the ESPecialist, vol. 30, nº 2 (141-173) 2009 ISSN 0102-7077

ANALYZING JOURNAL ABSTRACTS WRITTEN BY JAPANESE, AMERICAN, AND BRITISH SCIENTISTS USING COH-METRIX AND THE GRAMULATOR

Análise de Resumos de Periódicos Escritos por Cientistas Japoneses, Americanos e Britânicos com Coh-metrix e Gramulator

Philip M. McCarthy (University of Memphis, Department of English, Memphis, USA)
Charles Hall (University of Memphis, Department of English, Memphis, USA)
Nick D. Duran (University of Memphis, Department of Psychology, Memphis, USA)
Maki Doiuchi (University of Memphis, School of Audiology and Speech-Language Pathology, Memphis, USA)
Yuko Fujiwara (University of Tennessee, Department of Physiology, Knoxville, USA)
Benjamin Duncan (University of Rochester, College Writing Program, Rochester, USA)
Danielle S. McNamara (University of Memphis, Department of Psychology, Memphis, USA)

Abstract

This interdisciplinary study comprises two complementary analyses on a corpus of journal abstracts written in English by American, British, and Japanese scientists. The first analysis uses the computational tool Coh-Metrix to assess text at the discourse level. The second analysis uses the computational tool the Gramulator to compare the frequency of n-grams across the three sources of abstracts. The Coh-Metrix and Gramulator analyses both suggest significant differences between all three varieties of English. The greatest differences were apparent when comparing abstracts written by Japanese and English speakers; however, a number of differences were also apparent when comparing the British English and American English varieties. The results lend weight to the conclusion that native-English speakers (reviewer, editor, or reader) of either the British or American variety may interpret Japanese-English texts as *lacking in key areas of the proto-typical style of the English register. Our findings provide information for instructors, course developers, and scientists on how and where text might be modified in order to facilitate the production of more native-English-like representations.*

Key-words: *Coh-Metrix; Gramulator; corpus; discriminant analysis.*

Resumo

Este estudo interdisciplinar engloba duas análises complementares de um corpus de resumos de periódicos escritos em inglês por cientistas americanos, britânicos e japoneses. A primeira análise emprega a ferramenta computacional denominada Coh-Metrix para avaliar o texto em nível discursivo. A segunda análise emprega a ferramenta computacional denominada Gramulator para comparar a frequência de n-gramas nas três fontes de resumos. Tanto a análise com o Coh-Metrix como a análise com o Gramulator sugerem diferenças significativas entre as três variedades de inglês. As maiores diferenças vieram à tona ao comparar os resumos escritos por falantes de inglês e japonês; entretanto, algumas diferenças também foram observadas ao comparar o inglês britânico ao inglês americano. Os resultados contribuem para a conclusão de que os falantes nativos de inglês (críticos, editores ou leitores), tanto de sua variedade britânica como americana, podem considerar que os textos em inglês escritos por japoneses são deficientes em pontos-chave do estilo prototípico do registro em inglês. Nossos achados proporcionam informações para instrutores, criadores de cursos e cientistas em relação ao modo e aos pontos em que o texto poderia ser modificado para facilitar a produção de representações mais próximas das características apresentadas pelos falantes nativos de inglês.

Palavras-chave: *Coh-Metrix; Gramulator; corpus; análise discriminante.*

1. Introduction

Unfortunately, the prime directive of academia "Publish or Perish" fails to mention that publications, in order to count, should be *in*

English. In the majority of science fields, the most prestigious journals and especially those with the highest impact only publish in English. It is no surprise, therefore, that between 74% and 90% of scholarly work is now published in English (Lillis & Curry, 2006). Indeed, as these percentages appear to be growing (Canagarajah, 1996, 2002; Flowerdew, 1999; Gibbs, 1995; Jernudd & Baldauf, 1987; Tardy, 2004), they self-serve the impetus for even greater centralization on the English language as *the* language of academic research. There are, of course, respectable and prestigious journals that are published in other languages (Belcher & Connor, 2001; Canagarajah, 2002; Curry & Lillis, 2004); however, success in most fields depends on publishing in English; and as a consequence, most non-native English speaking academics submit their research for publication in English.

There are many issues in submitting a manuscript in English. First, do British and American journal editors and reviewers consider language differently? And, if so, should researchers decide where to submit their manuscript on whether their English is more similar to British or to American? This issue is important because, the degree to which an English-language text differs from an expected model (British-English or American English for example) may negatively affect the chances of the non-native English speakers having their manuscripts accepted (Flowerdew, 2001; Hewings, 2006). Our work addresses this issue by assessing English language abstracts written by Japanese scientists. The purpose of our study is to evaluate the degree to which and how Japanese-English texts differ from the text of their British and American counterparts. Our aim in highlighting such differences is to provide information regarding textual features that may facilitate non-native English speakers' writing so as to more closely mirror the writing of native English speakers. Moreover, the information gained may guide ESP [English for Specific Purposes] specialists as they work with non-native writers and develop curricula and materials.

To examine the question of whether there are explicit textual features that distinguish between writers from different countries, we analyzed three corpora of texts written by *native-* and *non-native* English speaking scientists: Japanese scientists, American scientists, and British scientists. We used the computational tool, Coh-Metrix (Graesser, McNamara, Louwerse, & Cai, 2004) to analyze these corpora at the discourse level and we conducted n-gram analyses (i.e. the two-word bigram and the three-word trigram) using the *Gramulator* tool (https://umdrive.memphis.edu/pmmccrth/public/index.htm) to analyze the corpora at the word level.

2. Motivation for the study

Our work stems from an informal question from a co-author of this study, Dr. Yuko Fujiwara, a *Japanese* bio-chemist. She felt that an analysis of her English writing might help her (and her Japanese colleagues) to more effectively present their research. Dr. Fujiwara explained that there are few official courses or seminars on writing scientific papers in English. Learning the register of scientific writing, she maintained, generally involved a laboratory member conducting a seminar in which the scientists chose a typical paper, translated it into Japanese, and discussed it, allowing the scientists to become familiar with representative expressions. The scientists were expected to copy those representative terms in their own papers.

Dr. Fujiwara stressed that the many books and native instructors that deal with EAP [English for Academic Purposes] rarely seem to be able to deal with the specialized language of scientific journals; even the professional native-English speaking proofreaders hired are often unspecialized in the discourse community. Finally, Dr. Fujiwara explained that after submission, the problem shifts from writing to dealing with reviewers' comments. Typically, the response is "revise and resubmit." Culturally, she explained that resubmitting a manuscript entails commenting on reviewer comments. Such a practice is seen by many Japanese as aggressive, immodest, and shameful. As a result, Japanese scientists are unlikely to resubmit such an article. As a consequence, Japanese scientists are more likely to opt for a journal of lower impact in order to avoid confrontation. Such choices, of course, can greatly affect both careers and entire lines of research.

Although Dr. Fujiwara's description is anecdotal, her claims are supported by theoretical and empirical research. For example, studies such as Gibbs, Kendall, and Pagel (2002) and Hewings (2006) discuss where non-native speakers have problems getting their work published, especially in the field of science. These studies confirm Dr. Fujiwara's experience, identifying the issue of communicating with journal editors and responding to reviewers' comments as particularly troublesome.

