Linguistic Research 34(3), 387-426
DOI: 10.17250/khisli.34.2.201712.007

What do you mean by contrast in syntax?*

Sanghoun Song[**a] · Eunjeong Oh**[***b]**
(Incheon National University[a] · Sangmyung University[b])

Song, Sanghoun and Eunjeong Oh. 2017. What do you mean by contrast in syntax? *Linguistic Research* 34(3), 387-426. The present work proposes a non-binary evaluation of contrast in syntax and elaborates the benefits of considering gradience and degree in syntactic study. Contrast is one of the most important notions in contemporary linguistics, but the field lacks consensus about its definition and its role. Syntactic contrast has long been used in the field as a binary means of distinguishing grammatical from ungrammatical sentences via introspective judgments. Given that not all (un)grammatical sentences sound equally good or bad, contrasts should also be posited as gradient, thereby being measurable on a continuum of acceptability. The present study argues that a gradient view of syntactic contrasts is often more informative, revealing a greater variety of syntactic underpinnings in human language. To substantiate the merits of the gradient view of contrast over the dichotomous view, the present study presents results from a series of experiments conducted on Korean. The test items consist of 287 sentence pairs randomly extracted from *Studies of Generative Grammar 1991-2014*. The experimental tasks include a two-alternative forced choice task, a binary yes/no task, and a 5-point Likert scale task. The analysis is four-pronged covering direction, position, distance, and intensity of contrast. First, the direction of syntactic contrasts is examined with respect to whether linguists' judgments and naïve speakers' judgments converge with each other. Second, the position of contrasts on acceptability continuum examines absolute goodness of sentences on the assumption that the 'good' and 'bad' sentence must absolutely sound good and bad to the vast majority of speakers. Third, the distance of contrasts pertains to strength of grammatical constraints. This indicates the magnitude of difference in acceptability between two pairwise sentences. Finally, the intensity of contrasts relates to judgment variation across speakers. This examines whether most naïve speakers agree with the acceptability of a particular sentence and if not, it estimates how much their judgments are scattered in terms of the gradience of acceptability. **(Incheon National University · Sangmyung University)**

Keywords contrast, dichotomy, gradience, acceptability judgments, language experiments

* We would like to thank two anonymous reviewers for their helpful comments, though it should be noted that we did not fully accommodate the comments in this article. Needless to say, all remaining errors and infelicities of expression are our own responsibility. This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2015S1A5A2A03048485).

** First author
*** Corresponding author

1. Introduction

The present study proposes that contrast in syntax be evaluated as not merely a matter of either-or distinction (i.e., dichotomy, ordinarily marked with or without asterisks) but also a matter of degree (i.e., gradience). Contrast is one of the most important measures in linguistic studies because it indicates how an isolated grammatical factor behaves and how different factors interact with each other. Yet, there seems to be no clear consensus about what the linguistic role of contrast is and what kind of contrast is crucial (Hall 2011). The present study argues that syntactic studies need to examine contrast in a nuanced way and from various angles.

Roughly speaking, contrast is identified when two linguistic elements occur in the same linguistic environment. When a linguistic difference results from replacing one with the other, we have identified what is commonly referred to as a minimal pair. The notion of minimal pairs originated in phonological studies and has been long extended to syntax to describe sentences that differ in acceptability. While contrast between minimally pairwise items serves as an essential tool in many fields of linguistics, establishing such pairs is not necessarily straightforward. Phonological studies in which contrast plays a particularly pivotal role have paid keen attention to the definition of contrast (*inter alia*, Flemming 2004; Dresher 2009; Hall 2011; Łubowicz 2012). The appropriate definition of contrast in syntax has been comparatively understudied, though the debate on categorical grammaticality and continuous acceptability has existed since the early 2000s (Schütze 1996; Edelman and Christiansen 2003; Phillips and Lasnik 2003; Newmeyer 2003; Featherston 2005a, 2005b; Wasow and Arnold 2005; Sorace and Keller 2005; Fanselow et al. 2006; Devitt 2006, 2010; Sampson 2007; Phillips 2009; Myers 2009a, 2009b; Culbertson and Gross 2009; Gross and Culbertson 2010; Gibson and Fedorenko 2010; Culicover and Jackendoff 2010; Cable and Harris 2011).

In one view, syntactic contrasts reveal grammaticality, a binary distinction between ‘good’ and ‘bad’ sentences. From another perspective, syntactic contrasts reflect acceptability of sentences, which exists along a mental spectrum admitting that the continuous property may be (partially) caused by extra-grammatical factors (Newmeyer 2003; Sprouse 2007; Phillips 2009). From this stance of gradience, contrast is such that can be measured on the continuum of acceptability. While the former regards contrast in syntax as a binary distinction between two (or more)

expressions in a minimal pair, the latter regards it as a degree of difference between them.

When marking contrast in syntactically constructed minimal pairs, syntacticians sometimes note directionality of contrasting items and sometimes note magnitude. Linguists often make a syntactic argument by means of simply presenting pairwise sentences with or without an asterisk, and sometimes make use of diacritics to show the degree of acceptability of the 'bad' sentence, such as '?', '??', '?*', '*?', and '*'. These two representations of contrast are often present in the work of a single author, and even within the context of single publication. Only a few studies have addressed this ambivalent presentation of contrast in syntax, but its prevalence certainly implies "the absence of clear criteria on how to record and interpret intermediate judgments which can lead to serious inconsistencies" (Sorace and Keller 2005:1500). Such co-existence of two different usages of contrast (i.e., directionality and magnitude of contrasts) may imply that neither of them is solely sufficient to verify syntactic arguments and they are complementary to each other.

The present study raises the following critical question: What do we mean by contrast in syntax? The dichotomy approach has been widely and successfully employed as a primary tool in syntax, and we do not propose abandoning its use entirely (Phillips 2009). We argue that the approach of regarding contrast as gradience reveals a more variety of syntactic underpinnings in human language. The present work covers a series of experiments substantiating the merits of the gradience view of contrast for syntactic study. The language the present study carries out a series of experiments within is Korean. The experimental tasks include a two-alternative forced choice task, a binary yes/no task, and a 5-point Likert scale task. The analysis is four-fold: direction (§4.1), position (§4.2), distance (§4.3), and intensity of contrast (§4.4).

This article is structured as follows: Section 2 goes over fundamental background for understanding the current analysis. Section 3 provides an explanation of the experiments the present study conducted, and through them Section 4 shows that the gradience view of contrast in syntax effectively discloses the linguistic factors underlying linguistic phenomena. Building upon these findings, Section 5 substantiates that definition of contrast as magnitude of difference between pairwise sentences elucidates the nature of human language in greater detail. Section 6 concludes with some final thoughts on syntactic contrasts from a perspective of experimental syntax.

2. Fundamentals

2.1 Contrasts in Syntax

As aforementioned, the correct treatment of contrast has been central to phonological studies for quite some time. While phonological factors have also traditionally been treated as Boolean features, such as [±stressed] and [±rounded], recent studies have begun to advocate cataloguing magnitude of contrast rather than simple binary distinction. Flemming (2004), proposing so-called dispersion theory, regards constraints that mandate perceptual distinctness in the phonetic realization as a key concept of contrast. In this theory, selection of phonological contrasts is subject to three functional goals, one of which is to 'maximize' the distinctiveness of contrasts (i.e., contrast maximization). Flemming approaches to contrasting segments in human language from a viewpoint of 'likelihood' of confusing the two sounds. Thus, distinctiveness in Flemming's theory is measureable and predicted to be relevant to the markedness of a contrast between two sounds. Bauer (2008), in a similar vein, criticizes the traditional view that regards contrast as an either-or decision without intermediate step, in that the notion of contrast basically refers to "a perceptual notion, one that psychologically real (Ibid., p. 93)" and also has to do with actual behavior of human language. Bauer further highlights that degree is an important aspect of contrast: "The contrast between /s/ and /t/ is much more central to the system of German than is the contrast between /ç/ and /x/. Contrast is a matter of how strong the contrast is, not just a matter of whether there is a difference. (Ibid., p. 98)"

The views presented above share three key understandings of contrast. First, contrasts are correlated with actual behavior of human language. Second, contrasts result from perceptual processing in our brain. Third, contrasts have to be treated as strength of constraints.[1] These views also hold true for syntactic contrasts. If we posit syntactic constraints as magnitude of difference in acceptability between two

1 This is relevant to the notion of so-called candidates in Optimal Theory, as Flemming (2004:232) says "In Optimal Theoretic terms, this means that there are constraints favoring less confusable contrasts over more confusable contrasts." Note, incidentally, quite a few previous studies that place emphasis on gradience of acceptability in syntactic studies take Optimal Theory as the grammatical framework (*inter alia*, Sorace and Keller 2005; Keller 2006; Myers 2015). Of course, the present study does not necessarily follow the basic framework of Optimal Theory.

(or more) pairwise sentences rather than an either-or distinction, contrast in syntax can be understood as directly correlated with the strength of constraints on the grammatical factor which have caused the difference in acceptability (Sorace and Keller 2005).

A direct comparison of contrast in syntax and phonology serves to clarify the property's role in syntactic argumentation. First, while phonological contrasts correspond to physically existing sounds, syntactic contrasts interact with abstract mental structures. For instance, two sounds that form a minimal pair can be phonetically transformed into discernibly visually distinct spectrograms. In contrast, there is no comparable means of undoubtedly representing differences of acceptability between pairwise sentences. Therefore, contrast in syntax is tacitly relevant to the cognitive process of perceptual distinctiveness in linguistic constraints. Second, in phonology, contrasted items in a minimal pair all separately and equally exist in the language, and native speakers can be presumed to have extensive experience using and hearing these sounds in their daily speech. In contrast, syntactic contrasts are normally represented with two sentences, one presumed to be grammatical and the other presumed to be ungrammatical. Since an ungrammatical sentence violates one or more grammatical constraints in the language, ideally it is non-existent in the language. Nonetheless, when syntacticians construct ungrammatical sentences to unfold their argumentation, although these constructed ungrammatical sentences are the same to the extent that they are inexistent in reality, we should note that not all ungrammatical sentences constructed sound equally bad (Keller 2000) and thus the issue of the magnitude of contrast arises. Likewise, not all grammatical sentences sound equally good. In actuality, acceptability forms a continuum from 'absolutely bad' to 'absolutely good' and as such, the magnitude of contrast ensues.

