A Coh-Metrix Analysis of Variation among Biomedical Abstracts

Benjamin Duncan*, Charles Hall**

* Texas A&M University, English Language Institute, **Department of English
*College Station, TX 77840-2130
**University of Memphis, Memphis, TN 38152
brduncan@neo.tamu.edu, cehall@memphis.edu

Abstract

Using the already validated Coh-Metrix tool, this study examines whether there are significant linguistic and discourse differences between biomedical abstracts for American and Korean English. Also, the current study accounts for variation among journals' countries of origin, distinguishing between biomedical journals published in the United States from biomedical journals published in South Korea. The significance of these studies regards the growing number of second language (L2) biomedical researchers attempting to publish their research in national and international English-language journals, but who find themselves locked out of the discussion because of differences in linguistic and discourse conventions. The present study aims to provide a more thorough and quantitative understanding of the prototypical linguistic components in biomedical rhetoric, and to suggest how word-, sentence-, and discourse-level structures can be researched, taught, and developed into materials. This improved understanding is expected to provide a powerful apparatus for the promotion of L2 English writers in the biomedical field.

Introduction

With between 74% and 90% of all scholarly work now being published in English (Lillis & Curry 2006), second language speakers (L2s) face growing difficulties in publication and professional promotion. This concern is especially true in the biomedical community, where a growing number of NNSs publish and wish to publish in international, English-language journals.

McCarthy et al. (2007) have used a natural language processing tool known as Coh-Metrix to verify that significant linguistic and discourse differences exist among the scientific abstracts of American, British, and Japanese authors. As Graesser (1994) points out, practitioners within specialized communities may implicitly or explicitly anticipate writing to follow certain linguistic and discourse guidelines. As a result, writers within the specialized community of biomedicine, whether L1 or L2s, who misuse and misunderstand the conventions of a genre are placed at a distinct disadvantage for international publication and academic promotion.

To date, little research has analyzed the language and discourse of scientific and technical writing for differences between first and second language authors. Perhaps one reason for this lack has been the difficulty collecting and processing a large corpus of samples within a particular genre, a difficulty which has recently been eased through computer software and natural language processing tools, such as Coh-Metrix.

The purpose of this study is to extend McCarthy and colleagues' earlier study using a larger and more specific corpus (N = 1500) that considers only a subfield of the scientific abstracts: biomedical abstracts. In addition, the corpus for this study differentiates between publications in top-tier biomedical journals from South Korea from those published in the United States over the past five years to gain a clearer understanding of how one specific group of NNSs may adapt their language and discourse to meet the current conventions of L1s within the biomedical community.

Coh-Metrix

Natural language processing has been at the forefront of determining how grammatical and lexical features are used to differentiate one community's shared discourse and shared texts from another (i.e., Biber 1988). Biber's initial processing of linguistic variation was able to uncover nine linguistic features that could accurately predict whether a text was written or spoken. By combining traditional and new linguistic metrics across word-, sentence-, and discourse-ranges, researchers have been able to uncover the grammatical and lexical features that separate one genre, or one specialized discourse community, from another. For instance, McCarthy, Graesser, and McNamara (2006) used Coh-Metrix (Graesser et al. 2004) to distinguish between science, social studies, and narrative texts. McCarthy et al. (2007) used the Coh-Metrix tool to differentiate among American, British, and Japanese scientific abstracts.

Coh-Metrix integrates Biber's linguistic features (Biber 1988), as well as indices of cohesion (Jurafsky & Martin 2000), readability (Flesch-Kincaid Grade Level; Klare 1974, 1975), syntactic parsers (Sekine & Grishman 1995), part-of-speech taggers (Brill 1992), Latent Semantic Analysis (LSA, Landauer, McNamara, Dennis, & Kintsch 2007), word and conceptual indices derived from the WordNet lexical database (Miller 1990), the

MRC database (Coltheart 1981), to produce over 600 metrics of language and text.

Corpus

Abstracts for the present study were collected from Medline, an online database of international biomedical publications. The corpus selection accounted for only the most recent publications (within the past five years) as well as the country of journals' origins, namely biomedical journals published in the U.S. and South Korea. Text samples were culled from several different first-tier American and Korean biomedical journals, covering the range of textual variation in the domain. The corpus for compared three groups: American authors published in top-tier American biomedical journals in their native English language (AE); Korean authors published in top-tier American biomedical journals in English, their second language (AK), and Korean authors whose work was published in top-tier Korean biomedical journals in English, their second language (KK). The corpus includes AE biomedical abstracts (n = 500), AK biomedical abstracts (n = 500), and KK biomedical abstracts (n = 500). Several selection constraints were applied to ensure uniformity and representation across individual journals and authors. The corpus was cleaned prior to Coh-Metrix analysis. For every abstract, headers, including authors' names, titles, journal names, dates of publication, were removed. Likewise, endings, including publishers' addresses, author affiliation, government grant numbers, were removed. Paragraph endings and extra spaces where also removed. Subheadings (background, methods, results) were included. The focus of the analysis, therefore, was only on the language of the actual abstract. All of the abstracts included in the corpus were written by medical professors or doctors with previous publishing experience and affiliations with medical departments at hospitals or universities.

