Proceedings of the Twenty-Third International Florida Artificial Intelligence Research Society Conference (FLAIRS 2010)

Identifying Varietals in the Discourse of American and Korean Scientists: A Contrastive Corpus Analysis Using the Gramulator

Hyunsoon C. Min & P. M. McCarthy

Department of English, University of Memphis
Memphis, TN, 38152
{hyunmin, pmmccrth} @ memphis.edu

Abstract

In this study, we identify systematic discourse characteristics in the academic writings of Korean and American scientists. Specifically, we employ *Contrastive Corpus Analysis* (CCA) using the computational tool, *The Gramulator* to extract statistically improbable “n gram” features that occur across texts. The results indicate that Korean scientists use non standard *varietal forms* of English (i.e., forms that are *correct* but relatively uncommon with their American counterparts). We conclude that the Korean scientists’ use of these varietals may be a key factor in interpreting and evaluating their work as non protypical in terms of discourse style. Our findings bring to light language characteristics and a methodology that may be helpful to language learners as well as materials developers.

Introduction

An old Korean proverb posits that “Good writers must endure the pain of shedding blood and bones.” But it may be fair to suggest that even these tribulations increase when the writing has to be done in a second language. Writing 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 (Hinkel 2002). That is, Non-Native-Speakers need to go beyond general linguistic knowledge of the target language into a specific writing register corresponding to the genre in which they intend to write. As such, Non-Native-Speakers often perceive that writing in English is the most difficult task to overcome (Reid 1992).

The burden of writing in a second language is especially important for Non-Native-English-Speaking-Researchers (NNESR). For these people, writing in English is critical to their careers, because it is seldom that another language is acceptable for publications in the prestigious academic journals (Duncan et al. in press; McCarthy et al. in press). But despite this burden, NNESR have surprisingly limited availability of resources to call upon. As McCarthy and colleagues demonstrate, texts providing information on linguistic features of academic writing are rare; where they exist at all. Instead, NNESR typically have to be satisfied with sessions of proofreading from non-specialized English speakers. As a result, the writings of NNESR reveal linguistic patterns that vary from the proto-typical model of native English speaking academic writing. This difference is problematic because manuscripts submitted to prestigious journals must first be reviewed by experts in the field. And, given that NNESR are more likely to produce *non*-prototypical texts (in terms of language choice, rather than content; what we refer to here as *non standard varietals*), 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 NNESR in issues of academic writing in English, McCarthy et al. (in press) 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 deployment 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 et al. (in press) analyzed writings from American scientists, Korean scientists publishing-in-Korea, and Korean scientists 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 these studies call for further research into NNESR writing. Specifically, there is the need to isolate and analyze tangible linguistic units that distinguish the writing of NNESR from that of native English speakers (i.e., varietals). The current study addresses that call, building directly from Duncan et al. (in press) by further investigating the writing of Korean scientists publishing-in-Korea. However, instead of quantifying discourse features, the current study seeks to identify varietals within the text that may be the driving force distinguishing the work of NNESR from their native English speaking counterparts. Through such an approach, we aim to address our primary research questions: *Do Korean scientists employ distinct linguistic features in comparison to a prototypical model from American scientists?* And, if so, *Do these features offer insights for the development of facilitative resources for the writing of NNESR*?

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). Further, in identifying those characteristics, we seek to offer insights into instructional resources for NNESR. To achieve this goal, we use a textual analysis approach from the field of Second Language Learning called *contrastive corpus analysis* (CCA: Cobb 2003; Granger 1998). CCA differs from more traditional corpus analyses inasmuch 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 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. scientific writing in English by Korean scientists vs. scientific writing in English by American scientists), whatever is characteristic of one corpus, relative to the other corpus, is what defines the text type.