2.2 Two Strands of Syntactic Theorizing

Although 'acceptability' has sometimes been used as a synonym with 'grammaticality', the two notions are distinct. In fact, this distinction has been addressed since the early days of contemporary linguistics (*inter alia*, Lakoff (1977)). In the intervening years studies have taken a variety of stances, but the current consensus appears to be that, while a wider range of (non-)linguistic phenomena can be interpreted along a spectrum of acceptability, grammaticality is an either-or question.

Given I-language *vs.* E-language distinction drawn by Chomsky's (1986), grammaticality is seen as of an internal generative language system property not directly accessible through conscious introspection. As such, grammaticality judgments are not possible. In contrast, 'acceptability' refers to a perceptual rating, and these ratings combine to form a continuum from the least acceptable to the most acceptable. Acceptability, then, is not equivalent to grammar, but it is a function of and reflection of grammar. As such, any native of a language can assess the acceptability of a sentence in that language. While grammaticality plays a pivotal role in understanding the nature of language competence, it is acceptability judgments which form the basis of linguistic theories, and it is acceptability judgments which syntacticians cite in support of those theories. Acceptability, when leveraged properly, serves as a window into how abstract grammar operates in our brain admitting that the window is not perfectly transparent (Phillips 2009).

2.2.1 Introspection *vs.* Experiment

One difference between the two strands of syntactic theorizing is thrown into sharp relief with respect to methodological framework. From the perspective of categorical grammaticality, the judgments on a set of sentences are intuitively made. From the stance of continuous acceptability, the language experiments in which acceptability judgments of native speakers are calibrated are emphasized. Wasow and Arnold (2005) criticize the heavy reliance on introspective intuitions of well-formedness, arguing that unsystematic data collection eventually has an adverse influence on the empirical basis of syntactic studies.[2] Concerns about the intuition-based nature of syntactic arguments have been raised by many researchers over the years, though not all of them take the stance that introspection is nugatory in linguistic study (Schütze 1996; Cowart 1997; Edelman and Christiansen 2003; Featherston 2005b; Sorace and Keller 2005; Fanselow et al. 2006; Sampson 2007; Myers 2009a, 2009b; Culbertson and Gross 2009; Gross and Culbertson 2010). Conversely, the value of introspection has also been reasserted by a number of syntacticians (Newmeyer 2003; Phillips and Lasnik 2003; Phillips 2009; Devitt 2006,

2 Sampson (2007) more pungently expresses a criticism of using the introspective method in the study of grammar, rather radically arguing that ungrammaticality is just an illusion (i.e, grammar without grammaticality). Although we do not follow such an argument, Sampson's claim about how data and experiments are importantly dealt with in many fields of contemporary science is noteworthy.

2010; Culicover and Jackendoff, 2010; Gibson and Fedorenko 2010). This latter group has gained some support from the results of recent studies showing that introspective judgments are often substantiated by subsequent empirical work (Featherston 2005a; Sprouse and Almeida 2012; Sprouse et al. 2013; Erlewine and Kotek 2016).

Our stance is that formal and empirical methods do not repudiate human introspection on grammar. In fact, formal judgments themselves consist of many instances of introspection as they are the result of large-scale, systematic collections of native-speaker intuitions. Most of previous experimental studies question only the coverage of introspective intuitions as noted by Erlewine and Kotek (2016:483): "(i) relative few participants, (ii) a relatively small number of target stimuli, and (iii) cognitive biases on the part of the researchers and participants". Moreover, the experimental findings draw us into a conclusion that there is little evidence for believing that the informal method itself is responsible for making a tenuous argument in syntactic studies. As Phillips (2009) argues, it is necessary that distinguish data and theory; Part of the problem is caused by the researchers who rely on less vetted data. One potential problem of informal methods, less vetted data, can be remedied by the formal experimentation involving a large sample size and population. In such a way, we believe that formal methods can supplement informal methods, although not all problems in syntactic data can be necessarily troubleshot by taking formal methods.

2.2.2 Dichotomy *vs*. Gradience

Syntactic contrasts, the primary focus of the current work, are another area in which marked differences arise between approaches that favor categorical grammaticality and those which highlight continuous acceptability. The assumption that language competence is categorically and essentially dichotomous results in the characterization of syntactic contrasts as an either-or decision (i.e., grammatical or not). These categorizations are obtained through the use of minimal pairs: Which sentence sounds good and which one bad? In contrast, gradience is central to investigations of acceptability which seek to reveal a continuum of judgments. Contrast in this view reveals degrees of well-formedness. Studies along these lines reveal that even linguistic phenomena that appear categorical show clear signs of gradience upon closer inspection.

Gradience itself has been conceptualized in two ways (Traugott and Trousdale, 2010: 22). The first approach understands gradient judgments as frequency effects working off of the assumption that grammar is quantitative and learned from exposure to other speakers (Bod et al. 2003; Bresnan et al. 2007; Bresnan and Hay, 2008). This first model holds that differences in frequency of particular types of utterances are central to assessments of acceptability. A second approach regards gradience as a function of the strength of the syntactic constraints implicated, often focusing on goodness of exemplar in syntactic literature (Sorace and Keller 2005; Fanselow et al. 2006). This view argues that because linguistic categories sometimes have blurred edges, dichotomous data fail to provide a clear-cut division between fully acceptable sentences and fully unacceptable sentences. Instead linguistic examples come in varying degrees of acceptability. Furthermore, syntactic constraints themselves may vary in the degree to which they must be strictly obeyed. Note, incidentally, that the very early approach in generative grammar makes an analysis of degrees of grammaticality (Chomsky 1964).

2.2.3 Our Standpoint

The theoretical and empirical work presented here make plain that that categorical grammaticality and continuous acceptability are correlated but distinct from each other in human language. That is, grammaticality and acceptability are not interchangeable designations. The present study conceptualizes the relationship between the two as a matter of conditioning: Grammaticality is not a necessary condition but a sufficient condition for acceptability. The violation of grammatical rules always involves unacceptability, but not vice versa; unacceptable sentences do not necessarily mean the violation of grammatical rules.

If a sentence in a minimal pair conflicts with a native speaker's mental grammar, this is experienced as a difference in acceptability between the sentences. It is important to note, however, that not all differences in acceptability are conditioned by syntactic factors (*inter alia*, Newmeyer 1983, 2003; Myers 2009a; Sprouse 2007; Phillips 2009).

With this in mind, we trust that several key points are evident to our readers: We are not arguing that informal methods of data collection are inherently

problematic and less trustworthy. We are not arguing that using the formal methods is the only viable means of establishing reliability and validity of syntactic arguments and that all linguistic studies should involve experiments. As clearly stated earlier, we believe that formal methods supplement or complement informal methods to the extent that the validity of syntactic data and analysis improves with great numbers of items and speakers, which is brought about by formal experimentation.

We further argue that incorporating gradience into our understanding of syntactic contrasts provides additional benefits and insights not obtainable from a strictly dichotomous approach. In other words, there exist some syntactic properties that only formal methods typically using gradient scales disclose. Myers (2009a:409) argues this succinctly when he states that "[T]he 'ideal speaker-listener' of Chomsky (1965:3), long a target of scorn from critics of traditional syntactic methodology, actually exists, but only after averaging across multiple speakers and sorting out random variation from the systematic." Unlike grammaticality, acceptability judgments of native speakers are essentially gradient. Understanding syntactic contrasts as a matter of degree facilitates analyzing syntactic phenomena more widely and thoroughly. In short, categorical grammaticality and continuous acceptability are complementary (not mutually exclusive) approaches to the better understanding of syntactic phenomena. While the worth of seeing syntactic contrasts in terms of categorical distinction has been widely discussed and accessed, there is still much to be discovered about the benefits of a gradient approach. It is to this crucial area of exploration that the present study contributes.

2.3 Syntactic Data in Korean

The experiments outlined here were conducted in Korean for two reasons. The first is simple: Korean is our native tongue and the native tongue in our current location, and this facilitated recruitment of a substantial number of native speaker participants. The second and foremost reason for our choice is that syntactic data in Korean have been less vetted so far than data in English. Korean has been widely and deeply researched since the early days of generative grammar. Nonetheless, studies which utilize methodologies of experimental syntax on a comprehensive scale are quite rare, and there are only a handful of previous studies in which a particular

syntactic phenomenon has been fragmentarily tested. In other words, the present study addresses new and lesser-known data within previously explored and well understood language.

Some asymmetry in reliability of introspective judgments across languages has already been reported. Sprouse et al. (2013) test acceptability judgments on English sentences within a methodological framework similar to what the current experiment employs. They measured acceptability judgments of the sentences exemplified in *Linguistic Inquiry 2001-2010*. The convergence rate for English data reported in Sprouse et al. (2013) is high enough: 93% using the Likert scale task, 95% using the forced choice task. In fact, this is not so surprising given that when introspective judgments are compared to formal judgments gleaned from language experiments, it has often reported that the two judgments are substantially convergent in English (Cowart, 1997; Clifton et al., 2006). In our view, such a convergence between two types of judgments is one of the characteristics of English data. Similar confirmation has been established for other languages, including German (Featherston 2005a) and Chinese (Myers 2009b, 2015). Nonetheless, introspective judgments are not universally replicated with naïve native speakers. For instance, Featherston (2005b) provides an experiment-based counterargument to an intuition-based syntactic argument regarding *that*-trace effects in German.

This asymmetry might stem from the culture of data-recycling, as Myers (2009a:408-409) notes: "The general reliability of informal judgments is at best contingent empirical fact. Perhaps English judgment claims do tend to be trustworthy, given the large number of English-speaking linguists critically examining them." The data used in the English studies have been recycled for a long time from various angles, and as a result the pairwise sentences presented in these recent studies have been particularly well-vetted. The syntactic data from other languages, including Korean, have generally not undergone such a long-term screening process. Linzen and Oseki (2015) follow this line of inquiry, using almost the same method on Japanese and Hebrew as Sprouse et al. (2013). When minimal pairs are theoretically subtle their results show a lack of empirical corroboration for one third of their Japanese examples and a full half of their Hebrew pairs. They theorize that a large number of English-speaking linguists have informally and thoroughly vetted all such theoretically subtle data, and this long-term process has eliminated overly controversial data.

