Contrastive Corpus Analysis on the Writing of American and Korean Scientists using Gramulator

Hyun-Soon Min
(University of Memphis)
Philip M. McCarthy
(Decooda International)

Min, Hyun-Soon & McCarthy M. Philip. (2013). Contrastive Corpus Analysis on the Writing of American and Korean Scientists using Gramulator. ***Language Research*** **49.3, 681-702.**

Non-Native-English-Speaking-Researchers face many challenges when attempting to publish in English-language-journals. One challenge is that non-conventional language selection by Non-Native-English-Speaking-Researchers may appear like the writing of nave native-speakers. Consequently, such work may be more closely scrutinized. In this study, we focus on the linguistic choices that distinguish the academic writing of American and Korean scientists. We employ Contrastive Corpus Analysis, using the Gramulator, to extract "indicative" n-grams that inform us of variations in writing practices. Our results indicate that Korean scientists (in comparison to their American counterparts) use fewer personal pronouns, more past tense reporting verbs (e.g., examine, demonstrate, show), and greater employment of passive voice. We conclude that our findings may benefit Korean scientists and material developers who seek to better learn and employ the writing style of their American counterparts.

Keywords: Korean, The Gramulator, N-grams, Corpus, Scientific writing

1. Introduction

Writing journal text in a second language means having to learn and apply specific rhetorical and discourse characteristics that may be quite different from the writers' native language (Ferris & Hedgcock 2005, Hinkel 2002). That is, Non-Native-Speakers need to go beyond general linguistic knowledge of the target language and into a specific writing genre that corresponds to the expectations of the intended au-

dience (van Dijk & Kintsch 1983, H-S Min & McCarthy 2010, Zwaan 1993). As such, it can be no surprise that Non-Native-Speakers often perceive that writing in English is the most difficult task to successfully accomplish (Reid 1992).

Although academic written English is challenging for Non-Native Speakers, the burden of writing in a second language is especially important for Non-Native-English-Speaking-Researchers (whom we will refer to here as NNSR). For these people, writing in English is critical to their careers, because relatively few prestigious journals accept languages other than English (Duncan & Hall 2009, McCarthy et al. 2009). Despite this burden, NNSR may not have all the kinds of resources available that they might need. As McCarthy and colleagues demonstrate, texts providing information on the subtleties of linguistic features of journal level academic writing are rare (*cf.* McNamara, Graesser, McCarthy, & Cai 2014). Instead, NNSR typically have to be satisfied with sessions of proofreading from Non-Specialized English Speakers. As a result, the writings of NNSR can often reveal linguistic patterns that vary from the prototypical features of academic writings of native English speakers. This difference may be problematic because manuscripts submitted to prestigious journals must first be reviewed by experts in the field. And, given that NNSR are more likely to produce a non-standard variation of English; it is reasonable to assume that such differences are unlikely to enhance the prospects of the manuscript being accepted for publication (see Glanville, Sengupta & Forey 1998).

To facilitate NNSR in issues of academic writing in English, McCarthy et al. (2009) analyzed English texts written by Japanese scientists, British scientists and American scientists. Their study found evidence of significant differences for 14 different quantitative text analysis measures, primary among which were the Japanese employment of more verb phrases, the selection of higher frequency words, and the use of higher syntactical similarity between sentences. Building from the work of McCarthy and colleagues, Duncan and Hall (2009) analyzed English writings from American scientists, Korean scientists who were publishing in Korea, and Korean scientists who were publishing in America. Their findings suggest that the texts from Koreans publishing-in-Korea were the most distinct, and therefore, presumably, the least prototypical.

Both the previously mentioned studies called for further research into NNSR' writing. Specifically, the authors argued that there is the need to isolate and analyze tangible linguistic units that distinguish the writing of NNSR from the writing of Native-English-Speaking-Researchers (whom we will refer to here as NSR). The current study addresses that call by further investigating the writing of Korean scientists who have published in Korea. However, unlike the previous studies mentioned, which have offered a predominance of quantitative discourse metrics (e.g., frequency of verb phrases), the current study seeks to identify the variation of linguistic features within the text (i.e., the linguistic choices in terms of words and groups of words). Thus, the current study assesses the degree to which linguistic features may be the driving force in distinguishing the work of NNSR (in this case, Koreans) from their native English speaking NSR counterparts. Through such an approach, we aim to address our primary research questions:

- Do American scientists employ distinct linguistic features in comparison to Korean scientists?
- Do Korean scientists employ distinct linguistic features in comparison to American scientists?
- Do any features of American scientists offer insights for the development of facilitative resources for NNSR' writing?

2. Difficulties of NNSR in Academic Setting

Although any number of journals and conferences may be held in any number of languages, the preeminent journals, conferences (and publishing houses, universities and so forth), remain English dominant. As such, NNSR are only likely to find an increasing disadvantage in publishing their work in these outlets (Cho 2009, Flowerdew 1999a, Li & Flowerdew 2007, McCarthy et al. 2009). Flowerdew (1999a) addressed this issue by surveying 585 NNSR (Cantonese as first language) in Hong Kong. Using questionnaires, he found that two thirds (68%) of the respondents felt at a disadvantage by publishing their work as non-native English speakers. Specifically, the respondents included as their reasons for feeling at a disadvantage: technical problems (51%), editors' prejudice (29%), organizational factors (14%), innovative thinking (11%),

difficulty incorporating existing literature (9%), and difficulty weighing value of literature (8%). Flowerdew (1999b) further elaborated this problem of publishing in non-Native English using a survey of 26 Cantonese professors. The survey revealed that they identified the problematic areas in writing English discourse as lengthy writing time, limited knowledge of English vocabulary, and negative L1 transfer. Such research suggests that the lack of appropriate linguistic knowledge of English has a negative impact for NNSR in terms of attaining publication.