3. Analyzing composition

We need to begin our analysis by discussing basic differences between the Japanese written styles and the British/American styles. For example, Dennett (1988) argued that the Japanese rhetorical structure aims at a goal quite different to the Aristotelian model, which underlies the British/American writing styles. Even in technical writing, Japanese writers consider the importance of such elements as *beauty* and *surprise*, elements a Western audience might consider unnecessary and confusing. Hinds (1990) also argued that the expository styles of Japanese writers were so different from those of their English speaking counterparts that native-English speakers evaluated English language texts written by Japanese as less coherent than similar texts written by native English speakers. Hinds concluded that the differing rhetorical principles were responsible for these differences in ratings.

Such subtle differences in writing style and culture have led many academics (e.g., Kaur & Sook, 2005) to call for greater attention to writing in English for Specific Purposes (ESP) classrooms, particularly in areas such as register awareness (Bhatia, 1997). Such an attention to register is often overlooked with the assumption that a field *in general* (e.g., science) is sufficiently representative of *a specific field* (e.g., biochemistry manuscripts). But empirical studies have shown that there are significant differences in text representation even between registers regarded as highly similar (Biber, 1988; Duran, Graesser, McCarthy, & McNamara, 2007; Hall, McCarthy, Lewis, Lee, & McNamara, 2007; Louwerse, McCarthy, McNamara, & Graesser, 2004). Studies such as these have paved the way for closer analyses of text through computational and statistical approaches. In this study, we build on such research by analyzing the texts of Japanese, American, and British scientists at the discourse level using the computational tool, Coh-Metrix, and also through a n-gram analysis of texts at the word level using a freely available on-line tool that we have developed, *the Gramulator*. Our study begins with the Coh-Metrix discourse analysis.

4. Coh-Metrix

Traditional approaches to the study of natural language typically do not extend beyond word-level features (e.g., grammatical class and frequency). This shallow analysis can be problematic, because possible differences in native language category texts are predicted to occur at the higher-order text components involved in cohesion and rhetorical style. As such, a much broader analysis is needed, one that also takes into account global text attributes and conceptual information. Advances in computational linguistics make it possible to collect comprehensive profiles of language and cohesion features (Jurafsky & Martin, 2000). At the forefront of the new computational techniques is a freely available, web-based software tool called Coh-Metrix (see cohmetrix.memphis.edu).

Coh-Metrix harnesses sophisticated developments in computational linguistics and discourse processing, featuring advanced syntactic parsers (Charniak, 1997; Sekine & Grishman, 1995), part-of-speech taggers (Brill, 1995), and Latent Semantic Analysis (LSA), (Landauer & Dumais, 1997; Landauer, McNamara, Dennis, & Kintsch, 2007). Word relationship indices are derived from the WordNet lexical database (Miller, 1990), and conceptual information from the MRC database (Coltheart, 1981). A variety of shallow metrics such as Flesch-Kincaid Grade Level (Klare, 1974/1975) are also added for purposes of comparison. These modules are integrated into the automated Coh-Metrix tool and used to generate over 600 indices of language, text, and readability (Graesser et al., 2004). Coh-Metrix has been involved in many research endeavors, ranging from learning assessment (Best, Rowe, Ozuro, & McNamara, 2005) to distinguishing segments of texts by their functional and rhetorical relationship (McCarthy, Briner, Rus, & McNamara, 2007). These successful applications of Coh-Metrix to rhetorical and linguistic variation provide us with a rich array of support in our current analysis.

Seven sets of *Coh-Metrix* measures were selected for this study. These metrics included 1) Word Information and Frequency, 2) Incidence of Part of Speech and Phrases, 3) Connectives, 4) Syntactic Complexity, 5) Event-Indexing Features, 6) Lexical Diversity, and 7) Coreference. Each metric is presented in an order that corresponds to information at the word, inter-clause, inter-sentence, and inter-paragraph level.

Word Information and Frequency. Coh-Metrix computes word information and frequency scores from established psycholinguistic and corpora analyses. The MRC database (Coltheart, 1981) a collection of human ratings of 150,837 words along four psychological dimensions: meaningfulness, concreteness, imaginability, and familiarity. Coh-Metrix derives scores for word abstractness and ambiguity by incorporating a module called WordNet (Fellbaum, 1998; Miller, 1990), an online lexicon tool that groups words into sets of synonyms that are connected by semantic relations.

Word frequency refers to the likelihood of a word being familiar to a reader due to its frequency in the world and subsequent likelihood of having been previously encountered by a reader. Word frequency is generally based on the frequency of words in a large corpus of printed texts.

Incidence of Part of Speech and Phrases. Coh-Metrix profiles the part of speech (POS) for every word contained in a text. There are over 50 POS tags derived from the Penn Treebank (Marcus et al., 1993). Coh-Metrix incorporates a sophisticated natural language processing tool, the Brill (1995) POS tagger, for assigning POS tags to each word. This assignment allows for an *incidence score* of POS categories, calculated as the occurrence of a particular category per 1,000 words. The incidence score is a useful index for identifying substantial linguistic features.

Connectives. Connectives help to increase the cohesion of a text by explicitly linking ideas at the clausal and sentential level. Cohesion is important as it has been shown to facilitate both comprehension and learning (Halliday & Hasan, 1976; Graesser et al., 2004). In Coh-Metrix, connectives are calculated in correspondence to subcategories of cohesion identified by Halliday and Hasan (1976) and Louwerse (2001), such as positive additive cohesion (e.g., *also, moreover*) or negative additive cohesion (e.g., *however, but*). Logical operators (e.g., variants of *or*, *and*, *not*, and *if–then)* are also cohesive links that influence the analytical complexity of a text.

Syntactic Complexity. Coh-Metrix measures syntactic complexity by analyzing the structural representation of a sentence in a parse tree. The Charniak (1997) parser is used to generate a tree from an underlying formal grammar. The parse tree is interpreted by Coh-Metrix with the assumption that syntactic complexity is characterized by a greater degree of embedded phrases, dense syntactic structure, and ambiguous syntax (Graesser et al., 2004). Finally, Coh-Metrix provides an estimate of the number of sentences with similar syntactic structure. A high score for syntactic similarity indicates consistency in style and form.

Event-Indexing Features. An important goal in reading comprehension is to build a mental representation of the state of affairs described in a text. The Event-Indexing Model (Zwaan, Langston, & Graesser, 1995) posits that readers monitor story events and link them to each other according to continuities in time, space, and causality. These continuities are made explicit by textual features that guide the "when" (i.e., *time*), "where" (i.e., *space*) and "why" (i.e., *causality*) of event integration. Coh-Metrix measures the relevant event-indexing features by calculating a variety of *repetition* (e.g., for temporality), *ratio* (e.g., for spatial features), and *incidence* (e.g., for causality) scores.

Lexical Diversity. Lexical diversity (LD) measures the range of vocabulary deployed by a speaker or writer. Greater LD is widely held to be indicative of greater linguistic skills, speaker competence, a speaker's socioeconomic status, or even textual difficulty (Avent & Austermann, 2003; Grela, 2002; McCarthy, 2005; Ransdell & Wengelin, 2003). Coh-Metrix offers several LD indices including M (Maas, 1972), K (Yule, 1944), and D (Malvern, Richards, Chipere, & Durán, 2004).

Coreference. Lexical coreference is an approximation of the conceptual redundancy between sentences. Coh-Metrix tracks four major types of lexical coreference: *noun overlap*, *argument overlap*, and *stem overlap*. Noun overlap is a proportion of all sentence pairs that share one or more common nouns. Argument overlap is a proportion of all sentence pairs that share common nouns or pronouns (e.g. *table/table* or *table/tables*). And stem overlap is the proportion of sentence pairs where a noun is in common with the same word of any grammatical category (e.g. the noun *photograph* and the verb *photographed*).