3. Experimental Methods

3.1 Data Source

We replicate the methodologies of data compilation employed by Sprouse et al. (2013). As a counterpart to the English data (excerpted from *Linguistic Inquiry 2001-2010*), the present study extracts data from *Studies in Generative Grammar* published by *The Korean Generative Grammar Circle*. The last 24 volumes of *Studies in Generative Grammar* (published since 1991) are freely downloadable from the official webpage of *The Korean Generative Grammar Circle* (*http://www.kggc.org*).

Gathering data began with downloading 564 articles from the 24 volumes of *Studies in Generative Grammar*. We discarded example sentences not in Korean, which left 8,656 sentences from 289 articles. A set of 'good' and 'bad' sentence (as judged by the authors) minimal pairs were assembled from the example sentences in each article. Sentences linguists had marked single question mark ('?') were not collected, since this diacritic might indicate that the author is not fully confident in their judgment. The sentences used for the current experiment where either unmarked (i.e., 'good') or marked with the diacritics '??', '?*', '*?', and '*' (i.e., 'bad'). This selection left us 1,554 sentence pairs, from which we randomly sampled 287 pairs to serve as our experimental set. For more information about our method of data compilation, see Song et al. (2017).

3.2 Tasks

We assigned participants three different experimental tasks in order to compare different views of syntactic contrasts; the two-alternative forced choice (abbreviated as FC), the Likert scale task (abbreviated as LS), and the binary yes/no task (abbreviated as Y/N).[3]

In the FC task, the two sentences in a minimal pair are presented to the

3 There is a competing task for acceptability judgment testing, viz. the magnitude estimation task. This task was not administrated in the current experiment following the previous studies which argue that the task is not of great use (*inter alia*, Sprouse 2008; Myers 2009b; Weskott and Fanselow 2011; Fukuda et al. 2012). Granting that the magnitude estimation task is informative, it would overlap with the LS task, since the data points we collected were all Z-transformed (i.e., converting discrete data into continuous data).

participant side by side. The order of presentation is randomly shuffled to avoid ordering and fatigue effects. Participants were forced to select one of the two sentences as ‘acceptable’. Sprouse et al. (2013) advocate this task as a best fit for binary informal judgments. Syntacticians construct the pairwise sentences that illustrate putative contrasts for a particular grammatical factor in question simultaneously, not individually.

In the LS task, each sentence in a minimal pair is presented separately and the order of presentation is random for each participant. Participants were solicited to rate the acceptability of each sentence using the five-point scale. The scale provides a five-way distinction of acceptability: R1 as ‘least acceptable’, R5 as ‘most acceptable’, and the midpoint R3 as ‘so-so’. This task facilitates expressing various shades of acceptability (i.e., gradience) rather than simply recording attitude for or against a specific stimulus (i.e., a dichotomous decision). We analyzed the five-point responses using Z-scoring, a statistical technique of standardizing data on one scale. The Z-transformation was applied to each participant in the present analysis.

In the Y/N task, stimulus sentences are presented in the same manner as in the LS task, but, similarly to the FC task, participants were asked to identify whether they viewed a given sentence as ‘acceptable.’ Participants selected either 1 (‘yes’) or 0 (‘no’). There was no number representing an intermediate option. This task is advocated by Myers (2009b) on the basis that the Y/N task is quite simple and also capable of partially capturing gradience in acceptability via the proportion of ‘yes’ and ‘no’ responses assigned to a given sentence. The Y/N task is a hybrid of the FC task and the LS task involving the categorical property of introspective intuitions while enabling us to identify gradience in acceptability judgments.

3.3 Experiments

Each stimulus set consists of a control set with 44 sentences and a pretest set with six sentences. The use of stimuli requiring uncontroversial answers can provide a measure of participant concentration on the task to protect against random answers. Pilot study responses guided the selection of control sentences. From the piloted sentences, 20 grammatical and 24 ungrammatical sentences were chosen from those with the least variation in acceptability ratings.[4] Before the main acceptability

judgement task was administered, the participants had a training session to familiarize them with the experimental tasks. The practice sentences were given to the participants in the same order to make sure that they had a similar understanding of acceptability ratings. Each participant responded to a total of 290 sentences: 6 pretest sentences, 240 (120×2) main sentences, and 44 control sentences. The pretest set was counterbalanced for acceptability as was the control set so that there were equal numbers of grammatical and ungrammatical sentences, which minimizes the effect of yes/no response bias.

The experimental environment for the current study was the open-source toolkit, OpenSesame (Mathôt et al., 2012) with the PsychoPy libraries (Peirce 2007). In contrast to the pen-and-paper or online testing (i.e., methods used by comparable studies), this methodology allows us to reliably capture response times which can further inform our analyses. The experimental environment of the present study was automatically built up in order for other researchers to facilitate reusing the programming script.

732 participants were recruited at three different colleges. 203 naïve native speakers participated in the FC task, 201 speakers participated in the Y/N task, and 328 speakers participated in the LS task. Basic instructions were provided prior to each participant beginning the task. During the pretest session in which unambiguously (un)acceptable sentences were presented, administrators verified that each stimulus item was responded to correctly. Participants who responded incorrectly received additional training.[5]

3.4 Two Types of Judgments

For the remainder of this article, informal and formal judgments are differently represented for ease of explication. Informal judgments (made by the linguists who originally produce the contrasted items) are in small caps, as in BAD and GOOD. Formal judgments created via experiments are italicized, as in *bad* and *good*.

---

4 For the unpaired control sentences in the FC task, in order to present sentences in pairs, we constructed their corresponding sentences. Out of 20 grammatical and 24 ungrammatical sentences, we randomly selected 10 sentences each.

5 Demographic information was gathered for each participant, including gender, age, region of dialects, major, and background in linguistic studies (i.e., whether or not he or she has taken a linguistics-related course at the university level). No other information was collected, and all information about participants was completely anonymized.

4. Analysis

Acceptability judgment testing allows us to measure qualitative linguistic factors in a quantitative way. The present study uses a similar calculation of convergence to that used in Sprouse et al. (2013). However, our methodology diverges from Sprouse et al. (2013) in one important aspect. They emphasized whether formal judgments can perceive syntactic properties which are expressed contrastively in two pairwise sentences along the same bipartite manner as informal judgments. Therefore, their research is only concerned with the directionality of difference − if a good sentence rating is higher than the rating for its corresponding bad sentence. In contrast, the present work is concerned with the magnitude of difference between syntactically contrasted items (i.e., to what extent the ratings of these two sentences are different) not simply the directionality of difference between them. Our analysis takes a position that accounting for gradience is more informative when it comes to syntactic contrasts. Understanding syntactic contrasts as a matter of degree facilitates assessing latent patterns of correspondence between informal and formal judgments (Myers, 2009a), establishing the strength of the constraints that provoke a contrast (Lakoff, 1977; Sorace and Keller, 2005), and identifying variation in judgments across speakers (Wasow and Arnold, 2005). Note also that dichotomous distinctions can be from gradient acceptability judgments, as ordinal variables can be converted into discrete variables (e.g., 'no' goes to 0 and 'yes' goes to 1), but not vice versa.

A gradience-based approach has the potential to represent our mental behavior as a spectrum, accurately reflecting the fact that acceptability judgments exist along a continuum from 'absolutely bad' to 'absolutely good'. As shown in Figure 1, acceptability judgments can be visualized as lightness in a spectrum: The lighter, the more acceptable.

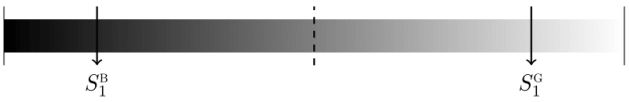


Figure 1. Spectrum of acceptability

In Figure 1 and subsequent figures, the solid lines on each end stand for the extreme points of acceptability judgments, and the dashed line in the middle stands for an ideal boundary between 'acceptable' and 'unacceptable'.[6] The arrow on the

spectrum indicates the representative value of the entire formal judgments for a particular sentence. S_i^B and S_i^G below (or over) the arrows indicate a sentence informally judged as BAD and the corresponding GOOD sentence in a minimal pair. Recall that GOOD and BAD refer to the informal judgments. In Figure 1, S_1^B and S_1^G are located on the negative side (left to the middle) and on the positive side (right to the middle), respectively. Thus, Figure 1 shows that the informal judgment fully converges with the experimentally obtained acceptability judgments.

From the perspective of both dichotomy and gradience, the following subsections address the various shades of syntactic contrasts in the visualization of acceptability spectrum.

4.1 Convergence: Directionality of Contrasted Items

Sprouse et al. (2013) compare informal and formal acceptability judgments using pairwise sentences adapted from *Linguistic Inquiry 2001-2010* to determine the extent of convergence between them. They make ample use of different experimental tasks and statistical measures for enhancing reliability and validity of their comparisons, but, as aforementioned, largely focus on directionality of contrasts. In other words, they calculate rate of convergence exclusively by whether participants discern the difference between syntactically contrasted items in the same direction as the linguists who provided the informal judgments.

4.1.1 Model

The directionality of contrasted items can be represented using an acceptability spectrum like that shown in Figure 2. As noted before, the locations of S_1^B and S_1^G in Figure 2 stand for the representative values of a BAD sentence and the corresponding GOOD sentence, respectively. Along the spectrum from the extremely

6 This spectrum translates acceptability data into a model reflecting probability. Probabilities are importantly used in the gradience model of frequency effects (Bod et al. 2003; Bresnan et al. 2007; Bresnan and Hay, 2008) because the continuous values from 0 to 1 are well suited to capturing the notion of gradience as they lie in a continuum between 0 (reflecting impossibility) and 1 (reflecting certainty).

bad and the extrememly *good*, S_1^G is situated on the right-hand side of S_1^B. Thus, these two items are on the direction predicted by the informal judgment. In contrast, S_2^G lies on the left-hand side of S_2^B, which means that the formal judgment on the contrasted items are on the direction opposite to the informal judgment.

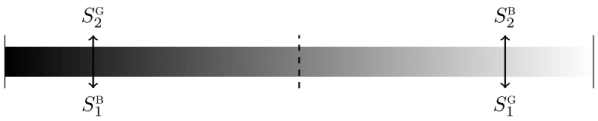


Figure 2. Directionality of contrasted items

Replicating the fundamentals of the methodological framework provided in Sprouse et al. (2013), the present study also investigates the degree of convergence between the two judgment types, and then computes the overall rate of that convergence. Our methods for measuring convergence differ from Sprouse et al. in three key ways. First, the Y/N task was added instead of the magnitude estimation task for the reasons provided in Section 3.2. Second, a five-point scale was used for the LS task because one of our pilot studies revealed that the seven-point scale used by Sprouse et al. was too demanding for our participants, and lead to score biases. Third, only one statistical measurement was carried out per experimental task. We chose the most suitable and well-attested inferential statistics for each task. For the FC and Y/N tasks, following Myers (2009b), we employed Fisher Yate's Exact test (hereafter, FYE) based on 2×2 contingency tables.[7] For the LS task, we carried out a paired T-test using Z-transformed responses (i.e., quasi-continuous scale).