NNSR also encounter academic challenges, especially those with insufficient knowledge of English writing. D-W Cho (2009) conducted a survey of 59 professors and 271 graduate students in Korea. Structured oral interviews were also administered to 3 professors and 5 Ph.D. students. The participants selected English article usage as the most troubling area of English, followed by syntax, prepositions, conjunctions, gerunds vs. to-infinitives, voice, and singulars vs. plurals. However, the results of the interview showed that NNSR' perception of feeling at a disadvantage differed according to their self-reported levels of English writing proficiency. That is, only those who had low confidence in their English argued that they were at a disadvantage in publishing their works in English journals. Correspondingly, D-W Cho (2009) suggests that linguistic knowledge of academic discourse in English would help NNSR' overcome their perceived problem. As such, NNSR appear to be convinced that they need to obtain better linguistic knowledge of English academic discourse.

In supporting the NNSR' claim, editors of academic journals also showed a concern that NNSR' non-standard linguistic choices might hinder the publication of their work. Flowerdew (2001) conducted an interview with 12 editors of social science journals (e.g., *Applied Linguistics*, *the Asian Journal of English Teaching*, and *the Journal of Phonetics*). The editors that were interviewed concurred that NNSR' English writing contained linguistic errors such as article usage and subject-verb agreements. Moreover, their writings lacked the structural moves that are necessary in introduction and literature review sections of research articles (Flowerdew 2001). Similarly, Gosden (1992) examined how editors in hard science journals (e.g., chemistry, physics, and biology) perceived NNSR's writing in the review process. By analyzing 154 returned surveys from editors in Canada, U.S. and U.K., his find-

ings demonstrated that a lack of linguistic knowledge among NNSR may play a significant role in leading to the rejection of the work, especially if their work is considered as mediocre or borderline. Taken as a whole, the research suggests that NNSR need better linguistic knowledge of English academic discourse so as to help them gain greater access to the international academic community.

3. Method and Materials

3.1. Method: Contrastive Corpus Analysis

The goal of our study is to identify the characteristics of one manifestation of scientific writing (Korean English) relative to a second manifestation of scientific writing (American English). Further, in identifying those characteristics, we seek to offer insights into instructional resources for NNSR. To achieve this goal, we use a textual analysis approach originated in the field of Second Language Learning that we refer to here as contrastive corpus analysis (CCA: see Cobb 2003, Granger 1998). CCA differs from more traditional corpus analyses inasmuch as the emphasis switches from what a single collection of texts can reveal about quantities or distributions of language features (e.g., Biber, Conrad, & Reppen 1998, Stubbs 1996) to an emphasis on what two (or more) highly related corpora (sister corpora) can reveal when their commonalities are excluded through computational and statistical techniques. Thus, the argument is that given two corpora that differ minimally (e.g., the sister corpora of scientific writing in English by Korean scientists and the scientific writing in English by American scientists), whatever is characteristic of one corpus, but non-characteristic of the corresponding sister corpus, is what is indicative of the text type.

Forms of CCA are rapidly establishing a strong reputation in Second Language Learning as the corpus analysis approach of choice. This reputation began with the research described in Granger (1998), and has been enhanced by the subsequent advancements made thanks to systems such as Coh-Metrix (Crossley, McCarthy, & McNamara 2007, Crossley et al. 2007, McCarthy et al. 2009). CCA reveals pervasive yet ill-defined underlying patterns of texts that not only reveal the constructs of text types (e.g. cohesion or readability values), but, also high-

light the linguistic features that are building blocks (i.e. lexical norms) of the text type. Subsequently, these linguistic features can be used as the basis for materials for language learners, allowing greater use of prototypical forms, and fewer instances of linguistic anomalies. An early example of the CCA using this approach in Second Language Learning is Conrad (1996), who found significant differences between the writing styles in the academic prose of text books on ecology, research articles on ecology, and general English compositional books. Her findings suggested that second language learners might not be able to acquire the appropriate writings styles for research articles with only the help of general English compositional books. Such research allows us to argue that CCA makes it possible to access a variety of linguistic features of academic writings and also provides instructional resources relevant to language learners. In a more specific example of materials development, Trebits (2009) uses computational tools such as Wordsmith and WordNet to identify language features in the Corpus of the European Union English (CEUE) relative to general English. The findings of Trebits led to proposals for several teaching activities such as gap-filling and paraphrasing using contextualization approaches. In a similar study, Gamon et al. (2009) developed ESL Assistant, the automated correction system by assessing three corpora of Chinese and Japanese writing and analyzing their errors relative to native English speakers. Finally, the research that forms the foundation of the current study (Duncan & Hall 2009, McCarthy et al. 2009) uses CCA techniques based on quantitative data provided by Coh-Metrix. Collectively, studies such as these support the notion that CCA can highlight systematic linguistic patterns of NNSR' writings and provide helpful information in language instructional activities.