CCA is rapidly establishing a firm 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 (e.g., Duncan et al., in press; McCarthy et al., in press). CCA reveals pervasive yet ill-defined underlying patterns of texts that not only characterize text types, but, in doing so, highlight the language features that make such a characterization. Subsequently, these features can be used as the basis for materials for language learners. An early example of this approach 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. 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. Trebits' findings 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. And indeed, the research that forms the foundation of the current study (Duncan et al. in press, McCarthy et al. in press) used 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 NNESR writings and provide helpful information in language instructional activities.

The Gramulator

This study employs contrastive corpus analysis using the textual analysis tool *the Gramulator* (http://tinyurl.com/5bwo64). The Gramulator is an "SIF n-gram" assessment tool designed to reveal underlying textual patterns through a process of identifying statistically relevant frequency distributions. The Gramulator's central unit of analysis is the n-gram: adjacently positioned lexical items in a text. In this study, we focus on two-word n-grams (or, *bigrams*); and, more exactly, on *SIF* bigrams. SIF bigrams are lexical features that are most common to one corpus (i.e., among the 50% most frequent bigrams, excluding *hapax legomena*) but *un*common to the contrasting corpus (i.e., *not* among the 50% most frequent bigrams). In identifying SIF bigrams, we highlight the most characteristic and least characteristic language sequences of the two corpora, and, based on these results not only ascertain whether a textual features difference exists between the two text types, but also demonstrate what some of those textual characteristics are.

Corpus

Our corpus comprises 669 abstracts culled from 26 experimental scientific journals; including genres of chemistry, biology, and physics; and all published since 2001. From these texts, two individual corpora were compiled: Korean scientists in Korea (KE) and American scientists in America (AE). The KE corpus comprises Korean scientists' abstracts (N=369), published exclusively in 15 different Korean journals. The AE corpus (the prototype model) comprised American scientists' abstracts (N=300), published exclusively in 11 American journals.

To establish confidence that the authors of the relative texts were either Korean or American, the model of McCarthy et al. (in press) was followed (see also Duncan et al. in press). The model has two major criteria: First, the primary author (typically in the field of science, the person who leads the projects and writes most of the paper) and the final author (typically the supervisor) are required to be from institutes within their respective countries. And second, the primary and the final authors' names must be 'typical' of the country of the classification. That is, the primary and final authors in the KE and AE corpora represent the typical names for Koreans and Americans respectively. 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 KE corpus, one native speaker of Korean (the first author) evaluated whether the names met the criteria. For the AE corpus however, 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.

The current study follows McCarthy et al. (in press) and Duncan et al. (in press) in focusing only on the abstracts of the articles. Abstracts form a suitable point of departure because they are relatively easy to collect, 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.

Gramulator Analysis

The Gramulator reveals explicit lexical structures in the form of *statistically improbable bigram features* (SIF bigrams). These lexical structures are useful because they provide explicit information to both NNESR and materials designers. However, it is also necessary to provide quantitative analysis of SIF bigrams so as to provide sufficient confidence that the Gramulator results are characteristic of a corpus, rather than simply an oddity of relatively few examples. To this end, we begin our analysis with a quantitative account.

The 669 total texts (300 American and 369 Korean) were processed using the Gramulator. The analysis produced 382 *American SIF bi grams* and 574 Korean *SIF bi grams*. The entire array of American SIF bigrams forms an index (i.e. a measure), which we refer to as Am-SIFb; and the same is true of the Korean SIF bi-grams, referred to here as Kr-SIFb. Used as an index, the quantity of the presence of SIFb indicates the degree to which a text, or corpus, is characteristic of the register from which the SIFb was derived. Using these indices, all 300 American texts and 369 Korean texts were processed through *the Counter*, a calculator module of the Gramulator that assesses the distribution of occurrences of the SIF bigrams (i.e. Am-SIFb and Kr-SIFb). The Counter *counts* how many times each SIF bigram occurs in each text, normalizing by text length and by total number of SIF bi-grams. The output is a value for each text; and because we process each text by each index, we are provided with one value for Am-SIFb and one value for Kr-SIFb. For example, one American text produced an Am-SIFb value of 1.173, whereas the same text produced a Kr-SIFb value of 0.772. Thus, predictably, this AE text produces a higher value of American characteristics (i.e., Am-SIFb) than it does of Korean characteristics (i.e., Kr-SIFb).