Predictions

The results reported by McCarthy et al. (2007), provide the basis for our research predictions (see Table 1).

Co-referential overlap. Coh-Metrix measures the incidence of various types of referential overlaps, such as *noun* and *verb overlap*. Overlap measures how often a common noun, verb, or stem appear in two adjacent sentences. Overlap increases cohesion and comprehension by directly linking topics, thus limiting a reader's reliance on inferencing and facilitating recall (McNamara 2001).

We predict that Korean-authored biomedical abstracts published in American journals and Korean journals (AK and KK) will contain significantly more co-referential overlap than American-authored biomedical abstracts published in American journals (AE), indicating a greater rate of redundancy.

Locational cues. Locational cues indicate spatiality within a text. Words such as *here, there, over, under* are considered to be locational cues.

We predict that AE abstracts will contain significantly fewer locational cues than AK and KK abstracts. The use of locational items may be an overly-elementary means of situating the reader rarely used in advanced, technical writing.

Intentional cues. Intentional particles and verbs indicate relations between agents and actions. Particles such as *because, consequence of*, and *as a result*, and verbs such as *kill, eat*, and *annoy* entail a related agent and action.

We predict that AE abstracts will contain significantly more intentional cues than AK and KK abstracts. Intentional cues represent a more advanced cohesive technique that requires a greater range of vocabulary and a greater knowledge of the relations between certain words. L2 speakers may struggle with the application of intentional cues.

Temporal cues. Temporal cues provide the reader with information concerning the passage of time within a text. Temporal cohesion can be calculated by analyzing the tense and aspect of particular verbs.

We predict that AK and KK abstracts will contain significantly more temporal cues than AE abstracts. American authors are presumably more familiar with the expository structures involved in composing an English abstract and should be able to alter tense and aspect as needed. Korean authors may feel more comfortable adhering to a single tense and aspect, suggesting an unwillingness to engage in risk taking.

Concreteness. Concreteness signifies how easily words can be pictured in the mind's eye. For example, readers can more easily picture a highly concrete word such as *train* than a less concrete word such as i*ntelligence*.

We predict that all three groups (AK, AE, and KK) will display a significantly high preference for concreteness of content words. A high preference for concreteness among all three groups is a key indication that Koreans' difficultly gaining entrance into the central discourse community is not a result of unfamiliarity with the terms and techniques of the field, but rather is located in other linguistic and rhetorical differences.

Syntactic complexity and diversity. Syntax refers to the complexity and diversity of sentence construction throughout the entire text.

We predict that AK and KK abstracts will exhibit significantly more similarity and simplicity in their sentence syntax whereas AE abstracts will be more diverse and complex in their syntax. Koreans may be less

adept in varying grammatical constructions, and as a result, use a narrower range of sentence types.

Word frequency. Some words, especially technical terms, are low frequency words. We predict that AE abstracts will contain significantly more low frequency words than AK and KK abstracts. The use of low frequency words suggests that Americans use a wider and more advanced range of vocabulary than do Koreans.

Incidence of third person-singular morphemes. We predict that AE and AK abstracts will contain significantly more third person-singular morpheme (e.g., *–s* in "she draw*s*") than KK abstracts. Central members of the biomedical community situate their research among the most current, most relevant topics of discussion. Therefore, present tense, emphasizing the present importance of the research, and third person, emphasizing the universality of the claims, is preferred. Note, for example, the difference between *This finding shows how to save lives* versus *My finding showed how to save lives*.

Incidence of cardinal numbers. We predict that KK and AK abstracts will contain more cardinal numbers than AE abstracts. McCarthy et al. (2007) argued that L2 speakers were more comfortable stating the results of their research without context rather than explaining the significance of the results and situating their data (in the form of cardinal numbers) within previous studies.

Word familiarity. Word familiarity is calculated via WordNet (Miller 1990) and is a strong indication of lexical decision making and ease of readability. For example, a word such as *nuptial* is familiar to fewer readers than is a word such as *marriage* or *wedding.*

We predict that KK and AK abstracts will have more variation and produce greater standard deviations in word familiarity than AE abstracts. A narrower range in word familiarity indicates L1's more extensive knowledge of vocabulary and the relations between certain words.