3.2. Materials: The Corpus

Our corpus comprises 750 abstracts culled from 31 experimental scientific journals; including the genres of chemistry, biology, and physics; and all published from 2001 to 2010. From these texts, two individual corpora were compiled: Korean scientists in Korea and American scientists in America. The Korean English corpus comprises Korean scientists' abstracts (n = 375), published exclusively in 15 different Korean journals. The American English corpus (the assumed

prototypical model) comprises U.S. scientists' abstracts (n = 375), published exclusively in 16 U.S. journals.

To establish confidence that the authors of the relative texts were either Korean or American, the model of McCarthy et al. (2009) was followed (see also Duncan & Hall 2009). The model has two major criteria: (1) the first author (in the field of science, typically the person who leads the projects and writes most of the paper) and the last author (typically the supervisor) are required to be from institutes within their respective countries. (2) The names of the primary and final authors must be 'typical' of the country of the classification. That is, the primary and final authors in the Korean English and American English corpora represent the typical names for Koreans and Americans respectively. Admittedly, these criteria cannot ensure that the papers were written exclusively, or even predominantly, by the primary and final authors. However, as McCarthy and colleagues demonstrate, these criteria of classification are effective in determining the language backgrounds of the writers. The authenticity of Korean names as 'typical' is not hard to establish, as relatively few Korean scientists are likely to have non-Korean names. Consequently, for the Korean English corpus, one native speaker of Korean (the first author of the current study) evaluated whether the names met the criteria. However, for the American English corpus many native English speakers have family names that are not of English origin. As such, three native speakers of English (all graduate students of linguistics) reviewed the selections to evaluate whether the collected texts represented names that could be described as typical. If any two of the evaluators agreed that a text did not meet the selection criteria, that text was excluded. When a total of 750 texts are assessed in this fashion, it is likely that some categorizational errors will occur; however, to fall outside of reasonable statistical assumptions, over 5% of the texts would have to be misaligned, meaning more than 35 texts in this case. Statistically, this eventuality is unlikely.

The current study follows McCarthy et al. (2009) and Duncan and Hall (2009) in focusing only on the abstracts of the articles. We acknowledge that abstracts may not be fully representative of all the linguistic features of the subsequent manuscript, and that abstracts themselves may differ from journal to journal; however, abstracts form a reasonable point of departure for our study because they are relatively

easy to collect, representative of the entire paper, and are provided freely to the public both in America and in Korea. More importantly though, abstracts are typically the first item of an article that is read, and also the most frequent item that is read, making abstracts perhaps the most important textual section of an article (McCarthy, Briner, Rus, and McNamara 2007).

3.3. Tool: The Gramulator

This study uses the textual analysis tool the Gramulator (McCarthy, Watanabe, & Lamkin 2012). The Gramulator facilitates contrastive corpus analysis through a process of machine differential diagnosis (MDD). MDD allows for the identification of indicative lexical features of correlative text types in the form of n-grams, or adjacently positioned lexical items in a text. In this study, we focus on two-word n-grams (or, bigrams); and, more exactly, on differential bigrams (D-bigrams). D-bigrams are bigrams that are typical to one corpus (i.e., among the 50% most frequent bigrams, excluding hapax legomena) but not typical to the contrasting corpus (i.e., not among the 50% most frequent bigrams). In identifying D-bigrams, we highlight the most indicative and least indicative language sequences of the two corpora, and, based on these results not only ascertain whether a difference exists between the two text types in terms of lexical features, but also demonstrate what some of those lexical features are and why they might be present.

4. Results

In the present analysis, we follow Witten and Frank (2005) by randomly dividing two third of texts in each corpus into training-set data (i.e., 275 texts) and one third into testing-set data (constituting the remaining 125 texts). This model of analysis allows us to attain two goals. First, a training set guards against excessive analysis that may lead to the violation of statistical validity (i.e., a type 1 error). Second, a test set allows us greater confidence that the findings from the training-set data can be generalized. Thus, if there is consistency in the findings, we can argue that our results are indeed systematic linguistic

features of specific groups.

The training-set data of the current study provided the differentials of Korean texts (relative to American texts), and American texts (relative to Korean texts). In Gramulator nomenclature, this is written as Korean (American) and American (Korean) respectively. From the analyses of these differentials, we build hypotheses to explain the usage of the differentials and then test our hypotheses using the held back testing-set data.

In the hypothesis testing, the Gramulator utilizes the Fisher's Exact test to assess the counts of differentials. Specifically, the test credits one abstract from the corpus as one unit if any number of occurrences is present. This procedure is appropriate because the texts in our corpus are not all of identical lengths.

Our analysis takes into account differentials in two forms: the simple bigram (i.e., the actual two words of the bigram) and the flexigram (i.e., the underlying or theoretical form of the bigram). Thus, for example, *my mother* and *my father* form two distinct bigrams but could also be represented as one flexigram: *my* + *parent*. Similarly, members of the same lemma can be represented as a flexigram as in *he* + *goes*/*went*. Flexigrams are useful in this kind of analysis because they describe different words and wordings that are performing a similar function. The Gramulator's Concordancer and Evaluator modules assist us in this analysis by highlighting examples of differentials in context.

The initial Gramulator results on the training-set data revealed 52,674 words and 150 D-grams for the American corpus and 49,331 words and 143 D-grams for the Korean corpus. Of these, our analysis centers on the most frequent occurring differentials of the American and Korean corpora.