Coh-Metrix also assesses conceptual overlap between sentences by a sophisticated computational model for word meaning, Latent Semantic Analysis (LSA) (Landauer & Dumais, 1997, Landauer, McNamara, et al., 2007). LSA represents word meaning by monitoring the type of contexts that words tend to occur. The basic premise of the model is simple: words that share similar contextual histories will be more similar in associative relationships. For example, the word *hammer* will be highly associated with words of the same functional context, such as *screwdriver*, *tool*, and *construction*.

5. The corpus

The corpus for this study comprised 602 abstracts downloaded from www.pubmed.gov. In total, 418 science journals were used, an average of 1.44 abstracts per journal, with no more than 6 abstracts taken from any one journal[1]. Once collected, a simple Visual Basic program separated the text of the abstracts from extraneous textual information (e.g. author information, affiliations) and the abstracts were automatically saved as TXT files, suitable for computational processing. Each native language category (Japanese-English, British-English, American-English) was represented by at least 200 texts. We employed two criteria for text classification. First, all of the authors were required to be from an institute within the country of classification. Second, each

[1] A complete list of journals used in this study can be found at http://tinyurl.com/6jm98t (last accessed on 12/09/09).

authors' name was required to be 'typical' of the country of classification. While such a technique is by no means perfect, we argue that the majority of texts will be appropriately categorized using these guidelines.

Our decision to include both British and American corpora stems from research suggesting that the two varieties differ significantly. The research begins with Biber (1987), who contrasted nine written genres, finding evidence that British texts are more formal but less interactive and abstract than those written by their American counterparts. But perhaps of most relevance to our present study is Hall et al. (2007). They compiled an American/British corpus in the specific language register of Legal English. Their study also used Coh-Metrix and found substantial difference across a wide range of discourse level variables, with results indicating that the British cases were more cohesive than the American cases. The results brought into question common assumptions (e.g. Johansson, 1985) as to similar genres varying little along language variety lines.

Such research suggests that not only might Japanese-English texts differ from native-English speakers' text, but that the native speakers' texts themselves (British and American texts) may also differ significantly. The question then becomes which of the two native-English varieties do the Japanese English texts more closely resemble? Thus, our study may shed light on which variety more closely reflects the English variety of native Japanese speakers. Such results could influence subsequent decisions as to which journals Japanese scientists might more successfully submit their manuscripts.

6. Coh-Metrix corpus analyses

For the Coh-Metrix analysis, the corpus was split into two random and approximately equally sized groups: the *training set* and the *test set*. The purpose of the training set was to identify which of the 400 Coh-Metrix variables best distinguished the three language categories: American-English, Japanese-English, and British-English. We would then use those selected variables to create a model to predict language categories. This model would be generated through a discriminant function analysis (detailed below). The accuracy of the model creaed by the discriminant analysis would then be assessed using the data held back in the test set.

Because each set (*training* and *test*) contained approximately 300 texts, we assumed a maximum of 15 variables could be selected for the analysis before concern for over-fitting the model occurred. Such a ratio (20:1) is typical of statistical analyses of this kind (e.g., Duran et al., 2007; McCarthy, Lewis, Dufty, & McNamara, 2006; Tabachnick & Fidell, 1989). Over-fitting is a concern because attempting to use too many variables in a complex model may result in fitting not just the *signal* of the predictors but also the unwanted *noise*. The effect of over-fitting typically results in a training model that fits the data well but when applied to new data (the *test set* in this case or any subsequent research data sets) the fit would lack accuracy because noise (by definition) will not be the same from data set to data set. Thus, to select 15 variables when multiple possible variables were available, the following procedure was undertaken. As this study hoped to shed light on areas of language that differed according to native language categories, we selected variables from each of the seven categories described above. Variables were selected based on results from an Analysis of Variance (ANOVA). Using native language categories as the between groups factor and each of the Coh-Metrix indices as the dependent variables, the resultant univariate F-values were ordered by effect size and the variable with the highest F-value was selected to represent its relevant category.

To obtain the other eight predictors variables (of the total 15 we had decided to allow ourselves), all the remaining variables were ranked in terms of F-Value. Unfortunately, we could not simply take the eight highest remaining variables because that would run the risk of incurring problems of collinearity. Collinearity refers to instances when two or more variables correlate at approximately *r* => .90. Such an outcome means that it is difficult to know which of the variables is contributing to the model; thus, interpretation of the data becomes difficult (Brace, Kemp, & Snelgar, 2006; Tabachnick & Fidell, 2001). An even more important concern with collinearity is that using two or more very similar variables wastes potential model power that could be capitalized on by variables measuring some other aspect of language difference. In this study, we take a conservative stand on issues of collinearity and ensured that no index pair correlated above *r* => .70 (e.g., Duran et al., 2007; McCarthy, Lewis, et al., 2006). As such, if the correlation between any two variables was *r*=>.70, then the variable with the weaker univariate relationship was removed. This process was continued until the eight further variables had been obtained. The final 15 selected variables are shown in Table 1. (Some examples demonstrating these variables are provided in the appendix).

	Means and Standard Deviations				
Predictor	American	Japanese	British	F	η^2
Verb Phrase Incidence	136.53 (39.78)	126.19 (25.36)	150.90 (31.78)	14.63 **	0.09
POS Incidence 3rd Person Verbs	11.63 (11.08)	6.71 (6.28)	13.38 (13.05)	10.89 **	0.07
Celex Frequency Value	2.75 (0.17)	2.84 (0.14)	2.82 (0.15)	10.36 **	0.07
POS Incidence Gerunds	16.11 (14.49)	10.58 (7.37)	17.22 (13.06)	8.85 **	0.06
POS Incidence Cardinal Numbers	35.03 (32.09)	52.67 (42.61)	33.14 (33.16)	8.75 **	0.06
Word Polysemy	2.83 (0.44)	2.78 (0.40)	3.01 (0.38)	8.51 **	0.05
SD Word Familiarity	1019.77 (316.23)	1168.02 (268.48)	1044.28 (256.26)	7.92 **	0.05
POS Incidence Infinitive-*to*	18.81 (12.27)	14.22 (8.59)	19.93 (11.27)	7.92 **	0.05
Lexical Diversity Index of D	83.09 (30.57)	76.45 (26.36)	90.26 (28.61)	5.93 *	0.04
POS Incidence W-adverbs	1.10 (2.77)	0.73 (2.02)	2.19 (4.37)	5.60 *	0.04
Tense and Aspect Repetition Index	0.69 (0.66)	0.87 (0.12)	0.83 (0.15)	5.32 *	0.03
Intentional Event Incidence	7.32 (9.96)	4.05 (5.31)	5.03 (6.99)	4.68 *	0.03
Location and Motion Ratio Scores	0.49 (0.74)	0.70 (0.33)	0.65 (0.42)	4.39 *	0.03
Argument Overlap	0.76 (0.30)	0.85 (0.18)	0.85 (0.23)	4.31 *	0.03
Sentence Syntax Similarity	0.08 (0.03)	0.09 (0.03)	0.08 (0.03)	3.79 *	0.02

Table 1: Results of ANOVA for the 15 Leading Predictors (Note: ** significant at *p* < .001; * significant at *p* < .05)

To more closely assess where differences lay between the native language categories, we conducted a *post-hoc* Bonferroni analyses. Such a test highlights the degree to which the relative native language categories differ and the direction of those differences (see Table 2).