Since the current comparison measures directionality between two contrasted items, we carried out only one-tailed tests for all the tasks (i.e., directional hypotheses, such as H_0: $\mu_B \geq \mu_G$, H_1: $\mu_B < \mu_G$). The informal and formal acceptability judgments converge with each other if and only if the null hypothesis is rejected ($\alpha = .05$).[8] In other words, the rejection of H_0 means that the contrasted

7 Myers (2009b) recommends the McNemar test for analyzing the Y/N task-based result. Since the McNemar test tallies only the mismatches in the 2×2 contingency table, the test is not suitable for the current comparison. For instance, {YY: 100, YN: 5, NY: 3, NN: 100} does not convey the same meaning with {YY: 20, YN: 5, NY: 3, NN: 20} in the current work, but the McNemar test treats them equally.

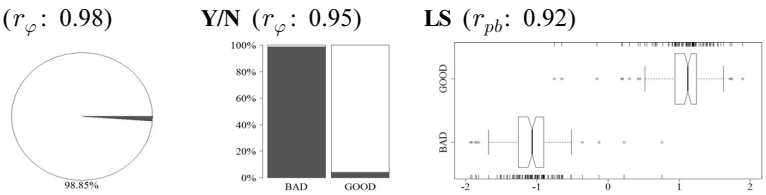
8 The significance codes used in this article are as follows: $p < 0.001$ '***', $0.001 < p < 0.01$ '**',

items lie on the acceptability spectrum in the predicted direction. The rate of convergence is measured by performing a one-tailed T-test repeatedly for each sentence pair and then measuring how many of those pairs reject the null hypothesis.

4.1.2 Examples

A typical example of perfect convergence between formal and informal acceptability judgments is presented in (1).[9] The sentence pair reveals that contrary to quantificational expressions (e.g., *ta* 'all' in (1b)), negative polarity items in Korean (e.g., *amwu-kes-to*, 'anything' in (1a)) cannot be licensed by positive predicates. The statistical results of the formal judgments on the contrasted items provided in (1) are presented in Figure 3.

(1) a. **Sue-nun nolay-wa kongpwu-lul amwu-kes-to cal-ha-n-ta.*
Sue-nom singing-and studying-acc anything well-do-pres-decl
b. *Sue-nun nolay-wa kongpwu-lul ta cal-ha-n-ta.*
Sue-nom singing-and studying-acc all well-do-pres-decl
'Sue does both singing and studying well.' (pair #79)



$p \approx$ 5.23e-48 (***) $p \approx$ 4.05e-39 (***) $p \approx$ 1.64e-59 (***), df = 100
predicted predicted predicted

Figure 3. Experimental result of (1)

The correlation coefficients are 0.98, 0.95, and 0.92 in turn, which mean that the informal and formal judgments are almost perfectly congruent to each other. The three plots represent the distributional properties of the responses. As seen in the pie plot, more

---

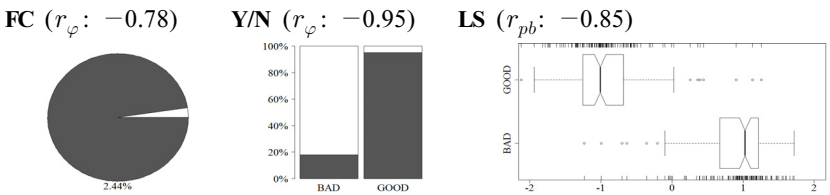
$0.01 < p < 0.05$ '*', $0.05 < p$ ' '.

9 The present study does not intend to point out the flaws in acceptability judgments provided by Korean linguists. As such, we want to be cautious to avoid any misunderstandings as to the purpose and scope of the current study when we provide the source of each pair in this article. For this reason, we decided to anonymize the source of each sentence pair using pair numbers.

than 98% of participants in the FC task agreed with the acceptability direction between (1a-b). The stacked bar plot represents how many responses of "yes" and "no" were collected from the participants in the Y/N task. In this and subsequent plots, the "yes" responses are represented as light and the "no" responses as dark. The dark section in the left bar representing the answers to the BAD sentence (1a) accounts for more than 95%. Likewise, the light section in the right bar representing the answers to the GOOD sentence (1b) is such larger than the dark section in the same bar. The box plot on the rightmost side represents the distribution of Z-scores obtained from the LS task. The box for (1a) is located on the right side of that for (1b). Additionally, all the inferential tests significantly take the alternative hypothesis ($\mu_B < \mu_G$) as '***'. In sum, we can say that the contrasted items are in the predicted direction.

Naïve native speakers' formal judgments are sometimes counter to those of linguists. In other words, the naïve native speakers judged some sentences that linguists considered GOOD as *bad*, and vice versa. This pattern is shown in (2). According to the original author (2a) sounds less acceptable than (2b) as there is no sentence-internal position for a null argument linked to the topic.

(2) a. ?**Chelswu-ka chayk Thoci-lul ilk-ess-e.*
Chelswu-nom book Thoci-acc read-pst-decl
b. *Chayk, Chelswu-ka Thoci-lul ilk-ess-e.*
book Chelswu-nom Thoci-acc read-pst-decl
'As for a book, Chelswu read *Thoci*.' (pair #867)



p = 1.0 opposite | p = 1.0 opposite | p = 1.0, df = 121 opposite

Figure 4. Experimental result of (2)

Contrary to the author's position, (2a) was found to sound fairly acceptable to most naïve native speakers. In addition, (2b) was found to sound less acceptable to naïve native

speakers as shown in the plots. The correlation coefficients are negative, which means that the two types of judgments are opposite to each other. Hence, this pair conclusively fails all the inferential tests regardless of experimental tasks. Thus, for this pair, the two types of judgement results diverge very significantly.

4.1.3 Result

Using the methods presented thus far, directionality was determined for all 287 contrasted items. We calculated the convergence rate in the experimental result as shown in Table 1.

Table 1. Rate of convergence

Task	Statistics	Raw	B-H
FC	FYE	90.94%	90.94%
Y/N	FYE	82.93%	82.93%
LS	T-test	88.85%	88.85%

Note that "α = .05" means that there is a 5% chance of type I errors (rejecting the null hypothesis even though the null hypothesis is true in reality). Because the convergence rate for each task comes from the sum of 287 tests (of which the significance level is 95%), the probability of wrongly rejecting the null hypothesis (i.e., a type I error) may or may not sum over the tests that we conducted. In order to control the false discovery rate, the Benjamini-Hochberg procedure (B-H in Table 1) was taken for our comparisons (Benjamini and Hochberg 1995). The convergence rates without and with controlling the false discovery rate are presented in the third and fourth columns respectively.

Table 1 reveals that the convergence rates of Korean syntactic data in the current work are about 4% lower than those of English syntactic data in Sprouse et al. (2013) when it comes to the FC and LS tasks (i.e., 91% and 89% in Korean vs. 95% and 93% in English). This difference might be accounted for by the fact that the Korean syntactic data have been less vetted than the English syntactic data as discussed in Section 2.3. Furthermore, such a difference can receive an explanation from divergent views about syntactic contrasts. While the current study is concerned with both directionality and magnitude of difference of syntactic contrasts, the primary concern of Sprouse et al. hinges on detecting the directionality of syntactic contrasts. As such, even with a

numerical rating task such as the LS task which can measure the magnitude of difference (i.e., gradience of syntactic contrasts), they test only directionality of contrasted items. Given that it is likely to be easier for two contrasted items to meet simple directional difference than to meet sizable difference between them, a relatively higher rate of Sprouse et al. can be accounted for along these lines. That is, the directionality of difference of syntactic contrasts is a necessary condition for magnitude of difference of the contrasts. We take a closer look at this issue again in Section 5.

4.2 Latent Patterns: Goodness of Contrasted Items

Myers (2009a:409) argues that "the quantitative analysis of many data-points necessarily provides a more reliable picture of latent patterns than the qualitative analysis of a few isolated data-points." The current subsection explores those latent patterns in syntactic contrasts, which can be identified when we test acceptability judgments on a comprehensive scale focusing on goodness of sentences.

In a dichotomous view of syntactic contrast, directionality in the relation between two sentences is crucial: The BAD sentence sounds relatively worse than the GOOD sentence. We expand on this approach by considering magnitude of acceptability in addition to directionality, allowing for an assessment of the absolute goodness of a sentence: A GOOD sentence must by definition sound acceptable to most native speakers, and likewise a BAD sentence must by definition sound unacceptable.

4.2.1 Model

Different patterns of syntactic contrasts can be represented on an acceptability spectrum like that in Figure 5.

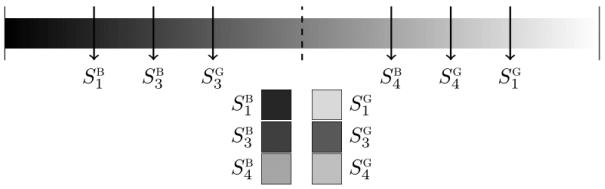


Figure 5. Goodness of contrasted items

In Figure 5, all sentence pairs including S_1, S_3, and S_4 are in the predicted direction with BAD-version sentences on the left side of the correponding GOOD-version sentence. Nonetheless, the positioning patterns of the pairs are different from each other. Both S_3^B and S_3^G are on the negative side (i.e., on the left hand of the middle represented as a dashed line), and both S_4^B and S_4^G are on the positive side. The lightness of each item is shown below the spectrum in Figure 5. On the BAD-version side, S_3^B looks almost as dark as S_1^B, but S_4^B looks lighter than the other two. On the GOOD-version side, S_1^G and S_4^G are close to each other but different from S_3^G in terms of lightness. The six items presented in Figure 5 can be represented as $S_1^B < S_3^B < S_3^G < S_4^B < S_4^G < S_1^G$, with S_3^G evaluated as less acceptable than S_4^B corresponding to the lightness boxes in Figure 5. We can see from these measures that the GOOD sentence in one minimal pair sounds worse than the BAD sentence in the other minimal pair.