4.1. American Differentials

Of the 150 D-bigrams in the training set-data of the American corpus, the most frequent American differential was *that are*. This differential is employed 29 times across 23 texts out of the 250 training-set texts. In contrast, Korean scientists employ *that are* significantly less frequently: 2 times across 2 of their corresponding 250 training-set texts (American: 9.20%, Korean = 0.80%, $p < .001$). Closer comparison of

the differentials shows that the Koreans also employ *which was* in a functionally similar form to *that are*. Note at that point that the Korean language does not distinguish the restrictive and non-restrictive relative clause, which is how these differentials are used in the corpora.

Broadening the analysis, we modified *that are* to include not only instances of *that are* but also instances of *that is*. In Gramulator nomenclature, this kind of either/or option is an example of a flexigram and it is written as *that* + *is/are*. The result of the search for the flexigram *that* + *is/are* called for further analysis as to whether American preference of *that* + *is/are* is consistent when the *be* verb is in the past tense form (i.e., *was/were*). To address this issue, we compiled an index of 8 bigrams for each of the possible variations of *that/which* + *be*: *that is*, *that are*, *that was, that were*, *which is*, *which are*, *which was*, and *which were*. In doing so, we closely examined each instance through the context provided by the concordancer and selected those occurrences that are used in a relative construction. We then reran this index on the testing-set data to assess whether this frexigram varies between the writing of American and Korean scientists. For the present tense (i.e., *that is*, *that are*, *which is*, and *which are*), the difference in frequency was approaching significance: Americans = 29 times out of the 125 testing-set texts; and Koreans = 17 times (American = 23.20%, Korean = 13.60%; p = 0.072). By contrast, the analysis of the past tense (i.e., *that was, that were, which was*, and *which were)* showed far lower frequencies and no significant difference across the corpora: Americans = 6 times and Koreans = 9 times (American = 4.80%, Korean = 7.20%; p = 0.596). Taken together, the results on the testing-set data suggest that American scientists may prefer the sequence of *that, which* + *be* verb in the present tense, as compared to their Korean counterparts. However, American scientists show no difference from Korean scientists in the use of *that, which* + *be* verb in the past tense. The potential difference in the preference of tense described here suggests that Korean scientists either lack the linguistic knowledge of conventions regarding tense choice in English journal text or, at least, that they do not employ those conventions to a similar degree as their native speaking counterparts.

To identify other indicative linguistic features of American scientists as compared to their Korean counterpart, we return to the results of the training-set data. Our analysis suggested that there was a system-

atic difference in a distinction in how American and Korean scientists used language to report their findings. Specifically, American scientists appeared to make greater use of the flexigram *we* + [*reporting verbs*], with 7 American differentials taking such forms as *we show, we demonstrate*, *we* and *we find*. Of these semantically related and syntactically identical bigrams (thus, a flexigram), *we show* is the highest ranked example (2^{nd}) in the training-set data, with *we report* (7^{th}), and *we demonstrate* (8^{th}) also ranking highly. In terms of comparative usage between the corpora of Americans and Koreans, *we show* is employed by Americans for 20 instances across 19 of 250 training-set data, whereas Koreans use it just 1 time in their corresponding 250 texts (American = 7.60%, Korean = 0.40%; $p < .001$). In contrast to the American differentials, only one Korean differential takes the pattern of *we* + [*reporting verbs*], *we examined*.

To guard against the possibility of type 1 errors caused by excessive analysis in one set of data (i.e., training set-data in this case), we reran the findings using the testing-set data. Combining the 7 *we* + [*reporting verbs*] bigrams into a flexigram, the difference in usage between the sister corpora was significant, 33 and 7 times in American and Korean corpus respectively (American = 26.40%, Korean = 5.60%; $p < .001$). However, the results of the difference for each differential (see Table 1) reveal that the American scientists' preference for *we* + [*reporting verbs*] center on just two of the differentials (i.e. *we show* and *we find*).

Table 1. The use of we + [reporting verbs] between American and Korean scientists

we + [*reportingverbs*]	Rank	American		Korean		*p*-value
		Present	absent	present	Absent	
we show	3	16(4.8%)	109(87.2%)	1(0.8%)	124(99.2%)	<.001*
we report	7	4(3.2%)	121(96.8%)	1(0.8%)	124(99.2%)	.370
we demonstrate	8	6(4.8%)	119(95.2%)	1(0.8%)	124(99.2%)	.120
we find	33	8(6.4%)	117(93.6%)	0(0%)	125(100%)	.007*
we identified	87	4(3.2%)	121(96.8%)	0(0%)	125(100%)	.122
we used	89	3(2.4%)	122(97.6%)	1(0.8%)	124(99.2%)	.622
we present	123	11(8.8%)	114(91.2%)	5(4.0%)	120(96%)	.195

Note: * indicates a level of significance.

The above results of the flexigram *we* and *each of the reporting verbs* prompted further investigation as to whether such language is also indicative at the uni-gram level. In other words, do the two groups of scientists differently employ the personal pronoun *we*? and/or do they differently employ the 7 reporting verbs?