	Japanese-American	Japanese-British	American–British
Verb Phrase Incidence	10.342	-24.713**	-14.370*
POS Incidence 3rd Person Verbs	4.915*	-6.673**	-1.758
Celex Frequency Value	-0.094**	0.016	-0.078*
POS Incidence Gerunds	5.527*	-6.635**	-1.108
POS Incidence Cardinal Numbers	-17.641*	19.529**	1.888
Word Polysemy	0.057	-0.229**	-0.171*
SD Word Familiarity	-148.252**	123.735*	-24.516
POS Incidence Infinitive-*to*	4.589*	-5.714**	-1.125
Lexical Diversity Index of D	6.641	-13.812*	-7.17
POS Incidence W-adverbs	0.373	-1.463*	-1.09
Tense and Aspect Repetition Index	-0.174*	0.039	-0.135*
Intentional Event Incidence	3.277*	-0.983	2.293
Location and Motion Ratio Scores	-0.211*	0.041	-0.17
Argument Overlap	-0.092*	0.006	-0.086*
Sentence Syntax Similarity	-0.011*	0.01	-0.001

Table 2: Bonferroni *Post-hoc* Analysis Showing Direction of Differences Between Languages (Note: ** *p* < .001; * *p* < .05)

The results of the Bonferroni analysis suggested that the majority of significant differences are between the native-English registers and Japanese-English (see Table 2). Such a result is not surprising given that both American scientists and British scientists are native speakers of English. However, as can be seen in the table, there are also five significant differences between the native language categories of American and British.

Locational Incidence. The negative value for the Locational index for American-Japanese differences informs us that American writers use significantly fewer locational items (e.g., *here*, *there*) than do their Japanese counterparts. However, there are no significant differences between the remaining comparisons.

Intentional Incidence. American writers appear to use far more intentional items (e.g. *drop*, *mix*) than do their British or Japanese counterparts. Intentional items are explicit markers of causality, widely held to be of significant importance for reading comprehension

(Zwaan, Langston, & Graesser, 1995). While it is interesting to note that British writers appear to circumvent such explicit cues, we may posit that Japanese writers could benefit from greater use of such textual elements.

Argument Overlap. The results for argument overlap suggest that both Japanese and British texts have significantly greater referential cohesion than American texts. The results are similar to those of Hall et al. (2007) who also founded greater cohesion in British texts. Such results suggest that British and Japanese writers employ greater redundancy than do American writers.

Temporal Incidence. As with the argument overlap measure, both Japanese and British writers appear to use more temporal cues in their texts. Thus, the results suggest that American writers organize their text more in terms of causal relations whereas the British and Japanese organize their texts more along co-referential and temporal relations.

Polysemy. The results generated for the Polysemy index suggest that British writers use significantly more high polysemy words than either of their counterparts. This result suggests that both American and Japanese scientists prefer to use concrete terms of low ambiguity.

Syntax. The results suggest that Japanese scientists may write significantly more syntactically similarly constructed sentences than American writers. One possible explanation of this result is that Japanese writers may feel less able to express their ideas in a variety of ways, preferring to stick to a structure they know reasonably well. This hypothesis is further supported by results on 3rd person-singular (see below).

Word Frequency. The results for word frequency suggest that American scientists use significantly more low frequency words than do either Japanese or British scientists. Such results suggest that Americans may assume an audience more familiar with specialized terms, whereas scientists from the other native language categories may be taking more care in the choice of lexicon.

Parts of Speech (3rd Person singular). The results for incidence of 3rd person-singular may be the most significant indicator of native language categories. Both American and British scientists use significantly more 3rd person-singulars than their Japanese counterparts. The result is predictable as the 3rd person-singular morpheme (the *–s* on *walk* in "he walks") is semantically empty and an English grammatical anachronism that non-native speakers have always struggled to master. It is not, therefore, surprising to learn that Japanese authors may be avoiding present tense usage (where the morpheme would occur) and instead are more likely to report their results in a steady flow of a more regular tense (such as the past tense). Such a conclusion is supported by the Syntax index described above. Thus, while the Japanese text may remain grammatically accurate, it is possible that the style (non-native like) may adversely affect the reading and the paper. To support this claim, we collected 50 Japanese language science journals[2] and calculated the choice of tense use in abstracts from those papers. Only 22% of the papers used past tense (22% present, 56% both).

While further research is necessary to assess tense choice, these initial results suggest that Japanese authors may be selecting tense based on convenience or simplicity rather than prototypical form.

Parts of Speech (Cardinal Numbers). The results for use of cardinal numbers are another significant indicator of where Japanese scientists may be significantly differing from their native English speaking counterparts. The Japanese appear to be relying a great deal on the use of numbers in their abstracts, an aspect that both Americans and British scientists appear to avoid.

Verb Phrase Incidence. The Verb Phrase Incidence index suggests that British writers put a heavier reliance on verb phrases. We can presume that a number of verb phrases across sentences correlates with the number of propositions. Such an outcome suggests that this feature may make British-English sentences more complex and subsequently harder to process (Kintsch, 1988; Kintsch & Vipond, 1979).

Word Familiarity. The familiarity variable used in this analysis reflects the standard deviation (or range) of values across the relative corpus. Considering these Familiarity values in conjunction with the

Polysemy values (reported above), the results suggest that Japanese writers use a diverse range of high frequency, concrete words together with a large number of low frequency (presumably technical) terms. The result would mean that Japanese writers may be tending to bind together a large number of specialist terms with relatively simple English structures. Such a text may once more prove to be less than optimal for native readers who are assessing the manuscripts.

Parts of Speech (Gerunds Incidence). The results for incidence of *gerunds* again suggest that Japanese writers may be avoiding difficult grammatical structures. Gerunds reflect a grammatical aspect that non-native speakers may tend to avoid for fear of making errors. The incidence here suggests the use is significantly lower for the Japanese compared to either the Americans or the British.

Lexical Diversity. Lexical diversity is a useful indicator of cohesion (via redundancy) and difficulty (via word range). The results in this study suggest that the Japanese use a significantly narrower range of vocabulary than do the British. The fact that this narrower range does not transfer into a significant difference for argument overlap (cohesion) suggests that British writers may be using their lexicon far more effectively.

Parts of Speech (*to* incidence). The incidence of Parts of Speech for the infinitive-*to* is yet another indicator of Japanese reluctance or inability to use more complex grammatical structures. Once again, there is no significant difference between the British and American writers, but the Japanese use significantly fewer instance than either of their native-English speaking colleagues.

Parts of Speech (W-adverbs). The incidence of Parts of Speech for W-adverbs (e.g. the *why* of *We do not know why this happened*) once more indicates a Japanese avoidance of more difficult syntactical structures. W-adverbs are typical of more complex multi-clausal sentences and while the Japanese use does not appear significantly different from the usage of Americans, the Japanese do use significantly fewer instances than the British.

7. Accuracy of the model

To test the accuracy of our findings, we conducted a series of discriminant analyses. A discriminate analysis is a statistical procedure that culminates with a prediction of group membership (in this case, native language category). In this study, as is typical of discriminant analyses' studies, the accuracy of the results are reported in terms of *recall* and *precision*. Recall shows the number of correct predictions divided by the total number of items in the group. Precision, on the other hand, is the number of correct predictions divided by the sum of the number of correct and incorrect predictions. The distinction between precision and recall is important because an algorithm that predicts everything to be a member of a single group will account for all members of that particular group (scoring 100% in terms of recall) but will also falsely claim many members of other group(s), thereby scoring poorly in terms of precision. Reporting both values allows for a better understanding of the accuracy of the model.