We conducted a repeated measures ANOVA using the produced indices (i.e., Am-SIFb and Kr-SIFb) to better establish the distribution of the SIF bigrams across the corpora. As predicted, for the AE corpus, the within-text factors of Am-SIFb was significantly higher than Kr-SIFb (Am-SIFb: *M*=1.786, *SD* = 0.976; Kr-SIFb: *M*=1.204, *SD* = 0.798; *F*(1,299) = 242.653, *p* < .001, η^2 p = .448). Also as predicted, for the KE corpus, the within-text factors of Kr-SIFb was significantly higher Am-SIFb (Kr-SIFb: *M*=1.336, *SD* = 0.931; Am-SIFb: *M*=0.863, *SD* = 0.455; F(1,368) = 4.139, *p* = .043, η^2 p = .011). The results suggest that American SIF bigrams are more broadly employed in the AE corpus than in the KE corpus, and the Korean SIF bigrams are more broadly employed in the KE corpus than in the AE corpus.

At first blush, the ANOVA results appear unremarkable; however, individual analyses of the paired text values reveal a more complex story. Although the AE corpus analysis is as predicted, with the Am-SIFb values higher than their corresponding Kr-SIFb values for 289 of the 300 texts (96.333%, *p* < .001 if chance = 0.5), the KE corpus results do not offer the same kind of finding. For this corpus, the Kr-SIFb values were higher than their corresponding Am-SIFb values for only 132 of the 369 texts (35.772%), a result easily attributable to chance (*p* > .999 if chance = 0.5).

Closer examination of the Korean texts reveals that 367of the 369 (99.5%) files produced a positive Kr-SIFb value, meaning that at least one Korean SIF bigram was present in the text. However, of the 132 files that produced higher Kr-SIFb values, no fewer than 93 of the corresponding Am-SIFb (70.5%) recorded a zero value. The result suggests that the primary difference between the American and Korean corpora is that a large minority of Korean writers do not sufficiently employ a broad range or quantity of *American* characteristics to make their texts proto-typical. That is, there is little evidence to suggest that the identified Korean characteristics are broadly and frequently employed in Korean texts; instead, perhaps as few as a third of Korean texts generate these features. In contrast, the American characteristics are broadly and frequently employed in both the American corpus (where it is to be expected) and also across most of the Korean corpus (where it is not).

Taken as a whole, the results suggest that the majority of Korean texts (74.8%) contain target language characteristics, but seemingly not in sufficient range or quantity to match the product of their American counterparts. Meanwhile, only 25.2% of Korean texts have

Table 1: Examples of *that are* in the American corpus

disulfido ligand that bridges two Mn(CO)3 groups **that are** joined by a Mn-Mn single bond, 2.6745(5) A
interspaced short palindromic repeats (CRISPRs) **that are** a hallmark of virus resistance. Virus population
effects of chronic methamphetamine abuse **that are** obscured by suppression of cortical glucose
expectation, cells enter different quiescent states **that are** determined by the initiating signal. However,
divergent species, functional sequences **that are** degenerate or biologically redundant will

any characteristics of the target language. Moreover, nearly 36% have a high degree of Korean-English features, which does not make their English *wrong*, although it is unlikely to make their work appear prototypical.

Our analysis of the *quantitatively* derived SIF bigrams can also be interpreted *qualitatively*. Our qualitative analysis operates similarly to a factor in a factor analysis study; and, like a factor analysis, our interpretations remain open to researchers with contrasting hypotheses. With this caveat in mind, we outline here some of the SIF bigrams of interest that resulted from our analysis. Space restrictions mean that we are limited to only the most frequently employed SIF bigrams for each corpus. We use the Gramulator's *Concordancer* module to assist us in this analysis by highlighting examples of bigrams in context.