Assessing the pronoun first, we predicted that Korean scientists would employ fewer instances of *we*. This lesser usage can be attributed to a lack of knowledge of the appropriate English academic writing style and/or because of issues of inter-language transfer (Hinkel 1997, C-K Kim 2009, Kabota 1998). Note that the Koreans' rhetorical style discourages the use of any first person pronoun forms in writings (C-K Kim 2009). To test our hypothesis, we conducted a Fisher's Exact test using the testing-set data to assess the frequencies between the American-English and Korean-English corpora. In doing so, we excluded the singular form of the personal pronoun *I*, which had only one incidence in the American corpus and none at all in the Korean corpus. The analysis on the personal pronoun *we* revealed a significant difference in the frequency of use: American= 63.2%; Korean= 32.8%; $p < .001$. The result suggests that a lower usage of the *we* personal pronoun is an indicative linguistic feature of Korean scientists as compared to their American counterparts. The lesser use of Korean personal pronoun use prompts us to ask how Koreans are functionally performing the act of reporting. We will address this issue in detail in a later section (Korean Differentials), specifically assessing the Korean preference for passive voice.

Turning to the use of the 7 individual reporting verbs (described earlier), we needed to consider the possibility that Korean scientists might employ each of the verbs without using *we* as the precedent pronoun. To assess this possibility, we included all possible forms that the scientific reporting verbs could take in a new analysis. For example, the lemma *show* can take five forms: *show*, *shows*, *showing*, *showed* and *shown*. Thus, we compiled a list of 29 forms and categorized them into four types: zero morpheme, third person singular, past participle, and present progressive. Note that there were no cases of present progressive form in either corpus, suggesting that such a form is not practiced in academic writings by either group of scientists. Having compiled the list of 29 forms into a single index, we then used this index to analyze the testing-set data. Our results were modified by identifying and re-

moving any incidences of the forms performing non-verbal functions (e.g., *using* as a gerund in the sentence of we report high-throughput analysis, using massively parallel signature sequencing) and differentiating the cases when past and past participle are identical (e.g., the past and past participle of the verb *find* are identical) by closely studying the context of the verbs. Analyses on the four types of morpheme indicate that American scientists appear to use significantly more instances of zero morpheme verbs (e.g., *show*, *demonstrate*) whereas the remaining types of morpheme (i.e. third-person singular and past participle) demonstrate no significant differences across the two groups of scientists (see Table 2). There was no significant difference in terms of the total usage of 29 verb forms (American: 87%, Korean: 80%; p = 0.171).

Table 2. The use of 7 reporting verbs on four types of morpheme between American and Korean scientists

Types of Morpheme	American		Korean		*p*-value
	Present	Absent	Present	Absent	
Zero morpheme (e.g. show, find)	66(52.8%)	59(47.2%)	23(18.4%)	102(81.6%)	< .001*
Third person singular (e.g. shows, finds)	13(10.4%)	112(89.6%)	9(7.2%)	116(92.8%)	.504
Past participle (e.g. shown/showed, found)	75(60.0%)	50(40.0%)	79(63.2%)	46(36.8%)	.697
Present participle (e.g. showing, finding)	0(0%)	125(100%)	0(0%)	125(100%)	1

Note: * indicates a level of significance.

Taking together the results of the personal pronoun use (*we*) and the verbs of reporting, our analysis suggests that American scientists (in comparison to their Korean counterparts) employ more instances of the personal pronoun *we* and also more instances of the 7 verbs of reporting with the zero morpheme. The two results are consistent, inasmuch as *we* takes the zero morpheme form of verbs. More importantly, the result demonstrates that Korean scientists are as prevalent with verbs of reporting as their American counterparts although they are using different linguistic forms.

4.2. Korean Differentials

The initial Gramulator analysis on the training-set data produced 143 D-grams of the form Korean (American). Of these differentials, *was not* ranked highest. This Korean differential has 24 instances across 21 of 250 training texts whereas Americans employ the differential bigram for 4 instances across 4 of 250 training texts (Korean = 1.60%; American = 9.60%; $p < .001$).

The past tense form of this bigram reminds us of the apparent Korean preference for past over present (see previous section). However, the presence of *not* in the bigram could signal that past tense use is more likely to be predominant in negative forms. To examine this observation further, we used the testing set data to compare the bigrams in their present and past forms. For the past tense structure (i.e., *was/ were* + *not*), there was a significant difference between the two groups of scientists, 17 and 2 times out of the 125 testing-set data for Koreans and Americans respectively (Korean = 13.06%, American = 1.60%, $p < .001$). The result supports the previous findings that Korean scientists prefer using the past tense. Correspondingly, for the present tense structure (i.e., *am/are/is* + *not*), Korean scientists showed significantly lower usage, 3 times and 20 times out of the 125 testing-set data for Koreans and Americans respectively (Korean = 2.40%, American = 16.00%, $p < .001$). Taken together, the results suggest that Korean scientists prefer the past tense structure of *was/ were* + *not* as opposed to their American counterparts' preference for *am/are/is* + *not*.

Returning to the training set data, further analysis of the Korean differentials led us back to the subject of reporting verbs. More specifically, it led us to an apparent Korean scientists' preference for these verbs in the past passive form, suggesting that the *was/were* prevalence is not merely a negative phenomenon. In total, 16 Korean differentials took some form of the flexigram *was/were* + [*reporting verbs*]. Of these bigrams, *were investigated* was the highest ranked example (11^{th}), with *was observed* (14^{th}), and *were observed* (16^{th}) also highly ranked. In terms of comparative usage, *were investigated* is employed by Koreans for 11 instances across 11 of 250 training-set data, whereas Americans use it just 1 time in their corresponding 250 texts (American = 0.4%, Korean = 4.4%; $p = .006$). In contrast to Korean differentials, only two American differentials use a similar pattern and both are in the present tense

form: *are associated* and *is known*.