Japanese English and British English. For the first discriminant analysis, the dependent variable (or *grouping variable*) was the native language categories of Japanese English/British English and the independent variables (or *predictor variables*) were the 15 selected variables discussed above. A total of 401 cases were analyzed. Univariate ANOVA revealed that the native language groups differed significantly on 11 of the 15 variables. When there are two groups in an analysis (as in this case, *Japanese-English* and *British-English*), a single discriminant function is calculated. This discriminant function is the essence of the model that differentiates the two groups, predicting group membership. The function works by determining whether groups differ with regard to the mean of each independent (or predictor) variable. A combination of these evaluations is used to predict group membership. In this analysis, the value of this function was significantly different for the two language groups ($\chi^2 = 79.395$, df = 15, $p < .001$), meaning that the combination of the predictor variables was sufficient to differentiate between the two native language categories to a degree above chance. Correlations between the predictor variables and the discriminant function suggested that higher values for *use of third-person* *singular*, *lexical diversity* were more indicative of British writers while a lower ratio for *sentence syntax similarity* was indicative of Japanese writers. Overall, the discriminant function successfully distinguished the two groups. Using the discriminant function algorithm generated from the training set to predict group membership of the *test set*, the model showed a significant distinction between groups ($\chi^2 = 12.086$, $p = .001$). The accuracy of the model for predicting Japanese-English texts was approximately 62% (recall = 63.63%; precision = 62.76%). The accuracy of model for predicting British-English texts was also approximately 62% (recall = 61.00%; precision = 62.89%). While the distinction may seem small, it is statistically above chance and points towards a number of differences between the writing styles of scientists in these language categories.

Japanese English and American English. For the second discriminant analysis, the dependent variable was the native language categories of Japanese English/American English. A total of 401 cases were analyzed. Univariate ANOVA revealed that the native language groups differed significantly on 12 of the 15 variables. The fact that the non-significant variables in this analysis differ from those in the analysis between Japanese-English and British-English supports previous ANOVA results from the training set and further suggests that Japanese-English writers may have to be aware of two significantly different forms of English. The value of the discriminant function for this pairing was significantly different for the two language groups ($\chi^2 = 64.258$, df = 15, $p < .001$). Correlations between the predictor variables and the discriminant function suggested that *use of third-person singular*, *word frequency*, and *incidence of the to-infinitive* were the three most predictive variables of language category. The results suggest greater grammatical variation by the American-writers and more frequent use of uncommon words; however, as the *lexical diversity* was not significantly different, we might speculate that Japanese writers compensated for grammar and low frequency words with a wider range of words. This indeed is supported by the data with American-English writers' text size ($M = 101.67$, $SD = 43.51$) and Japanese writers' text size ($M = 112.35$, $SD = 29.00$) showing a significant difference: $F(1, 399) = 8.36$; $p = .004$).

Overall, the discriminant function successfully distinguished the two groups. An analysis of the *test set* alone showed a significant distinction between groups (χ^2 = 16.998, p < .001). The accuracy of model for predicting American-English texts was approximately 61% (recall = 65.22%; precision = 56.07%). The accuracy of model for predicting Japanese-English texts was approximately 86% (recall = 80.20%; precision = 92.05%). The distinction is statistically above chance and points towards a number of differences between the writing styles of scientists in these language categories.

British English and American English. For the third discriminant analysis, the dependent variable was the native language categories of British English/American English. A total of 402 cases were analyzed. Univariate ANOVA revealed that the native language groups differed significantly on only 7 of the 15 variables: The 8 non-significant variables were *incidence of third-person singulars*, *incidence of gerunds*, *incidence of cardinal numbers*, *incidence of infinitive to*, *lexical diversity*, *incidence of intentional event*, *standard deviation of word familiarity, Sentence syntax similarity*. Although 8 variables were not significant, the finding that half the variables in the *test set* significantly distinguished between two dialects of native speakers provides more confidence in the results of the Bonferroni analysis (discussed above). The value of the single discriminant function was significantly different for the two language groups (χ^2 = 30.435, df = 15, p = .010). Correlations between the predictor variables and the discriminant function suggested that *word frequency*, (high for British) *polysemy* (greater for British), and *incidence of intentional events* (lower for American) were the three most predictive variables of language category.

Overall, the discriminant function successfully distinguished the two groups. An analysis of the *test set* alone showed a significant distinction between groups (χ^2 = 8.119, p = .004). The accuracy of model for predicting American-English texts was approximately 56% (recall = 52.08%; precision = 60.98%). The accuracy of model for predicting British-English texts was approximately 64% (recall = 68.00%; precision = 59.65%). In this analysis, the distinction between the groups is small but statistically above chance, pointing towards a number of differences between the writing styles of scientists in these language categories.

Native English Speakers (American and British combined) and Japanese English. For the final discriminant analysis, the dependent variable was the native language categories of native English speaker/Japanese-English speaker. A total of 602 cases were analyzed. Univariate ANOVA revealed that the natural language groups differed significantly on 14 of the 15 variables: The one non-significant variable was *mean of location and motion ratio scores.* Compared to the previous analyses, the high number of significant differences between these groups suggests that British and American writers of English have more in common with each other in their writing styles than Japanese writers have in common with either. Such an outcome would be predictable. The value of the discriminant function was significantly different for the two language groups: (χ^2 = 113.142, df = 15, p < .001). Correlations between the predictor variables and the discriminant function suggested that *verb phrase incidence score*, *word frequency*, and *sentence syntax similarity* were the three most predictive variables of language category.

Overall, the discriminant function successfully distinguished the two groups. An analysis of the *test set* alone showed a significant distinction between groups (χ^2 = 43.842, p < .001). The accuracy of model for predicting native-English texts was approximately 76% (recall = 68.89%; precision = 82.82%). The accuracy of model for predicting Japanese-English texts was approximately 63% (recall = 71.72%; precision = 53.79%). The distinction between texts written by Native English speakers and those written by Japanese writers in English is moderate but statistically well above chance, pointing towards a number of differences between the writing styles of scientists in these language categories.

8. The Gramulator: N-Gram analysis

Tools such as Coh-Metrix examine texts at the word, sentence, and discourse level; however, to analyze text as *sequences of probabilistically occurring words* at the sub-sentential level, we developed an n-Gram tool called *the Gramulator*.

An n-gram is a string of adjacent words, where the *n* represents the number of adjacent words. For instance, two adjacent words are a *bi-gram*; and three words are a *tri-gram*. Typically, the most common n-grams (e.g., *of the*) do not differ from corpus to corpus. With this in mind, the Gramulator was developed to analyze n-grams in terms of *statistically improbable features* (SIF). SIF are those n-grams that are common to one corpus (i.e. among the 50% most frequent n-grams) but uncommon to another (i.e. among the 50% least frequent n-grams). By identifying SIF, we are able to identify the most common and least common language sequences.

As expected, Japanese, American, and British scientists' most frequent bi-grams and tri-grams were similar (see Tables 3 and 4).