It is highly unlikely for linguists to unfold their syntactic argumentation by presenting a contrast within a single minimal pair. Rather, in order to effectively present and stregthen their argumentation, a multiple number of contrasts within multiple pairs examining the same syntactic constraint are likely to be presented. As such, the issue of consistency of good (bad)ness (i.e., prototypicality) of two contrasted items arises in any paper with multiple pairs provided. We argue that (i) it is important to maintain the relative goodness of a GOOD-version of the pair not only as compared to its BAD-version counterpart but also to all BAD-version sentences of multiple pairs presented throughout the paper and that (ii) more importantly, all GOOD-version sentences of the pairs provided should be evaluated higher than 0 (the centering value on the scale). That is, we believe that absolute goodness or badness of two contrasted items should be essential to syntactic contrasts; any given GOOD-version sentence of a pair should sound better than its BAD-version counterpart (satisfying directionality of contrasts) but also sound good enough to be evaluated higher than the centering point 0 (satisfying magnitude of difference of contrasts). The size of difference is an integral part of syntactic contrasts and once the magnitude of difference detected, directionality of contrasts follows by default. This issue of the consistency of good/badness of two items is not restricted to informal methods typically measuring the directionality of syntactic contrasts. Any methods or approaches solely focusing on detecting the directionality of contrasts cannot

be free from this problem. In the following subsection, we illustrate this point with two relevant examples.

4.2.2 Examples

The minimal pair in (3) is an instance of S_3 in Figure 5. According to the author, the causative particle in Korean can co-occur with the passive particle if and only if a separator *key* is attached to the causative particle, as seen in (3b). However, the author's informal judgment is different from the formal judgments of naïve native speakers, as shown in Figure 6.

(3) a. **Ai-ka* *John-ey uyhayse* *ppang-ul* *mek-i-hi-ess-ta.*
child-nom John-by bread-acc eat-cause-pass-pst-decl
b. *Ai-ka* *John-ey uyhayse* *ppang-ul* *mek-i-key* *toy-ess-ta.*
child-nom John-by bread-acc eat-cause-key pass-pst-decl
'The child was caused to eat the bread by John.' (pair #342)

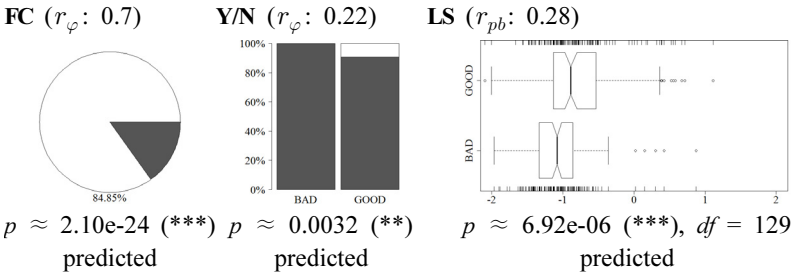


Figure 6. Experimental result of (3)

Notice that the pie plot shows that about 85% of the participants in the FC task answered that (3b) was more acceptable than (3a). In contrast, if the two sentences were separately presented to the participants, (3b) was mostly evaluated as *bad*, as shown in the second bar of the stacked bar plot and the upper box of the box plot. Thus, the Y/N and LS tasks yield much smaller correlation coefficients (0.22 and 0.28) than the FC task (0.7). The dark section of the second bar is smaller than that of the first bar in the stacked bar plot. Likewise, the upper box lies more towards

the positive side than the other box in the box plot. These results indicate that the direction between syntactic contrasts can be ascertained using the current model. In this case, all the hypothesis tests (which correspond to directionality) say that the two types of judgments are significantly similar in direction.

The minimal pair presented in (4) is an instance of S_4 in Figure 5. Korean employs two resultative markers such as *-tolok* and *-key*, and the author claims that *-tolok* expresses an expected result whereas *-key* expresses an entailed result. The author further argues that the latter is seldom used for conveying a meaning of 'clausal resultative'. Our results, represented in Figure 7, indicate that formal judgments compete with the linguist's informal judgment.

(4) a. ??*Ku-nun Mary-lul chim-i malu-key chingchanha-yess-ta.*
he-nom Mary-acc saliva-nom dry.out-key praise-pst-decl
b. *Ku-nun Mary-lul chim-i malu-tolok chingchanha-yess-ta.*
he-nom Mary-acc saliva-nom dry.out-tolok praise-pst-decl
'(lit) He praised Mary (his) saliva dried out.'
'He spoke in the highest terms of Mary.' (pair #508)

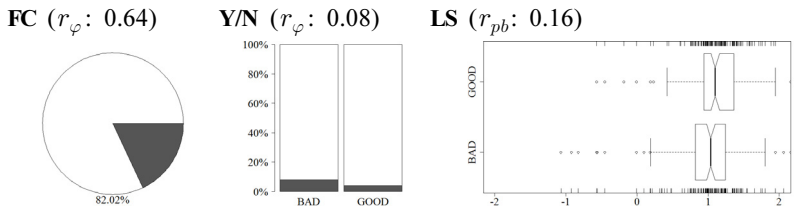


Figure 7. Experimental result of (4)

The pie plot shows that if the two sentences were provided simultaneously, more than 80% of participants agree with the informal judgment, indicating that the predicted direction between the contrasted items is accurate. However, the stacked bar plot in the middle indicates that naïve native speakers judged both sentences as fairly *good*, though the BAD sentence was judged slightly worse than the GOOD sentence. Likewise, both boxes in the right plot go toward the positive side, which means that native speakers mostly accept both sentences. Thus, the correlation

coefficients indicate that the two types of acceptability judgments are not correlated to each other, for both binary decision and gradient scales. This is a clear example of different experimental tasks producing different results with respect to hypothesis testing (Table 1). The pair given in (4) significantly passes the hypothesis tests with the FC and LS task, and yet fails the test with the Y/N task.

It is worth comparing the formal acceptability judgments for (3b) and (4a). They are reported as GOOD and BAD within each minimal pair but evaluated as being *bad* and *good* individually. Along with its BAD counterpart (3a), the box for (3b) is located on the left hand side of the centering point (Z-transformed score, 0). Along with its GOOD counterpart (4b), the box for (4a) is located on the right hand side of the centering point. Results such as these highlight the importance of determining which type of information about syntactic contrasts, between directionality and magnitude of difference of contrasted items, is more useful for understanding syntactic contrasts. In the next subsection we outline some key evidence that the magnitude should be considered at least as important as directionality.

4.2.3 Result

Naïve native speaker judgments sometimes diverge from the judgements of linguists. Analyzing specific patterns of correspondence in informal and formal acceptability judgments creates a better understanding of syntax data. To this end, four patterns were identified as exemplified in (1-4). These four patterns can be represented using the 2×2 contingency table in Figure 8. In Figure 8, symbols '+/−' indicate results of judgments of the majority of native speakers on a given sentence. If the majority of native speakers judge a given sentence as *bad*, that particular sentence is presented with a minus symbol. If the opposite holds, it is presented with a plus symbol. In four plus-minus combinations, the first symbol represents judgments on a BAD sentence and the second symbol represents judgments on a GOOD sentence. For instance, Pattern # 1 (−/+) means that most of native speakers judge the BAD as *bad* and the GOOD sentence as *good* (i.e., the most preferable correspondence between two types of acceptability judgments).

Pattern #1 (−/+)

	bad	*good*
BAD	✓	
GOOD		✓

Pattern #2 (+/−)

	bad	*good*
BAD		✓
GOOD	✓	

Pattern #3 (−/−)

	Bad	*good*
BAD	✓	
GOOD	✓	

Pattern #4 (+/+)

	bad	*good*
BAD		✓
GOOD		✓

Figure 8. Four patterns

The check mark in Figure 8 notes that the formal judgments cluster toward either *good* or *bad* for each sentence pair. In the FC and Y/N tasks, the threshold for cluster identification of judgments is 50% of the total judgments for each sentence. For the LS task, if the median of the Z-scores on a particular sentence is over 0, the sentence is classified as clustering toward *good*. The shaded cells indicate discrepancies in the two types of acceptability judgments in Figure 8.

Table 2 shows the proportion for each pattern. Note that the last two patterns cannot be captured with the FC task.

Table 2. Proportion of patterns

Task	**#1 (−/+)**	**#2 (+/−)**	**#3 (−/−)**	**#4 (+/+)**
FC	91.99%	8.01%	N/A	N/A
Y/N	58.89%	0.7%	26.13%	14.29%
LS	59.23%	0.7%	24.74%	15.33%

Directionality of difference can be easily determined by comparing the centering values of two distributions of data on a significant level ($p < .05$). In other words, as long as a GOOD-version sentence is rated higher than its corresponding BAD-version sentence, the directionality of contrasts is met. Therefore, for the directionality-based analysis, the size of acceptability judgment clusters cannot be an issue. For the FC task, the proportion of Pattern #1 is almost the same as the rate of convergence (91% and 92%). On the other hand, if we focus on the magnitude of difference using the 2×2 contingency table, the proportion of the syntactic contrasts on which the two types of acceptability judgments converge is at best 60% as shown in the second column of Table 2. Note that Y/N and LS tasks produce almost the

same results from Pattern #1 to #4. These analyses clearly demonstrate that directionality (captured only by the relation between two items in a minimal pair) and goodness (pertaining to whether an individual sentence does indeed sound acceptable) do not necessarily tap into the same property of syntactic contrasts.

4.3 Strength of Constraint: Repulsion of Contrasted Items

Lakoff (1977) argues against the use of either-or decisions in syntactic analysis, pointing out that many sentences can be declared as neither purely 'out' nor purely 'in'. As an alternative, Lakoff posits a delicately shaded and rather interminable hierarchy of acceptability. More recently, Sorace and Keller (2005) classify linguistic constraints into hard constraints and soft constraints depending on whether the violation triggers strong unacceptability (and whether the constraint is subject to context effects and developmental optionality). In a similar vein, Bauer (2008:95) argues that not all contrasts are equal as evidenced by the German example provided in Section 2.1 (/s/ and /t/ vs. /ç/ and /x/): "[S]ome contrasts are much more thoroughly integrated into the language system than others." These studies (and several others) frame acceptability of sentences from a gradient perspective with an emphasis on the strength of contributing constraints. This approach is further explored in the following subsections.

4.3.1 Model

If we frame acceptability as a matter of strength, strength in the relation between syntactic contrasts can be calibrated as repulsion force. The repulsion between contrasted items can be modeled as in Figure 8.