The Korean scientists' preference for the structure of *was/were* + [*past participle reporting verbs*] is supported by the testing-set data. For that analysis, we compiled a list of 16 bigrams (i.e., *was/were investigated*, *was/were observed*, *were prepared*, *was/were identified*, *was/were examined*, *were used*, *was obtained*, *was increased*, *was studied*, *were compared*, *was conducted*, and *were measured*). Combining the structure of the 16 *was/were* + [*past participle of reporting verbs*] bigrams into a flexigram, the difference in usage between the sister corpora was significant (American = 9.60%, Korean = 36.8%; $p < .001$). Therefore, the results suggest that Korean scientists prefer the structure *was/were* + [*past participle of reporting verbs*] when compared to American scientists.

The analyses described above led us to ask (1) Do the two groups of scientists differently employ the reporting verbs that we extracted from the Korean differentials? (2) Do the two groups of scientists differently employ tense? and/or (3) Do the two groups of scientists differently employ voice? We predicted that the Korean scientists would employ the 12 reporting verbs as often as their American counterparts. However, we predicted that Korean scientists would prefer the past tense because of the lack of familiarity to conventions in English academic writings. We also predicted that Korean scientists would favor the passive voice, transferring the rhetorical preference of their first language in academic writings by suppressing the agent of action (C-K Kim 2009).

Assessing the use of the reporting verbs, we first extracted the 12 lemma from the 16 instances of was/were + [*past participle of reporting verbs*]. From these lemma, we then constructed an index of 48 verb forms: each lemma can take four verb forms (e.g., *investigate* can take *investigate*, *investigates*, *investigating*, and *investigated*). We used the concordancer module of the Gramulator to assess each instance in context, and excluded any instance of a target form performing a non-verb function (e.g., using *study* as a noun in natural and artificially created samples. This study demonstrated the need for caution in the direct). The results indicate that there was no significant difference in using the 12 *reporting verbs* (American = 69.60%, Korean = 69.60%; $p = 1$).

We then sought to determine whether the two groups of scientists use the 12 verbs of reporting with functionally different verb forms (i.e., *tense* and/or *voice*). To examine the comparative use of tense, we

used the 12 lemma of verbs of reporting, which were extracted from the 16 *was/were* + [*past participle of reporting verbs*] structures. Theses lemmas were then transformed into their present tense forms. Although the *be*-verb has three forms of present tense (i.e., *am*, *are*, and *is*), we only used *are* and *is* because *am* does not occur in either of the corpora. Also, we added the present perfect form (i.e., have/has been) in this analysis. As a result, we compiled 48 structures of 12 *is/are/has been/have been* + [*past participle of reporting verbs*]. The results indicate that there is no difference between Korean and American scientists in terms of the 48 *is/are* + [*past participle of reporting verbs*] (American = 8.00%, Korean = 6.40%; p = 0.807). As such, Korean scientists do not appear to significantly differ from their American counterparts in their employment the present passive voice (see Table 3).

In examining the use of *past* passive voice, we compiled 48 verbs because each verb can take two past passive voices forms (e.g., *was/were investigated*). We excluded the instance of the past perfect form (e.g., *had been investigated*) in this index because it was absent in both corpora. The findings indicate that Korean scientists are significantly more likely to use the *past* passive structure than their American counterparts (American = 17.60%, Korean = 49.60%; p <.001). Taken together, the analyses on the usage of tense in the passive structure of 12 reporting *verbs* reveal that Koreans implement the past tense significantly more often than their American counterparts. However, in the use of the present tense, there is no significant difference across the corpora (see Table 3).

Table 3. The comparative use of tense in the passive voice for 12 reporting verbs used by American and Korean scientists

Tense	American		Korean		*p*-value
	Present	Absent	Present	Absent	
Present passive voice (e.g., is/are investigated)	10(8.0%)	115(92.0%)	8(6.4%)	117(93.6%)	0.807
Past passive voice (e.g., was/were investigated)	22(17.6%)	103(82.4%)	62(49.6%)	63(50.4%)	< .001*

Note: * indicates a level of significance.

Turning to the examination of the comparative use of voice (as opposed to tense), we first compared the usage of active voice across the corpora. To do so, we changed the 12 lemma of reporting verbs (i.e., *investigate*, *observe*, *prepare*, *identify*, *examine*, *use*, *obtain*, *increase*, *study*, *compare*, *conduct*, and *measure*) into active structures. The final index comprised a list of 48 verb forms, following the same procedure as previously discussed for the American flexigram of *we* + [*reporting verbs*]. Moreover, we included the present perfect structure (i.e., *has/have* + verb) in this index of active voice. As a result, we constructed the index with a total of 72 verb forms. We then extracted all cases of the 72 verb forms from each corpus and from these incidences we selected the cases that used the main verbs in active voice. That is, we excluded any non-verb function (e.g., using *study* as a noun in the sentence of "··· this study demonstrated that the cell-occupied method ···") by closely examining the context of each instance. Analyses on the active voice indicated a significant difference across the corpora (American = 52.00%, Korean = 32.00%; p =.002).