JS		AS		BS	
Bi-gram	Freq.	Bi-gram	Freq.	Bi-gram	Freq.
of the	283	of the	241	of the	290
in the	279	in the	225	in the	258
patients with	118	to the	79	and the	90
And the	105	associated with	73	to the	77
to the	75	for the	52	for the	58
on the	73	and the	50	that the	56
for the	50	from the	50	on the	55
That the	50	on the	48	of a	53
with the	45	with the	46	to be	53
in a	41	as a	42	With a	53

Table3: Most frequent (freq) Bi-Grams for Japanese Scientists (JS), American Scientists (AS), and British Scientists (BS)

JS		AS		BS	
Tri-Grams	Freq	Tri-Grams	Freq	Tri-Grams	Freq
in patients with	24	as well as	17	study was to	18
of this study	17	the number of	16	there was no	16
study was to	17	the presence of	13	of this study	15
family history of	16	of this study	12	this study was	15
the number of	16	In patients with	11	the aim of	14
the risk of	16	study was to	11	in patients with	13
as well as	14	a total of	10	the presence of	13
the presence of	13	associated with the	10	aim of this	12
this study was	13	in response to	10	the use of	11
to investigate the	13	this study was	10	a range of	10

Table 4: Most frequent (freq) tri-Grams for Japanese Scientists (JS), American Scientists (AS), and British Scientists (BS)

However, the analysis of the statistically improbable features (SIF) is more revealing of lexical choices made by Japanese scientists as a condition of language register (see Table 5 and 6). First, the bi-gram *in Japan* is common to the Japanese scientists' papers only. While it is not surprising that the Japanese should report work in Japan, it is worth noting that neither British scientists nor American scientists recorded a single instance of a corresponding bi-gram (e.g. *in England*, *in Britain*, *in the UK*, *in the USA*, *in America*). Thus, this lexical choice clearly identifies the researchers' location (and nationality) and may (unconsciously) indicate to reviewers that the research is limited, focused, or difficult to generalize from.

A second bi-gram of interest is *among the* as in "among the 103 patients enrolled…." *Among the* featured in over 9% of Japanese abstracts (less than 2% for British and just 1% for American.) The 'over-use' of *among the* could be an inter-language transfer issue. For instance, a native English speaker might ask "Which is the most expensive, Tokyo, New York, or London?" but the Japanese translation would be "*Among* [*no naka de*] Tokyo, New York, or London, which is the most expensive city?" In all instances of the use of *among* by Japanese scientists, it would be hard to argue that the *wrong* word was used; however, it is clear that *grammatically acceptable* does not entail *commonly used.*

The tri-gram *of the patients* appears in eight Japanese abstracts but only once in American and British abstracts. Further investigation showed that the word *patients* occurred in 42% of Japanese abstracts, whereas American use was 23% and British use was 21%. These data suggest that the Japanese use of *patients* and *of the patients* may be overly redundant. McNamara (2001) argues that when readers are skilled and knowledgeable (as presumably they are in the domain of this study) redundancy can be counter-productive. McNamara suggests that comprehension is enhanced for higher knowledge readers when they are induced by the text to generate inferences (i.e., when the text is not replete with explicit cohesive features such as obvious repetitions of known concepts). Thus, it is possible that Japanese scientists' over-use of common or obvious terms may negatively affect the reviewing process of their manuscripts.

	AS		BS
in Japan	27	history of	36
between the	25	in Japan	27
in Japan.	24	in Japan.	24
was performed	24	was performed	24
family history	19	the present	23
The risk	19	these results	20
to investigate	19	family history	19
The two	18	should be	19
The mean	17	than in	19
among the	15	(p <	18

Table 5: Japanese Scientists' Statistically Improbable Bi-Grams Relevant to American Scientists (AS) and British Scientists (BS)

	AS		BS
family history of	16	family history of	16
the risk of	16	the present study	12
to investigate the	13	the prevalence of	11
the present study	12	a history of	10
based on the	11	in this study,	10
the aim of	11	of the patients	10
the effect of	11	patients with a	10
the expression of	11	wbc count and	10
the prevalence of	11	between the two	9
of the patients	10	from patients with	9

Table 6: Japanese Scientists' Statistically Improbable Tri-Grams Relevant to American Scientists (AS) and British Scientists (BS)

9. Discussion

From the result of both analyses, we can see a wide variety of distinctions between abstracts written by Japanese scientists and those written by native writers of English. Given that the differences cover a wide variety of variables at each of the text analysis levels (*discourse*, *sentence*, and *word*), it is reasonable to conclude that a native-English speaker (reviewer, editor, or reader) may interpret the Japanese texts as lacking in key areas of the proto-typical style of the register. Such a conclusion supports the claims of Hinds (1990) that texts written in

English by Japanese writers are regarded differently from those written by native English speakers. While this study produced no evidence to claim that these differences would produce a negative effect on readers of the text (or any effect at all for that matter) it is reasonable to assume that the differences in writing style of the Japanese scientists are unlikely to enhance their chances of gaining optimal reviews.

This study also produced findings pointing towards significant differences between British and American texts. These findings support such research as Crossley and Louwerse (2007) and Hall et al. (2007). The findings are important as they demonstrate that the variety of English taught in schools, and subsequently presented in composition, is not a trivial question. That is, reviewers from both British and American journals may have certain expectations as to the text: expectations that go beyond mere grammatical and lexical correctness. These expectations could include diversity within the text's structure or lexicon, a diversity that may serve to enhance the reader's interest. Japanese writers need to be familiar with these aspects in order to produce manuscripts that increase their chances of optimally conveying their research. Of course, this conclusion may mean that if non-native English speakers (such as the Japanese) aim to publish in both American and British arenas, then they may be faced with the prospect of learning *Englishes*, and adapting their English for the journal in question. The algorithms described in this study may go a long way to assisting these writers in assessing the degree to which their texts have met those standards.

One further benefit from this study addresses materials development. As Orr (2001) acknowledges, teaching ESP is often highly time consuming because of the difficulty of collecting suitable amounts of material. However, as demonstrated in this study, large numbers of natural examples of target texts are often quite freely available. Thus, one immediate pedagogical implication for this study is that educators (as well as researches themselves) may collect corpora and use the techniques highlighted in this study to determine better the degree to which their relevant text corresponds to desired target text type. Indeed, workshops have already been prepared and delivered to non-native instructors of English on collecting, preparing, and examining relatively large corpora for ESP classes (Hall, 2006). Additionally, professional organizations have begun to realize the importance of this topic and have included papers and demonstrations in their programs on these methods (Hall, McCarthy, Lewis & McNamara, 2007). Even publishers, such as Cambridge University Press, have begun to encourage the use of corpus examination in the preparation of commercial materials (Moor, 2005).

Both the data we gathered and the n-gram analysis technique in general is particularly useful for instructors and developers because it is both easily conducted and easily understood. As we saw, one SIF *in Japan* could easily limit the apparent universality of a Japanese scientist's research. It would be very easy to train ESP specialists on collecting the data, applying a simple n-gram, and then including the results in highly focused instruction for specific discourse communities.

Two of the major limitations of our study are the focus on abstracts and the limitation of non-native English speakers' texts to Japanese. Abstracts provide an easily assessable and comparable example of text to compare. Abstracts have also been shown to be highly indicative of trends across science papers (McCarthy, Briner, et al., 2007). However, to better understand the differences between writing styles in English, future research must consider other sections of texts such as *introductions*, *methods sections*, and *discussions*. Future research must also consider other English language varieties' and other language groups' (such as Chinese or German) production of scientific texts.

This study sheds light on a number of features of English and how these features differ according to language variety. The findings offer guidance to teachers, students, and researchers as to how and where text might be modified in order to facilitate the production of more native-English like representations. While such a study cannot hope to completely level the playing field on which non-native speakers of English are forced to compete, it does at least offer some hope that computational analyses (such as those produced by Coh-Metrix and the Gramulator) will better facilitate those whose careers depend on written production in a foreign language.

Summary

In this study, we used two tools (Coh-Metrix and the Gramulator) to analyze three corpora of science journal abstracts written by American, British, or Japanese scientists. The purpose of our study was to explore differences between the writing of Japanese scientists and the writing of their native English speaking colleagues. To conduct the analysis, we first used Coh-Metrix to analyze text on cohesion, readability, and difficulty. We then conducted n-gram analyses using the Gramulator for information at the phrasal and word level.