The most frequently employed Am-SIFb is *that are*. In the AE corpus, this bigram is employed for 32 instances across 26 of 300 texts (8.667%), whereas Koreans use it just 5 times across 5 of 369 texts (1.355%). The examples (see Table 1 above) suggest a greater American use of relative pronouns, and even when the bigram of *that are* is extended to include instances of *which are* (note that the Korean language does not distinguish the restrictive and non-restrictive relative clause), the difference is maintained (American: total = 42, texts = 34 [11.3%]; Korean: total = 18, texts = 18 [4.9%]; $F(1{,}667) = 69.445$, $p < .001$; $\eta^2 p$ = .094). The result might be explained by *avoidance* on the part of the Koreans (i.e., avoiding the complex structure of a relative pronoun when a more simple, although less common, strategy is available). However, if the Koreans are avoiding relative pronouns, the question becomes, *what language patterns are they using to convey their ideas?* To address this question, we formed two hypotheses. The first is the *use connectives instead* hypothesis, which predicts that Korean will use more connectives than Americans to compensate for complex syntax. The second hypothesis is the *use shorter sentences* hypothesis, which predicts that Koreans will have a tendency to simply convey their ideas with simpler (and therefore shorter sentences). ANOVA results showed no significant differences between AE and KE for the *use connectives instead* hypothesis, but there was evidence supporting the *use shorter sentences* hypothesis (Korean: M = 24.033, SD = 6.563; American = 25.604, SD = 5.299; F (1,667) = 11.231, p < .001, $\eta^2 p$ = 0.017). Taken as a whole, the analysis suggests that American writers are more likely to use longer and more syntactically complex sentences.

The second-most employed Am-SIF bigram is *patients with*. This bigram is employed for a total of 26 instances across 17 of 300 texts (5.7%). The bigram does not occur at all in the KE corpus; however, the unigram *patients* occurs 15 times across 6 of 369 texts (1.6%). The difference here (see Table 2) might be an American preference for the structure *patients with* + *issue* (100%) and a Korean preference for a compound of *issue* + *patients* (78.6%). We speculate that the choice of structures may be explained by inter-language transfer. That is, Korean uses only the structure of *issue* + *patients*; there is no equivalent structure of *patients with* + *issue*. Indeed, in Korean, all modifiers precede that which is modified.

The third-most employed AE bigram is *the human*, featuring 25 times across 21 texts (7.0%). In the KE corpus, the bigram occurs just 4 times (1.1%), although there are 52 cases across 21 texts for the unigram *human* (8.4%). A likely candidate for explaining the difference is prepositional avoidance. Specifically, Americans use *the human* as part of a prepositional phrase in 92% of cases, whereas this is the case for just 1 in 5 (20%) of the Korean examples (see Table 3). However, closer examination suggests a more complex story. In the KE corpus, the unigram *human* occurs no fewer than 19 times in prepositional phrases (36.539%), perhaps indicating that *determiner avoidance* trumps prepositional avoidance.

Table 2: Contrasting examples of the use of *patient* in the AE and KE corpora

AE	accurate disease staging of **patients with** pancreatic cancer is essential to divide
AE	has decreased from 85 to 29%, and that in **patients with** central nervous system disease has decreased
AE	for further improvements in outcome for **patients with** this potentially devastating disease lie
KE	sera of 177 active pulmonary tuberculosis **patients** and 323 healthy individuals revealed that the
KE	virus (hbv) or by superinfection of hbv **patients**. to date, there is no vaccine available for
KE	were obtained from rheumatoid arthritis **patients**, and flss were isolated. the cells were stimulated

Table 3: Contrasting examples of the use of *the human* the AE and KE corpora

A	for a mutation in a zebrafish paralog of **the human** and mouse tumor suppressor gene
A	comprehensive view of promoter function in **the human** genome
A	nine base-pair genomic "words" throughout **the human** genome. results identify previously unknown
K	those transgenic rice lines that expressed **the human** cytokines in small quantities were able to survive
K	transformants used in this assay contain **the human** estrogen receptor along with the appropriate
K	interaction between the endocrine disruptors and **the human** hormone system.