We then examined the comparative use of passive structure that employs the 12 *reporting verbs* across the corpora. Each verb can have eight forms of passive structure. For example, the verb *investigate* can take *am/are/is investigated*, *was/were investigated*, and *has/have/had been investigated.* However, we exclude the structure of *am* + *reporting verbs* because it does not occur in either of the corpora. As a result, we compiled a total of 84 passive structures in this analysis. The results indicate that there is a significant difference in employing passive structure across the corpora (American = 25.60%, Korean = 53.60%; p <.001). Collectively, the analysis on the comparative usage of voice using the 12 reporting verbs suggests that Koreans employ significantly fewer active voice structures but greater passive structures as compared to their American counterparts (see Table 4).

Table 4. The comparative use of voice in 12 reporting verbs between American and Korean scientists

Voice	American		Korean		*p*-value
	Present	Absent	Present	Absent	
Active voice (e.g., investigate, investigates)	65(52.0%)	60(48.0%)	40(32%)	85(68.0%)	.002*
Passive voice (e.g., is/was investigated)	32(25.6%)	93(74.4%)	67(53.6%)	58(46.4%)	< .001*

Note: * indicates a level of significance.

Taken together, the results of *be* verb + *not* and the reporting verbs suggest that Korean scientists (in comparison to their American counterparts) employ more instances of the past tense *be* verb + *not* and also more instances of the 12 reporting verbs in passive structure, especially in past tense. These results lead us conclude that Korean scientists use reporting verbs as commonly as their American counterparts, although the Koreans appear to be using different linguistic forms of verbs.

5. Discussion

In this study, we explored systematic linguistic features in the academic writings of Korean scientists as compared to the academic writings of American scientists. In doing so, this study serves to inform Korean researchers and prospective material designers as to the (presumably facilitative) discourse characteristics of English, and the correspondingly (and presumably deleterious) discourse characteristics that are commonly employed by Korean scientists. The study addressed three central research questions: 1) Do American scientists employ distinct linguistic features in comparison to Korean scientists? 2) Do Korean scientists employ distinct linguistic features in comparison to American scientists? 3) Do any features of American scientists offer insights for the development of facilitative resources for NNSR' writing?

Addressing the first question, our response is that American scientists appear to employ linguistic features in different forms from their

Korean counterparts. Specifically, American scientists preferred the personal pronoun *we*, the present tense, and active voice in the use of reporting verbs (e.g., *show*, *demonstrate*), as compared to the Korean scientists. Importantly, the results show that while Korean scientists' deployment of reporting verbs is different from their American counterparts, the employment of the verbs (e.g., their frequency) is highly similar.

To address our second question, our response is that Korean scientists share commonalities on the choice of words with their American counterparts but they differed in terms of how they presented the words. That is, Korean scientists preferred fewer personal pronouns (i.e., *we*), the past tense, and passive voice in the use of reporting verbs as compared to their American counterparts. Additionally, our results show that Korean scientists employ more instances of the past tense *be* verb in conjunction with the negation *not*.

To answer our third question, we find that linguistic features of American scientists generated by the Gramulator can help us better understand characteristics of the prototypical writings (e.g., personal pronouns, present tense, and active voice in the use of reporting verbs). Thus, the findings from this contrasting corpus analysis may benefit material developers and researchers who aspire to learn prototypical linguistic features of the native English speaking scientific community.

Although this study has important implications for identifying the linguistic differences between American and Korean scientists, many questions remain. Primary among these questions are how well these findings can generalize to different types of reporting verbs (e.g., *explore*, *reveal*), and different sections of research journals (e.g., the introduction section, the discussion section etc.), and we also need to know how well these findings can generalize into different areas of academic research (e.g., psychology articles, computer science articles etc.). Additionally, the issue of L1 transfer on these linguistic choices needs to be better understood. Thus, future research needs to examine the breadth of the linguistic structures in diverse sections and genres of research articles across a number of non-English languages.

To be sure, whatever the identified linguistic differences between NSR and NNSR usage, the most important element is whether these differences actually have an effect on the readership, and more precisely, whether this effect is positive or negative. Thus, future studies

need to assess whether changes made to texts as a result of such differential analysis has an effect on reviewers and the subsequent success of Non-Native-English-Speaking-Researchers.

Acknowledgments

The authors also acknowledge the contributions of Charles Hall, Diana Lam, George Min, Hojin Lee, Scott Healy, and Travis Lamkin.

References

Biber, D., Conrad, S., & Reppen, R. (1998). Corpus linguistics: Investigating language structure and use. Cambridge, Cambridge University Press.

Cho, D.W. (2009). Science journal paper writing in an EFL context: The case of Korea. *English for Specific Purposes* 4.28, 230-239.

Cobb, T. (2003). Analyzing late interlanguage with learner corpora: Qubec replications of three European studies. *The Canadian Modern Language Review/La Revue canadienne des langues vivantes* 59.3, 393-423.

Conrad, S. (1996). Investigating academic texts with corpus-based techniques: an example from biology. *Linguistics and Education* 8, 299-326.

Crossley, S.A., Louwerse, M., McCarthy, P.M., & McNamara, D.S. (2007). A linguistic analysis of simplified and authentic texts. *Modern Language Journal* 91, 15-30.

Crossley, S.A., McCarthy, P.M., & McNamara, D.S. (2007). Discriminating between second language learning text-types. In D. Wilson and G. Sutcliffe. eds., *Proceedings of the twentieth International Florida Artificial Intelligence Research Society Conference*, 205-210. Menlo Park, The AAAI Press.