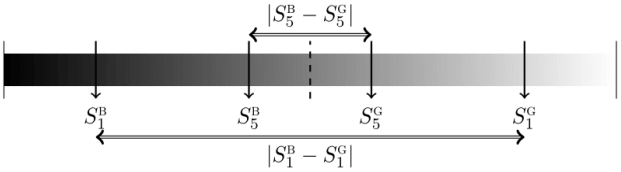


Figure 8. Repulsion of contrasted items

S_5^B and S_5^G in Figure 8 are not only in the predicted direction but also on the left and right hands of the middle respectively. Thus, the syntactic contrast is good enough in terms of goodness (Pattern #1) as well as directionality (convergence). Nevertheless, the S_5 does

not share the same status as the S_1 in terms of the distance between two items: The repulsion force between S_1^B and S_1^G is stronger than that between S_5^B and S_5^G. Simply put, the farther two pairwise sentences are from each other on the acceptability spectrum, the stronger the constraint related to the contrast is.

4.3.2 Example

The minimal pair provided in (5) is concerned with which case marker can be attached to the main verb affixed with the morpheme *-ci-* in the long negation form in Korean (realized with the negative auxiliary *anh-*): If the main predicate dependent upon the auxiliary is intransitive (e.g., *yeyppu-* 'pretty' in (5)), the nominative marker is preferred to the accusative marker.

(5) a. *??Yenghuy-ka yeyppu-ci-lul anh-ta*
Younghee-nom pretty-conn-acc not-decl
b. *Yenghuy-ka yeyppu-ci-ka anh-ta*
Younghee-nom pretty-conn-nom not-decl
'Younghee is not pretty.' (pair #1302)

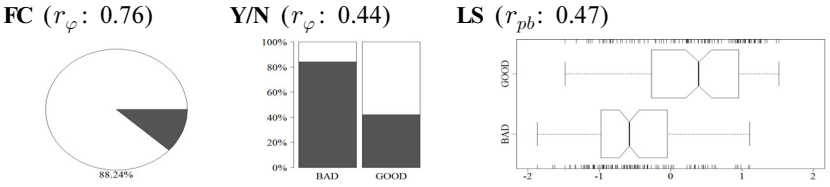


Figure 9. Experimental result of (5)

The sentence pair significantly passes all the tests of detecting direction irrespective of experimental tasks, and the goodness across 'bad' and 'good' is identified as Pattern #1. The sentence pair (5) behaves similarly to the previous pair (1) as such, but comparing the plots for (1) and (5) reveals some differences. Although the pie plot in Figure 9 indicates that more than 88% of the participants in the FC task agree that (5a) sounds less acceptable than (5b), the stacked bar plot shows that the gap in dichotomous judgments

on (5a-b) is much smaller than that on (1a-b). The distances between the two centering points in the box plots for (1) and (5) are 2.189 and 0.9721, respectively.

Furthermore, the difference between them is numerically evident in the correlation coefficients of the Y/N and LS tasks. The coefficients (0.44 and 0.47) imply that the two types of judgments do not have an equally strong correlation when contrasted items are judged separately. Recall that the correlation coefficients between the two types of judgments for sentence pair (1) are 0.95 and 0.92. In Lakoff (1977)'s framing, the contrast exemplified in (1) occupies a higher position in the hierarchy than the contrast exemplified in (5). Using the terms of Sorace and Keller (2005), the constraint reflected in (1) seems to be harder than that reflected in (5).

4.4 Variation across Speakers: Intensity of Contrasted Items

There exists variation in linguistic phenomena, and the same goes for acceptability judgments. Unlike linguists in other fields, many theoretical syntacticians tend to have traditionally overlooked variation. They postulate that linguistic behavior is categorical and therefore syntactic studies should treat variation as irrelevant noise. However, Wasow and Arnold (2005) present variation in acceptability judgments across speakers as evidence that informal judgments made by a single native speaker may not always be accurate. Wasow and Arnold assert that the degree of variability among speakers is such that speakers of the same language might disagree on whether or not an expression is ill-formed. The following subsections address why examining variation across speakers is valuable in the analysis syntactic contrasts and how that variation can be detected using a gradience-based analysis of acceptability judgments.

4.4.1 Model

In Figure 11, S_1^B and S_6^B and their counterparts S_1^G and S_6^G share the same representative value of acceptability judgments. Thus, the syntactic contrasts S_1 and S_6 are both in the predicted direction and follow the same pattern of goodness. However, the potential variability in acceptability judgments between the two is sufficiently revealed by the analyses presented thus far. The bars over S_1^B and under

S_6^B indicate the difference in variability between these two measures: The acceptability judgments on S_1^B are densely gathered, while those on S_6^B are scattered. In other words, S_1 and S_6 differ in intensity of acceptability judgments.

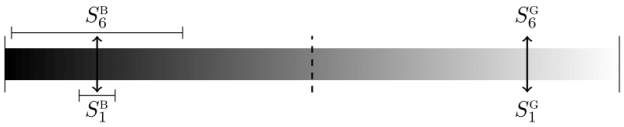


Figure 11. Intensity of contrasted items

Such variation often occurs across dialects in a language, but can be seen even in a homogeneous language community (Honeybone, 2011). Prior literature has classified variation in acceptability judgments into two types, viz. intraspeaker variation and interspeaker variation. Most variation in linguistic behavior has been categorized as intraspeaker variation due to non-grammatical factors, including superficial analogy, discourse context, lexical semantics, etc. In order to focus on the underlying patterns behind speaker's linguistic performance, such an intraspeaker variation has been often ignored in syntactic studies. The causes of interspeaker variation are distinct from those outlined above. This type of variation arises when speakers acquire different syntactic rules for the same linguistic phenomenon (Han et al., 2007; Han et al., 2016).

Approaching syntactic contrasts from a gradience perspective has utility for identifying both intraspeaker and interspeaker variation. Formal experiments representing acceptability judgments within spectrum facilitate distinguishing grammatical factors from non-grammatical ones in syntactic data. As Myers (2009a:415) argues, "formal judgment experiments do a better job at separating the wheat from the chaff." As aforementioned, we fully admit that a variety of non-grammatical factors can manifest themselves in acceptability judgments and they must be accounted for separately when establishing syntactic rules. The problem is that in many cases it is not easy to identify whether interpretation of an expression is influenced by a non-grammatical factor and, if so, which non-grammatical factor is at fault. If one of the pairwise sentences (or both) is affected by non-grammatical factors, the syntactic data and any assertions drawn from it, will be incorrect. These biasing factors should be minimized as much as possible in order to ensure the

syntactic data we use in literature is trustworthy. At the same time, preserving true interspeaker variation is crucial because these data provide important insights into how syntax is acquired and how syntax operates. If interspeaker variation exists, viewing the syntactic contrasts in terms of gradience helps to reveal it and makes it available for further studies.

4.4.2 Example

The minimal pair in (6) exemplifies variation across speakers. The unacceptability of (6a) is accounted for by arguing the Korean demonstrative *ku* cannot be used to express uniqueness as 'the' does in English.

(6) a. **Minsu-kaI tongari-uy ku hoycang-i-ta.*
Minsu-nom this club-gen the president-cop-decl
b. *Minsu-ka I tongari-uy hoycang-i-ta.*
Minsu-nom this club-gen president-cop-decl
'Minsu is the president of this club.' (pair #1253)

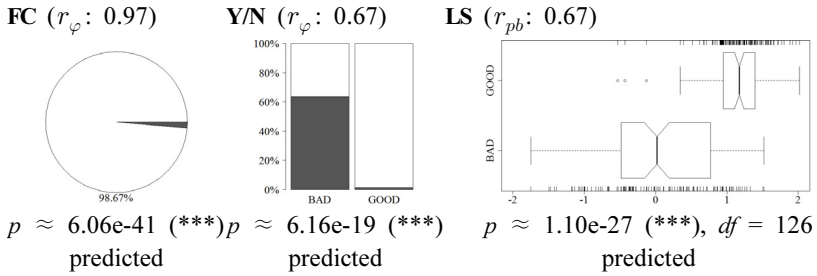


Figure 12. Experimental result of (6)

The participants in the current experiments mostly agreed with the linguists' judgments on the GOOD sentence provided in (6b), but the judgments on the BAD sentence (6a) are mixed as visualized in the stacked bar plot in the middle and the box plot on the right side. Note that the FC-based results of (6) are the same as those of (1): The light portion in the pie plot accounts for almost 100%. In terms of directionality of contrasted items, the sentence pair (6) significantly passes all the

inferential tests, and the pattern of goodness is identified as #1 irrespective of the experimental tasks. However, the light portion in the left bar of the stacked bar plot accounts for about 40%, which means that there exists some variation in acceptability judgments on the BAD sentence. The variation is more clearly seen in the box plot. The BAD sentence (6a) comprises the widest range of acceptability judgments among the sentences exemplified hitherto.

As presented above, the main difference between the two sentence pairs (1) and (6) is in the intensity of acceptability judgments on the BAD sentences. For ease of comparison, the acceptability judgments of the sentences can be visualized using density plots as shown in Figure 12. Note that out of the three tasks in the current experiment only the LS task which allows the participants to use intermediate values enables us to see the intensity of acceptability judgments.

