
7 Hongji-dong Jongno-gu, Seoul, 110-743 Korea
E-mail: eoh@smu.ac.kr

Received: June 30, 2014
Revised version received: August 11, 2014
Accepted: August 15, 2014