Duncan, B. & Hall, C. (2009). A coh-metrix analysis of variation among biomedical abstracts.

Proceedings of the twenty second International Florida Artificial Intelligence Research Society Conference, 237-242. Menlo Park, The AAAI Press.

Ferris, D. & Hedgcock, J. (2005). Teaching ESL composition. New Jersey, Lawrence Erlbaum Associates.

Flowerdew, J. (1999a). Writing for scholarly publication in English: The case of Hong Kong. *Journal of Second Language Writing* 8, 123-145.

Flowerdew, J. (1999b). Problems in writing for scholarly publication in English: The case of Hong Kong. *Journal of Second Language Writing* 8, 243-264.

Flowerdew, J. (2001). Attitudes of journal editors to nonnative speaker contributions. *TESOL Quarterly* 35, 121-150.

Gamon, M., Leacock, C., Brockett, C., Gao, J., & Klementiev, A. (2009). Using statistical techniques and web search to correct ESL errors. *CALICO Journal* 26.3, 491-511.

Glanville, R., Sengupta, S., & Forey, G. (1998). A (cybernetic) musing: Language and science and the language of science. *Cybernetics and Human Knowing* 5.4, 61-70.

Gosden, H. (1992) Research writing and NNSs: from the editors. *Journal of Second Language Writing* 1.2, 123-139.

Granger, S. eds. (1998). Learner English on computer. Longman.

Hinkel, E. (1997). Indirectness in L1 and L2 academic writing. *Journal of Pragmatics* 27.3, 361-386.

Hinkel, E. (2002). Second language writers' text linguistic and rhetorical features. Mawhwa, Lawrence Erlbaum Associates.

Kabota, R. (1998). An investigation of L1L2 transfer in writing among Japanese university students: Implications for contrastive rhetoric. *Journal of Second Language Writing* 7.1, 69-100.

Kim, C.K. (2009). Personal pronouns in English and Korean texts: A corpus-based study in terms of textual interaction. *Journal of Pragmatics* 41, 2086-2099.

Li, Y.,& Flowerdew, J. (2007). Shaping Chinese novice scientists' manuscripts for publication. *Journal of Second Language Writing* 16, 100-117.

McCarthy, P.M., Briner, S.W., Rus, V., & McNamara, D.S. (2007). Textual signatures: Identifying text types using latent semantic analysis to measure the cohesion of text structures. In: A. Kao, & S. Poteet (Eds.). *Natural Language Processing and Text Mining*, 107-122. London, Springer-Verlag UK.

McCarthy, P.M., Hall, C., Duran N.D., Doiuchi, M., Duncan, B., Fujiwara, Y., & McNamara, D.S. (2009). A computational analysis of journal abstracts written by Japanese, American, and British scientists. *The ESPecialist* 30, 141-173.

McCarthy, P.M., Myers, J.C., Briner, S.W., Graesser, A.C., & McNamara, D.S. (2009). A psychological and computational study of sub-sentential genre recognition. *Journal for Language Technology and Computational Linguistics* 24, 23-55.

McCarthy, P.M., Watanabi, S., & Lamkin, T.A. (2012). The Graumlator: A tool to identify differential linguistic features of correlative text types. In P.M. McCarthy & C. Boonthum-Denecke (Eds.), *Applied natural language processing: Identification, investigation, and resolution*, 312-333. Hershey, IGI Global.

McNamara, D.S., Graesser, A.C., McCarthy, P.M., & Cai, Z. (2014). Coh-Metrix: Automated Evaluation of Text and Discourse. Cambridge University Press.

Min, H.C. & McCarthy, P.M. (2010). Identifying Varietals in the Discourse of American and Korean Scientists: A Contrastive Corpus Analysis Using the Gramulator. In H. W. Guesgen & C. Murray (Eds.), *Proceedings of the 23rd International Florida Artificial Intelligence Research Society Conference*, 247-252. Menlo Park, The AAAI Press.

Notre Dame, University of Notre Dame Press.

Reid, J. (1992). A computer text analysis of four cohesion devices in English discourse by native and nonnative writers. *Journal of Second Language Writing* 1.2, 79-107.

Stubbs, M. (1996). Text and corpus analysis. Oxford, Blackwell.

Swales, J.M. (1990) *Genre analysis: English in academic and research Settings*. Cambridge, Cambridge University Press.

Trebits, A. (2009). The most frequent phrasal verbs in English language EU documents: A corpus-based analysis and its implications. *System* 37, 470-481.

Witten, I. & Frank, F. (2005). Data mining: Practical machine learning tools and techniques (2nd ed.). San Francisco, Morgan Kaufmann.

Van Dijk, T.A. & Kintsch, W. (1983). Strategies of discourse comprehension. New York, Academic Press.

Zwaan, R.A. (1993). Aspects of literary comprehension. Amsterdam, John Benjamins.

Hyun-Soon Min
Ph.D. student
Department of English, University of Memphis
Memphis, TN, USA 38152
E-mail: hyunmin@memphis.edu

Philip M. McCarthy
Chief Scientist
Decooda International (Rm 416)
516 Tennessee St.
Memphis, TN 38103
E-mail: philmccarthy1@gmail.com

Received: October 26, 2013
Revised version received: December 10, 2013
Accepted: December 13, 2013