(1) Var(BAD): 0.06, Var(GOOD): 0.0559 (6) Var(BAD): 0.6161, Var(GOOD): 0.0998

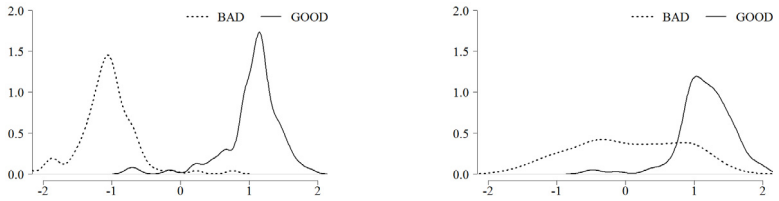


Figure 13. Density plots (variance of BAD and GOOD)

In the left plot of Figure 13, both lines form a steep curve, which indicates that the acceptability judgments on the sentences are narrowly distributed. In the right plot, the dashed line for (6a) forms a gentle slope and the solid line for (6b) exhibits a mountain-shape. This indicates that naïve native speakers do not share a common acceptability judgment on (6a). The different behavior exhibited by (6a) can be numerically determined using the variances (i.e., 0.6161): The variance of the Z-transformed responses on (6a) is much bigger than the variances on the others. Hence, these two syntactic contrasts are not on a par with each other, although they are similar in terms of the direction analysis and the pattern analysis. This level of differentiation is only possible when acceptability judgments are measured with a methodology that captures gradience (e.g., using response scales).[10]

10 Ideally, variation across speakers would be measured via more sophisticated statistical skills, such as mixed effect models. Nonetheless, in our experience not all syntacticians are

Bauer (2008) argues that variation across speakers is of importance with respect to gradience in syntactic contrasts and an either-or decision is essentially incapable of explaining variation across speakers in linguistic contrasts. This argument is empirically supported by the data distribution and statistical results of the sentence pair (6). If we had cataloged only either-or categorizations of syntactic contrasts, we would not have discovered that not all native speakers judge (6a) to be acceptable. Even in cases where we chose to emphasize dichotomous distinctions in our formal experiments, using a task like the Y/N task allows for the capture of at least some variation data (Myers, 2009b). This variation can be more precisely discerned through the use of the response scales that permit intermediate values. However, if we test syntactic contrasts exclusively by comparing either one sentence sounds relatively better than another, the complexities in speakers' judgments of sentences like (6a) will remain absent from linguistic studies (though the underlying complexity may not be purely syntactic).

5. Direction, Position, Distance, and Intensity

Gries (2015) argues that the adequate criticism of a model has to include a set of specific examples with which the model cannot cope and also provide a verification that a drawback arises from how the model operates. Likewise, if we want to advocate a model, we must offer a set of specific examples which the model can handle well and other models cannot, giving an account of how the model processes those examples. The present study has posited the advantages of the gradience-based view towards syntactic contrasts over the dichotomy-based view. We conducted three formal experimental tasks on 287 sentence pairs in Korean excerpted from a syntactic journal. Amongst them, the previous section overviewed six sentence pairs which the different approaches to syntactic contrasts can or cannot

---

experienced in the use of high-end techniques of statistics. Furthermore, syntacticians do not need to use the skillful techniques even though they want to investigate the variation in syntax or eliminate the latent non-grammatical factors from the syntactic data. As Myers (2009b) emphasizes conducting formal experiments for syntactic studies, using a simpler model, is sufficient for revealing the features of interest hidden within syntactic data. Drawing a density plot as in Figure 13 and/or computing the variance of the responses is enough to capture an amount of intra/inter-speaker variation.

cope with. The previous section also posits a mental spectrum of acceptability judgments and approaches syntactic contrasts from different angles, including direction, position, distance, and intensity. The results of these analyses are summarized in Table 3.

Table 3. Summary

Ex	Pattern	FC(r_{φ})	Y/N(r_{φ})	LS(r_{pb})	Var(BAD)	Var(GOOD)
(1)	#1 (−/+)	*** (0.98)	*** (0.95)	*** (0.92)	0.06	0.0559
(2)	#2 (+/−)	(−0.78)	(−0.95)	(−0.85)	0.1731	0.1516
(3)	#3 (−/−)	*** (0.7)	** (0.22)	*** (0.28)	0.1049	0.1853
(4)	#4 (+/+)	*** (0.64)	(0.08)	*** (0.16)	0.1031	0.0908
(5)	#1 (−/+)	*** (0.76)	*** (0.44)	*** (0.47)	0.4514	0.5932
(6)	#1 (−/+)	*** (0.97)	*** (0.67)	*** (0.67)	0.6161	0.0998

The syntactic contrasts exemplified in (1) and (2) are indicative of direction of contrasted items following Sprouse et al. (2013). The second and third rows in Table 3 show that all experimental tasks are capable of identifying the directionality, although they result in different rates of convergence. This corroborates our earlier assertion that categorical grammaticality can be discerned even from formal experiments which use response scales (i.e., quantitative analysis of qualitative variables).

The syntactic contrasts exemplified in (3) and (4) pertain to position on the continuum of acceptability judgments. The sentence pairs match the predicted direction just as with (1), but the representative values for the BAD and GOOD sentence are inclined towards the same side. Since native speakers judge both sentences either *good* or *bad*, we cannot say that (3) and (4) represent the same phenomenon as (1) in spite of the likeness in directionality. Notice that (3) and (4) pass all inferential tests with the exception of the Y/N task for (4). This implies that the direction-based inferential tests are not able to capture the different patterns of acceptability judgments. If we take the absolute goodness of individual sentences into consideration, only about 60% of the formal judgments (out of 287 sentence pairs) match the corresponding informal judgments.

The syntactic contrast exemplified in (5) shows the importance of gradience in acceptability in revealing that the contrast there involves a softer constraint than the contrast exemplified in (1). The contrast exemplified in (5) is on the predicted direction as marked with '***' and like (1) is analyzed as Pattern #1. The difference between these two contrasts is a function of the distance between the representative

values of the acceptability judgments for each sentence in the minimal pair: The farther apart the contrasted items are on the continuum, the stronger the constraint is. The correlation coefficients numerically represent the strength of constraint implicated. Note also that the correlation coefficient in the Y/N task is computed using a contingency table divided in a binary way, showing once more that dichotomous distinctions are not incompatible with experimental methods.

The syntactic contrast exemplified in (6) reflects intensity of acceptability judgments. Many crucial examples used in syntactic literature are not clearly 'good' or 'bad', and native speakers' judgments of those sentences may be mixed. Such variation across speakers needs to be clarified even if only for the purpose of excluding non-grammatical factors from syntactic discussion. The LS task in which gradience in acceptability is measured performs this function. As shown in the last row of Table 3, the BAD sentence in (6) shows greater variance than the others. Although the sentence pair passes all the inferential tests and is an example of Pattern #1, the variance shown in (6a) should be accounted for in descriptions of the syntax and reveals the need for further syntactic analysis.

The results of the present study differ from those of Sprouse et al. (2013). Sprouse et al. report convergence rates of more than 90% for all experimental tasks. In contrast, the results in the current work prove quite sensitive to experimental task. There are two likely reasons for this discrepancy. First, as aforementioned, Korean syntactic data have been less vetted than English syntactic data. When data is less stable, subtle changes in experimental setting may have larger effects on experimental results (Linzen and Oseki, 2015). Second and more importantly, we consider both directionality and magnitude of syntactic contrasts in the analysis of our results, whereas Sprouse et al. largely focus on directionality. In the context on the present study, Sprouse et al. place a high value on tasks like the FC task.

Lastly, we would like to emphasize that we are not arguing that tasks like the FC task are not of great use in experimental syntax. The choice of experimental task hinges on theoretical issues at stake and the language phenomenon in question. The FC task has an advantage over the others when it comes to detecting the directionality of syntactically contrasted items as Sprouse et al. (2013) argue. Not all syntactic studies must take the gradience view. If gradience is irrelevant to the research questions, and a researcher wishes to carry out a formal experiment focusing on directionality, the FC task serves as a simple and reliable method. Recall that simplicity should be a key consideration in

experimental studies (Myers, 2009b). We believe that the results of the current experiments clearly show the irreplaceable merits of experimental tasks using response scales, Y/N and LS tasks. The Y/N and LS tasks that partially and entirely facilitate measuring gradience provide several additional merits, such as revealing latent patterns, calibrating strength of constraints, and identifying variation across speakers.

6. Conclusion

Since contrasts play a pivotal role in syntactic argumentation, a precise characterization of syntactic contrasts is crucial. There are two views of syntactic contrasts, viz. dichotomy and gradience. The former typically pertains to categorical grammaticality, an either-or decision, and directionality, whereas the latter relates to continuous acceptability, degree, and magnitude. The present study has compared the informal and formal judgments on pairwise sentences in order to examine various shades of syntactic contrasts within the context of Korean. All the available volumes of *Studies in Generative Grammar 1991-2014* were used to collect informal judgments. We then randomly selected 287 from 1,554 sentence pairs extracted from the original data source. The present study used the OpenSesame toolkit for three distinct experimental tasks. First, the FC task was carried out with 203 participants. Like traditional informal judgments, this task involves dichotomous decision making. Second, the Y/N task was carried out with 201 participants. This task requires participants to respond to each sentence in a minimal pair as either ‘acceptable’ or ‘not acceptable’. The third task, which was conducted with 328 participants, involves a five-point Likert scale ranging from ‘1’ to ‘5’. 732 participants each responded to 290 stimulus items which consisted of six pretest, 240 pairwise, and 44 control sentences. For each participant, 120 sentence pairs were randomly selected from 287 sentence pairs. In addition, the pairwise and control sentences were randomly reordered for the presentation of stimuli.

All the experimental tasks used in the current experiment allow us to measure qualitative linguistic factors in a quantitative way, but they differ in which aspect of syntactic contrasts is measured. Visually representing the acceptability spectrum, the present study has analyzed four shades of syntactic contrasts, viz. directionality (i.e., convergence between two types of judgments), latent pattern (i.e., absolute goodness of

individual sentences), strength of constraint (i.e., how far the two sentences are from each other in terms of acceptability), and variation across speakers (i.e., the intensity of acceptability judgments). The current analysis shows that the dichotomous view towards syntactic contrasts is not incompatible with a gradience view as an ordinal variable can be easily converted into a numerical variable. More crucially, the current analysis empirically justifies that the gradience view toward syntactic contrasts has several advantages over the dichotomous view; it is better equipped to provide various shades of contrasts such as strength of constraint and intensity of judgments which only methods quantifying magnitude of differences can offer. In other words, there exist some syntactic properties that only formal methods disclose.

Since acceptability judgments of native speakers are essentially gradient, understanding syntactic contrasts as a matter of degree facilitates analyzing syntactic phenomena more deeply and thoroughly. The gradience view serves as a catalyst for making syntactic studies bear more abundant fruit in the long term. Furthermore, this formal method of collecting introspective intuitions on a comprehensive scale allows for capturing many more shades of our underlying conscious awareness, a much-needed improvement from simply revealing an attitude for or against a specific stimulus. As such, conducting an experiment from the gradience view enables us to see both the direction and magnitude of acceptability judgments.

References

Bard, Ellen G., Dan Robertson, and Antonella Sorace. 1996. Magnitude Estimation of Linguistic Acceptability. *Language* 72(1): 32-68.

Bauer, Laurie. 2008. Contrast in Language and Linguistics. *Journal of English Linguistics* 36(1): 93-98.

Benjamini, Yoav, and Yosef Hochberg. 1995. Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society B* 57: 289-300.