Turning to the Korean SIF bigrams, the two most commonly employed examples are *in order* and *order to*. These examples combine to form the trigram of *in order to*, employed by Koreans for 31 instances across 31 of 369 texts (8.4%). In contrast, Americans employ the trigram for 4 instances across 4 of 300 texts (1.3%). Closer examination reveals that *in order to* may be more informative as a form of quadgram (i.e., [period] + *in order to*). In this sentence-opening-role, the quadgram appears for 14 instances across 14 texts (3.8%), whereas it is used just once by Americans. We hypothesized that *[period] + in order to* was employed by Koreans where Americans would use just [period] + *to*, (if what followed the construction was a verb phrase). The results confirmed our hypothesis with Koreans employing the quadgram for 26 instances across 34 texts (76.47%), whereas American used the shorter alternative (i.e., [period] + *to*) 28 times across 30 texts (93.333%). The result suggests that Koreans may be employing *self grounding* (see Langacker 1991). That is, longer (and more complex) language is chosen to firmly impress upon the reader the backdrop against which the case will be made. A similar argument for *reverse avoidance* is also possible. That is, the writer is over-compensating for fear that the potentially more ambiguous shorter version will be misunderstood. A third and more pragmatic view is simply that English language text books focus first on the fullest form, therefore making it the most likely to be produced.

Our final KE example is the SIF bigram *to investigate*. Koreans employ this infinitive 30 times across 29 of 369 texts (7.9%). In contrast, Americans use it just twice across 300 texts (0.7%). We hypothesized that Koreans may be employing a *heavier is better* preference, similar to the way native English speakers may prefer the heavier form of *usage* over *use*. The results did not support our hypothesis, Koreans produced four instances of *to examine*, whereas Americans produced just two. Indeed, finding any alternative for Americans proved difficult: using a combination of *to investigate*/*to examine*/*to explore*/*to look into*/*to study* a total of just 9 instances occurred in 9 of the 300 American texts (3%). We also reduced the bigram examples to unigrams (i.e., *investigate* and *examine*) hypothesizing that [pronoun] + finite verb may be more common to Americans. However, even here we find higher Korean use (Korean: investigate = 92 instances across 84 of 369 texts [22.8%]; examine = 55 instances across 52 texts [14.1%]; American: investigate = 19 instances across 18 of 300 texts [6%]; examine = 22 instances across 20 texts [6.7%]. Looking within the Korean examples of *to investigate* we found a possible explanation. Specifically, nine of the *to investigate* examples were preceded by *this study was* (see Table 4). The result suggests a form of *flexi gram* consisting here in all nine cases of the trigram *this study was* + [verb meaning *do*] + the bigram *to investigate*. We speculate that such flexi-grams may be characteristic of non-native speakers' discourse. We also speculate that such flexi-grams may be accepted by the discourse community as "not wrong," even if the use is non proto-typical. Characteristics such as these might be explained in several ways. First, in terms of inter-language transfer, the style may simply be more *Korean like*. Second, the use of past tense may simply reflect differing convention styles for abstracts. And third, the use might be an example of *self grounding, over compensating*, or *reverse avoidance*: that is, the *inclusion* of non proto-typical language for the (unconscious) purpose of foregrounding a topic or avoiding potential misunderstanding. Obviously, a combination of these elements is also possible.