The results of ANOVA for the Coh-Metrix data suggested significant differences between the Japanese texts and native English speakers' texts. However, the analysis also showed significant differences between the British and the American texts, supplying further evidence to studies such as Hall et al. (2007) that the major English language varieties feature substantial and consistent textual differences.

Finally, we used the Gramulator to explore both the most frequent and statistically improbable features (SIF) of bi- and tri-grams for the three forms of English (Japanese, British, American). As expected the most common n-grams were similar across registers, but the SIF analysis showed that there are n-grams that may be the result of inter-language transfer (such as *among the*). There are also other n-grams (such as *in Japan*) which by seemingly limiting the scope of the paper or by providing too much identifying information about the authors could be a source of low ratings by reviewers.

Acknowledgements

This research was supported in part by the Institute for Education Sciences (IES 305G020018-02) and in part by Counter-intelligence Field Activity (CIFA H9c104-07-C-0014). Opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the IES or CIFA. Some findings from this analysis were previously reported in Foreign Language for Special Purposes. The authors would also like to acknowledge the contributions made to this paper by Brian Lehenbauer, James West, Art Graesser, Zhiqiang Cai, Gabor Tigyi, and Ryoko Tsukahara.

Recebido em: 07/2008; Aceito em: 10/2008.

References

Avent, J.R. & Austermann, S. 2003 Reciprocal scaffolding: a context for communication treatment in aphasia. *Aphasiology*, **17**: 397-404.

Belcher, D. & Connor, U. (eds.) 2001 *Reflections on multiliterate lives*. Multilingual Matters.

Best, R.M.; Rowe, M.; Ozuru, Y. & McNamara, D.S. 2005 Deep-level comprehension of science texts: the role of the reader and the text. *Topics in Language Disorders*, **25**: 65-83.

Bhatia, V. 1997 Applied genre analysis and ESP. IN: T. Miller (ed.) *Functional approaches to written text: Classroom applications*. USIA.

Biber, D. 1987 A textual comparison of British and American writing. *American Speech*, **62**: 99-119.

_______ 1988. *Linguistic features: algorithms and functions in variation across speech and writing.* Cambridge University Press.

Brace, N.; Kemp, R. & Snelgar, R. 2006 *SPSS for psychologists: a guide to data analysis using SPSS for Windows (Versions 9, 10, & 11)*. Erlbaum.

Brill, E. 1995 Transformation-based error-driven learning and natural language processing: a case study in part-of-speech tagging. *Computational Linguistics*, **21**: 543-565.

Canagarajah, A.S. 1996 Nondiscursive requirements in academic publishing, material resources of periphery scholars, and the politics of knowledge production. *Written Communication*, **13**: 435-472.

________ 2002 *A geopolitics of academic writing*. University of Pittsburgh Press.

Charniak, E. 1997 Statistical parsing with a context-free grammar and word statistics. IN: *Proceedings of the Fourteenth National Conference on Artificial Intelligence*. AAAI/MIT Press.

Coltheart, M. 1981 The MRC psycholinguistic database. *Quarterly Journal of Experimental Psychology*, **33**: 497-505.

Crossley, S.A. & Louwerse, M.M. 2007 Multi-dimensional register classification using bi-grams. *International Journal of Corpus Linguistics*.

Curry, M. & Lillis, T. 2004 Professional academic writing by multilingual scholars. *Written Communication*, 23: 3-35.

DENNETT, J.T. 1988 Not to say is better than to say: how rhetorical structure reflects cultural context in Japanese-English technical writing. *IEEE Transactions on Professional Communication*, **31**: 116-119.

DURAN, N.D.; MCCARTHY, P.M.; GRAESSER, A.C. & MCNAMARA, D.S. 2007 Using temporal cohesion to predict temporal coherence in narrative and expository texts. *Behavior Research Methods*, **39**: 212-223.

FELLBAUM, C. (ed.) 1998 *WordNet: an electronic lexical database*. MIT Press.

FLOWERDEW, J. 1999 Writing for scholarly publication in English: the case of Hong Kong. *Journal of Second Language Writing*, **8**: 123-145.

________ 2001 Attitudes of journal editors to nonnative speaker contributions. *TESOL Quarterly*, **35**: 121-150.

GIBBS, W.W. 1995 Lost science in the third world. *Scientific American*, **273**: 76-83.

GIBBS, H.; KENDALL, F. & PAGEL, W. 2002 Self-identified publishing needs of nonnative English-Speaking faculty and fellows at an Academic Medical Institution. *Science Editor*, **25**:111-114.

GRAESSER, A.; MCNAMARA, D.; LOUWERSE, M. & CAI, Z. 2004 *Coh-Metrix*: analysis of text on cohesion and language. *Behavioral Research Methods, Instruments, and Computers*, **36**: 193-202.

GRELA, BERNARD G. 2002 Lexical verb diversity in children with Down Syndrome. *Clinical Linguistics & Phonetics*, **14**: 251-263.

HALL, C. 2006 *Workshops in legal English*. FEELTA Conference, Vladivostok, Russia.

_______; MCCARTHY, P.M.; LEWIS, G.A.; LEE, D.S. & MCNAMARA, D.S. 2007 A Coh-Metrix assessment of American and English/Welsh Legal English. *Coyote Papers: Psycholinguistic and Computational Perspectives. University of Arizona Working Papers in Linguistics*, **15**: 40-54.

_______; MCCARTHY, P.M.; LEWIS, G.A. & MCNAMARA, D.S. 2007 Simple steps to corpus analysis of genres. Paper presented at *The 41st Annual TESOL Convention and Exhibit*, Seattle, Washington.

HALLIDAY, M.A.K. & HASAN, R. 1976 *Cohesion in English*. Longman.

HEWINGS, M. 2006 English language standards in academic articles: attitudes of peer reviewers. *Revista Canaria de Estudios Ingleses*, **53**: 47.

HINDS, J. 1990 Inductive, deductive, quasi-inductive: expository writing in Japanese, Korean, Chinese and Thai. IN: U. CONNOR & A. JOHNS (eds.). *Coherence in writing: research and pedagogical perspective*. TESOL.

JERNUDD, B.H. & BALDAUF R.B. 1987 Planning science communication for human resource development. IN: B.K. DAS (ed.). *Communicative language teaching*. RELC.

JOHANSSON, S. 1985 Some observations on word frequencies in three corpora of present-day English texts. *I.T.L. Review of Applied Linguistics*, **67-68**: 117-26.

JURAFSKY, D.S. & MARTIN, J.H. 2000 *Speech and language processing*. Prentice Hall.

KAUR, S. & SOOK, P. 2005 Towards a process-genre based approach in teaching of writing for business English. *ESPWorld*. Retrieved September 25, 2006. Available on-line at: http://www.esp-world.info/Articles_11/Sarjit-poon2.htm.

KLARE, G.R. 1974–1975 Assessing readability. *Reading Research Quarterly*, **10**: 62-102.

KINTSCH W. 1988 The use of knowledge in discourse processing: a construction-integration model. *Psychological Review*, **95**: 163-182.

_______ & VIPOND, D. 1979 Reading comprehension and readability in educational practice and psychological theory. IN: L.G. NILSSON (ed.). *Perspectives on memory research*. Lawrence Erlbaum Associates.

LANDAUER, T.K. & DUMAIS, S.T. 1997 A solution to Plato's problem: the latent semantic analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review*, **104**: 211-240.