Bod, Rens, Jennifer Hay and Stefanie Jannedy. 2003. *Probabilistic Linguistics*. Cambridge, MA: The MIT Press.

Bresnan, Joan, and Jennifer Hay. 2008. Gradient Grammar: An Effect of Animacy on the Syntax of *give* in New Zealand and American English. *Lingua* 118(2): 245-259.

Bresnan, Joan, Anna Cueni, Tatiana Nikitina, and R. Harald Baayen. 2007. Predicting the

Dative Alternation. In Gerlof Bouma, Irene Krämer, and Joost Zwarts (eds.) *Cognitive Foundations of Interpretation*, 69-94. Amsterdam: Royal Netherlands Academy of Science.

Cable, Seth and Jesse A. Harris. 2011. On the Grammatical Status of PP-pied-piping in English: Results from Sentence-rating Experiments. In Margaret Grant and Jesse A. Harris (eds.) *University of Massachusetts Occasional Papers in Linguistics: Processing Linguistic Structure, Vol. 38*, 1-22. Amherst, MA: GLSA Publications.

Chomsky, Noam. 1964. Degrees of Grammaticalness. In Jerry A. Fodor and Jerrold J. Katz (eds.) *The Structure of Language: Readings in the Philosophy of Language*, 384-389. Englewood Cliffs, NJ: Prentice-Hall.

Chomsky, Noam. 1965. *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.

Chomsky, Noam. 1986. *Knowledge of Language: Its Nature, Origin, and Use*. New York: Praeger.

Clifton Jr, Charles, Gisbert Fanselow, and Lyn Frazier. 2006. Amnestying Superiority Violations: Processing Multiple Questions. *Linguistic Inquiry* 37(1): 51-68.

Cowart, Wayne. 1997. Experimental Syntax: Applying Objective Methods to Sentence Judgments. Sage publications.

Culbertson, Jennifer, and Steven Gross. 2009. Are Linguists Better Subjects?. *The British Journal for the Philosophy of Science* 60(4): 721-736.

Culicover, Peter W. and Ray Jackendoff. 2010. Quantitative Methods Alone are Not Enough: Response to Gibson and Fedorenko. *Trends in Cognitive Sciences* 14(6): 234-235.

Devitt, Michael. 2006. Intuitions in Linguistics. *The British Journal for the Philosophy of Science* 57(3): 481-513.

Devitt, Michael. 2010. Linguistic Intuitions Revisited. *The British Journal for the Philosophy of Science* 61(4):833-865.

Dresher, B. Elan 2009. *The Contrastive Hierarchy in Phonology*. Cambridge: Cambridge University Press.

Edelman, Shimon and Morten H. Christiansen. 2003. How Seriously Should We Take Minimalist Syntax?. *Trends in Cognitive Sciences* 7(2): 60-61.

Erlewine, Michael Yoshitaka, and Hadas Kotek. 2016. A Streamlined Approach to Online Linguistic Surveys. *Natural Language & Linguistic Theory* 34(2): 481-495.

Fanselow, Gisbert, Caroline Féry, Ralf Vogel, and Matthias Schlesewsky. 2006. Gradience in Grammar. In Gisbert Fanselow (ed.) *Gradience in Grammar: Generative Perspectives*, 1-23. Oxford: Oxford University Press.

Featherston, Sam. 2005a. Magnitude Estimation and What It Can Do for Your Syntax: Some *wh*-constraints in German. *Lingua* 115(11): 1525-1550.

Featherston, Sam. 2005b. *That*-trace in German. *Lingua* 115(9): 1277-1302.

Flemming, Edward. 2004. Contrast and Perceptual Distinctiveness. In Bruce Hayes, Robert

Kirchner, and Donca Steriade (eds.), *Phonetically Based Phonology*, 232-276. New York: Cambridge University Press.

Fukuda, Shin, Grant Goodall, Dan Michel, and Henry Beecher. 2012. Is Magnitude Estimation Worth the Trouble?. In *Proceedings of the 29th West Coast Conference on Formal Linguistics*, pp. 328-336. Somerville, MA: Cascadilla Proceedings Project.

Gibson, Edward and Evelina Fedorenko. 2010. Weak Quantitative Standards in Linguistics Research. *Trends in Cognitive Sciences* 14(6): 233-234.

Gries, Stefan Th. 2015. More (Old and New) Misunderstandings of Collostructional Analysis: On Schmid and Küchenhoff (2013). *Cognitive Linguistics* 26(3): 505-536.

Gross, Steven, and Jennifer Culbertson. 2011. Revisited Linguistic Intuitions. *The British Journal for the Philosophy of Science* 62(3): 639-656.

Hall, Daniel Currie. 2011. Phonological Contrast and Its Phonetic Enhancement: Dispersedness without Dispersion. *Phonology* 28(1): 1-54.

Han, Chung-hye, Jeffrey Lidz, and Julien Musolino. 2007. V-Raising and Grammar Competition in Korean: Evidence from Negation and Quantifier Scope. *Linguistic Inquiry* 38(1): 1-47.

Han, Chung-hye, Julien Musolino and Jeffrey Lidz. 2016. Endogenous Sources of Variability in Language Acquisition. In *Proceedings of the National Academy of Sciences of the United States of America*, 113:4, pp. 942-947.

Honeybone, Patrick. 2011. Variation and Linguistic Theory. In Warren Maguire and April McMahon (eds.) *Analysing Variation in English*. Cambridge: Cambridge University Press.

Keller, Frank. 2000. Gradience in Grammar: Experimental and Computational Aspects of Degrees of Grammaticality. Ph.D. thesis, University of Edinburgh.

Keller, Frank. 2006. Linear Optimality Theory as a Model of Gradience in Grammar. In Gisbert Fanselow (ed.) *Gradience in Grammar: Generative Perspectives*, 270-287. Oxford: Oxford University Press.

Lakoff, Robin. 1977. You Say What You Are: Acceptability and Gender-related Language. In Sidney Greenbaum (ed.) *Acceptability in Language*, 73-86. The Hague, The Netherlands: Mouton Publishers.

Linzen, Tal and Yohei Oseki. 2015. The Reliability of Acceptability Judgments across Languages. Unpublished Manuscript.

Łubowicz, Anna. 2012. *The Phonology of Contrast.* Sheffield and Bristol, Conn.: Equinox Publishing Ltd.

Mathôt, Sebastiaan, Daniel Schreij, and Jan Theeuwes. 2012. OpenSesame: An Open-Source, Graphical Experiment Builder for the Social Sciences. *Behavior Research Methods* 44(2): 314-324.

Myers, James. 2009a. Syntactic Judgment Experiments. *Language and Linguistics Compass* 3(1): 406-423.

Myers, James. 2009b. The Design and Analysis of Small-scale Syntactic Judgment Experiments. *Lingua* 119(3): 425-444.

Myers, James. 2015. Markedness and Lexical Typicality in Mandarin Acceptability Judgments. *Language and Linguistics* 16(6): 791-818.

Newmeyer, Frederick J. 1983. *Grammatical Theory: Its Limits and its Possibilities*. Chicago, IL: University of Chicago Press.

Newmeyer, Frederick J. 2003. Grammar is Grammar and Usage is Usage. *Language* 79(4): 682-707.

Peirce, Jonathan W. 2007. PsychoPy – Psychophysics Software in Python. *Journal of Neuroscience Methods* 162(1): 8-13.

Phillips, Colin and Howard Lasnik. 2003. Linguistics and Empirical Evidence: Reply to Edelman and Christiansen. *Trends in Cognitive Sciences* 7(2): 61-62.

Phillips, Colin. 2009. Should We Impeach Armchair Linguistics?, In Shoishi Iwasaki, Hajime Hoji, Patricia M. Clancy, and Sung-Ock Sohn (eds.), *Japanese/Korean Linguistics*, Vol. 17, 49-64. Palo Alto, CA: CSLI Publications.

R Core Team. 2017. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Sampson, Geoffrey R. 2007. Grammar without Grammaticality. *Corpus Linguistics and Linguistic Theory* 3(1): 1-32.

Schütze, Carson T. 1996. *The Empirical Base of Linguistics: Grammaticality Judgments and Linguistic Methodology.* Chicago, IL: University of Chicago Press.

Song, Sanghoun, Sang-Geun Lee, Jae-Woong Choe, and Eunjeong Oh. 2017. Measuring the Degree of Convergence between Informal and Formal Acceptability Judgments for Korean. *Korean Journal of Linguistics* 42(4).

Sorace, Antonella and Frank Keller. 2005. Gradience in Linguistic Data. *Lingua* 115(11): 1497-1524.

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.

Sprouse, Jon. 2007. Continuous Acceptability, Categorical Grammaticality, and Experimental Syntax. *Biolinguistics* 1: 123-134.

Sprouse, Jon. 2008. Magnitude Estimation and the Non-Linearity of Acceptability Judgments. In *Proceedings of the 27th West Coast Conference on Formal Linguistics*, 397-403, Somerville, MA: Cascadilla Proceedings Project.

Traugott, Elizabeth Closs, and Graeme Trousdale. 2010. Gradience, Gradualness and Grammaticalization: How Do They Intersect. In Elizabeth Closs Traugott and Graeme Trousdale (eds.) *Gradience, Gradualness and Grammaticalization.* Amsterdam/

Philadelphia: John Benjamins Publishing.
Valian, Virginia. 1982. Psycholinguistic Experiment and Linguistic Intuition. In Simon, Thomas W., and Robert J. Scholes (eds.), *Language, Mind, and Brain*, pages 179-188, Hillsdale, NJ: Lawrence Erlbaum.
Wasow, Thomas and Jennifer Arnold. 2005. Intuitions in Linguistic Argumentation. *Lingua* 115(11): 1481-1496.
Weskott, Thomas and Gisbert Fanselow. 2011. On the Informativity of Different Measures of Linguistic Acceptability. *Language* 87(2): 249-273.

Sanghoun Song
[22012] Department of English Language and Literature
Incheon National University
119 Academy-ro, Yeonsu-gu, Incheon, Korea
E-mail: sanghoun@inu.ac.kr

Eunjeong Oh
[03016] Department of English Education
Sangmyung University
20 Hongimun 2-gil, Jongno-gu, Seoul, Korea
E-mail: eoh@smu.ac.kr

Received: 2017. 10. 18.
Revised: 2017. 11. 20.
Accepted: 2017. 11. 26.