The Kr-SIF bigrams and related language characteristics identified here need to be distinguished from performance 'mistakes' and competence 'errors' (Corder 1967). Both terms imply that something is 'wrong,' 'confusing' or 'inappropriate' with the language used, whether in the sense of grammar, typography, or syntax. Similarly, we do not use the word *over generalization* for Korean SIFs because, again, the implication is that the use is often 'wrong.' We also distinguish these SIFs from collocations, arguing that collocations imply idiomatic usage, and that they never exist across phrase boundaries as they do here in examples such as *this study was*. Indeed, the strongest reason for emphasizing that Korean characteristics are not "wrong" is because 90% of them are *also* used by Americans. What appears to be problematic, however, is the breadth and quantity of the Korean use of these characteristics, and the fact that a semantically equal alternative is available. More

Table 4: Four examples of the flexi-gram *this study was* + (synonyms) + *to investigate*

this study was performed to investigate the effects of glycine on the development
this study was conducted to investigate the sex pheromone composition of the variegated
this study was carried out to investigate the proximate composition, soluble sugar,
this study was aimed to investigate the nitric oxide (no)-induced cytotoxic mechanism

importantly, this alternative is the form primarily selected by native speakers. Thus, we use the term *non standard varietal* to describe the Korean characteristics, and we use *standard varietal* to describe the corresponding American language. More specifically, we define *non standard varietal* as "'An example of self-grounding, avoidance, or language transfer manifested by the common (although maybe unconscious) employment of correct but non proto-typical language, presumably because the language selected is simpler to construct or less likely to be misinterpreted."

Discussion

This study used the Gramulator to conduct a contrastive corpus analysis. Its purpose was to reveal systematic linguistic features in the academic writings of Korean scientists as compared to the academic writings of American scientists. In so doing, this study aimed to inform Korean researchers and prospective materials 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. This study addressed two central research questions: *Do Korean scientists employ distinct linguistic features in comparison to a prototypical model from American scientists?* And, if so, *Do these features offer insights for the development of facilitative resources for the writing of NNESR*? Addressing the first question, our response is that Korean scientists appear to employ acceptable but less commonly employed structures (i.e., non-standard varietals). Infrequent use of such structures appears to be acceptable; however, in combination such variants may signal a discourse style that is *non* proto-typical. To address the second question, our response is that the Gramulator and our contrastive corpus analysis approach have revealed numerous avenues of interest for materials developers. Further analysis will be required to fully reveal the breadth of what we have termed here as *standard* and *non standard varietals*, and further experimentation will be required to assess whether changes made to texts as a result of such studies has an effect on reviewers and the subsequent success of non native English speaking researchers.

Acknowledgments

This research was supported by the Institute for Education Sciences (Grants R305GA080589, R305G020018 02, and R305G040046), by the National Science Foundation (Grant IIS 0735682) Department of Defense/CIFA (H9C104 07 C0019). The authors also acknowledge Charles Hall, Diana Lam, George Min, Hojin Lee, John Myers, Rebekah Guess, Scott Healy, and Travis Lamkin.

References

Biber, D., Conrad, S., and Reppen, R. 1998. *Corpus linguistics: Investigating Language Structure and Use*. Cambridge, UK: Cambridge University Press.

Cobb, T. 2003. Analyzing Late Interlanguage with Learner Corpora: Québec 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.

Corder, S.P. 1967. The significance of learners' errors. *International Review of Applied Linguistics* 5: 161-170.

Duncan, B., McCarthy, P.M., Hall, C., and McNamara, D.S. in press. Language Variation among Biomedical Abstracts. *TESOL Journal*.

Gamon, M., Leacock, C., Brockett, C., Gao, J., and Klementiev, A. 2009. Using statistical techniques and web search to correct ESL errors. *CALICO Journal* 26 (3): 491-511.

Glanville, R., Sengupta, S., and Forey, G. 1998. A (cybernetic) musing: Language and science and the language of science. *Cybernetics and Human Knowing* (5).

Granger, S. eds. 1998. *Learner English on computer*. London: Longman.

Hinkel, E. 2002. *Second language writers' text linguistic and rhetorical features*. Mawhwa, NJ: Lawrence Erlbaum Associates.

Langacker, R (1991). Concept, Image, and Symbol. Berlin: Mouton de Gruyter.

McCarthy, P.M., Hall, C., Duran N.D., Doiuchi, M., Duncan, B., Fujiwara, Y., and McNamara, D.S. in press. A computational analysis of journal abstracts written by Japanese, American, and British scientists. *The ESPecialist*.

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.

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