LANDAUER, T.; MCNAMARA, D.S.; DENNIS, S. & KINTSCH, W. (eds.). 2007 *Handbook of latent semantic analysis*. Erlbaum.

LILLIS, T. & CURRY, M. 2006 Re-Framing notions of 'competence' in multilingual scholarly writing. *Revista Canaria de Estudios Ingleses,* **53**: 63-78.

LOUWERSE, M.M. 2001 An analytic and cognitive parameterization of coherence relations. *Cognitive Linguistics*, **12**: 291-315.

Louwerse, M.M.; McCarthy, P.M.; McNamara, D.S. & Graesser, A.C. 2004 Variation in language and cohesion across written and spoken registers. IN: K. Forbus, D. Gentner & T. Regier (eds.) *Proceedings of the 26th Annual Meeting of the Cognitive Science Society*. Erlbaum.

Maas,s H.D. 1972 Zusammenhang zwischen Wortschatzumfang und Länge eines Textes. *Zeitschrift für Literaturwissenschaft und Linguistik*, **8**: 73-79.

Malvern, D.D.; Richards, B.J.; Chipere, N. & Duran, P. 2004 *Lexical diversity and language development: quantification and assessment*. Palgrave Macmillan.

Marcus, M.; Santorini, B. & Marcinkiewicz, M. 1993 Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics*, **19**: 313-330.

McNamara, D.S. 2001 Reading both high-coherence and low-coherence texts: effects of text sequence and prior knowledge. *Canadian Journal of Experimental Psychology*, **55**: 51-62.

Miller, G. 1990 WordNet: An on-line lexical database. *International Journal of Lexicography***, 3**235-312.

McCarthy, P.M. 2005 An assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual, lexical diversity (MTLD). *Dissertation Abstracts International*, **66**: 12, (UMI No. 3199485).

_______; Lewis, G.A.; Dufty, D.F. & McNamara, D.S. 2006 Analyzing writing styles with Coh-Metrix. IN: *Proceedings of the Florida Artificial Intelligence Research Society International Conference (FLAIRS)*.

_______; 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*. Springer-Verlag.

Moore, J. 2005 *Common mistakes at proficiency*. Cambridge University Press.

Orr, T. 2001 English language education for Specific Professional Needs. *IEEE Transactions on Professional Communication*, **44**: 207-211.

RANSDELL, S. & WENGELIN, Å. 2003 Socioeconomic and sociolinguistic predictors of children's L2 and L1 writing quality. *Arob@se*, **1-2**: 22-29. Available on-line at: http://www.arobase.to/somm.html.

SEKINE, S. & GRISHMAN, R. 1995 A corpus-based probabilistic grammar with only two nonterminals. IN: *Fourth International Workshop on Parsing Technologies*. Prague: Karlovy Vary. pp. 260-270.

TABACHNICK, B.G. & FIDELL, L.S. 1989 Using multivariate statistics. 2nd edition. HarperCollins.

_______ 2001 *Computer-assisted research design and analysis*. Allyn and Bacon.

TARDY, C. 2004 The role of English in scientific communication: lingua franca or Tyrannosaurus rex? *Journal of English for Academic Purposes*, **3**: 247-269.

YULE, G.U. 1944 *The statistical study of literary vocabulary*. Cambridge University Press.

ZWAAN, R.A.; LANGSTON, M.C. & GRAESSER, A.C. 1995 The construction of situation models in narrative comprehension: an event-indexing model. *Psychological Science*, **6**: 292-297.

Appendix: Examples of usage for American, British, and Japanese English

3rd person singular common to British English:
In a minority of patients, lung transplantation provides the only hope of long-term survival. The median survival of patients with IPF is approximately 3 years, which in turn emphasizes the need for further investigation into its pathogenesis and potential disease-modifying pharmacological therapies.

3rd person singular *not* common to Japanese English:
A 38-year-old man presented with a progressive swelling of the entrance of left external auditory meatus. The patient underwent a surgical removal of the tumor.

High cardinal number usage common to Japanese English:
The mean early- (40-60 min after injection) and delayed (100-120 min)-phase ablated lesion-to-muscle ratios were, respectively, 2.9 +/- 1.0 and 3.3 +/- 0.8

(1 d), 4.1 +/- 0.6 and 5.2 +/- 0.9 (1 wk), 4.1 +/- 1.0 and 5.3 +/- 1.5 (2 wk), 3.1 +/- 0.5 and 3.6 +/- 1.1 (4 wk), and 1.8 +/- 0.1 and 2.3 +/- 0.1 (8 wk).

High overlap common to Japanese (note use of *blood* and *urine*/*urinary*):
We examined how the influence of smoking on *blood* and *urinary* cadmium (Cd) concentrations was modified by the level of environmental Cd. We measured *blood* and *urinary* Cd concentrations of 1134 men over 50 yr of age in three areas in Japan that were exposed to different levels of environmental Cd. Analysis of variance was used to compare Cd concentrations in *blood* and *urine* of smokers with those of nonsmokers living in the three areas.

Low overlap common to American English:
To examine the contributions of Archaea to digestive health, we colonized germ-free mice with Bacteroides thetaiotaomicron, an adaptive bacterial forager of the polysaccharides that we consume, with or without M. smithii or the sulfate-reducing bacterium Desulfovibrio piger. Whole-genome transcriptional profiling of B. thetaiotaomicron, combined with mass spectrometry, revealed that, unlike D. piger, M. smithii directs B. thetaiotaomicron to focus on fermentation of dietary fructans to acetate, whereas B. thetaiotaomicron-derived formate is used by M. smithii for methanogenesis.

Philip McCarthy is assistant professor at The University of Memphis. He is a computational linguist, primarily interested in devising, writing, and testing algorithms for text disambiguation. He has numerous publications in several fields including linguistics, artificial intelligence, and cognitive psychology. pmccarthy@mail.psyc.memphis.edu

Charles Hall is associate professor of applied linguistics at the University of Memphis. His major areas are English teaching methodology, ESP for Law, and curriculum development for o nline courses. He has led sponsored workshops in almost 30 countries. He was chair of the TESOL ESP Interest Section in 2005. cehall@memphis.edu

Nicholas Duran is a researcher at the University of Memphis in the Department of Psychology and the Institute for Intelligent Systems. His research interests include action dynamics, language use and representation, corpora analysis, and deceptive behavior. He holds a

MS in Experimental Psychology (Cognitive Emphasis) and is pursuing a Ph.D. in Experimental Psychology. nickd3ps@gmail.com

Maki Doiuchi, M.A., is a doctoral candidate in the School of Audiology and Speech-Language Pathology at The University of Memphis. She is primarily interested in speech acoustics and vocal development in infancy. Her native language is Japanese but she is fluent in English. bm2829@gmail.com

Yuko Fujiwara is an Assistant Professor at University of Tennessee, Health Science Center in Memphis. She received Ph.D. in cell biology and her research focuses on the function of bioactive lipids. Her native language is Japanese but she writes and speaks English fluently. yuko@physio1.utmem.edu

Benjamin Duncan, Ph.D. teaches at the College Writing Program at the University of Rochester. His research interests include technical and science writing and English second language issues. bduncan3@mail.rochester.edu

Danielle McNamara is a Professor at the University of Memphis and Director of the Institute for Intelligent Systems. Her work involves the theoretical study of cognitive processes as well as the application of cognitive principles to educational practice. Her current research ranges a variety of topics including text comprehension, writing strategies, building tutoring technologies, and developing natural language algorithms. d.mcnamara@mail.psyc.memphis.edu