Incidence of gerunds. A gerund is a verb acting as a noun, such as the noun *swimming* in *Swimming is good exercise*. Gerunds may be an "advanced" grammatical structure that L2 speakers of English may be wary to use in their research. For example, Kleinmann (1977) relates that second language learners tend to avoid grammatical structures that differ markedly from their mother tongues. Avoidance of certain grammatical structures is often difficult to notice because there is usually an alternative structure to express the target language. We predict that such avoidance will cause KK and AK abstracts to contain fewer gerunds than AE abstracts.

Lexical diversity. Lexical diversity (LD) assesses the range of vocabulary employed in a text. We predict thatAK and KK abstracts will contain a significantly narrower range of vocabulary than AE abstracts because L1 speakers presumably have a greater number of English words available.

Incidence of infinitives. An infinitive is the root form of the verb, starting with *to*. Infinitives are prevalent in Western languages, but are non-existent n Korean. As such, Koreans may likely avoid using *–to* infinitives to the extent that L1 speakers use them (see Kleinmann 1977). We predict that AE abstracts will demonstrate a higher incidence of the infinitive *–to* + verb structure than AK or KK abstracts.

Incidence of Wh-adverbs. *Wh*-adverbs involve the use of *who, what, when, where* and *why* as modifiers of verbs, adjectives or other adverbs. For example, the *why* of *The results indicate why there may be a predilection toward one group. Wh*-adverbs may be another "advanced" grammatical structure that L1 speakers of English may be wary to use in their research, but can add to the syntactic diversity of the writing. We predict that the incidence of *parts of speech* for Wh-adverbs will be more prevalent in AE abstracts than in AK and KK abstracts.

Table 1. Summary of Findings in Relation to McCarthy and Colleagues (2007)

	McCarthy et al. (2007)	Prediction
1. Co-reference	Japan > US	Korea > US
2. Locational cues	Japan > US	Korea > US
3. Intentional cues	US > Japan	US > Korea
4. Temporal cues	Japan > US	Korea > US
5. Concreteness	US & Jap. high;	US& Kor. High
6. Syn. diversity	US > Japan	US > Korea
7. Low freq.words	US > Japan	US > Korea
8. 3rd p.-sin.- mor	US > Japan	US > Korea
9. Cardinal #	Japan > US	Korea > US
10. Word Familiar	Japan > US	Korea > US
11. Gerunds	US > Japan	US > Korea
12. Lexical Divers	US > Japan	US > Korea
13. *to* infinitives	US > Japan	US > Korea
14. *Wh*-adverbs	US > Japan	US > Korea

Predicted hierarchy. Korean speakers who have published in American biomedical journals are expected to more successfully demonstrate coherence to the conventional use of linguistic and discourse elements. In contrast, Korean researchers who have not published in American journals are expected to deviate from the expected linguistic and discourse conventions of the biomedical abstract genre. Thus, we predict that the linguistic and rhetorical elements of AE abstracts will appear more similar to those of AK abstracts (published

in American journals) than to those of KK abstracts (published in Korean journals).

Results

We processed the corpus using Coh-Metrix. Values for the predicted linguistic and rhetorical elements among the 650+ variables were isolated. Statistical descriptive indices (i.e., *M*, *SD, F, p*, and n_p^2 values) were calculated for each linguistic variable using SPSS software (2001). *F*, or effect size, is a measure of the strength of the relationship between variables. The n_p^2, or partial Eta squared values, calculates the proportion of the effect plus error variance that is attributable to the effect. The linguistic and rhetorical components among the three groups were compared using a planned contrast Bonferroni analysis for significant differences.

The results of McCarthy et al.'s (2007) Coh-Metrix analysis suggested 14 factors that revealed significant differences among American, British, and Japanese scientific abstracts. The results are presented according to the highest univariate *F*-values in Tables 2 and 3 below.

Table 2. F and n_p^2 values for the 3 groups: American abstracts published in American journals (AE), Korean abstracts published in American journals (AK), and Korean abstracts published in Korean journals (KK)

Dependent Variable	*F*	np^2
# Words/ Abstract	100.913***	0.119
Cont. Words / Sent.	65.755***	0.081
Int. Event/1000 Wo	58.842***	0.073
High-lConst./Word	52.614***	0.066
Hyper. Value Noun	35.757***	0.046
Card.#/1000 Words	34.042***	0.043
Loc. Prep./1000wor	26.241***	0.034
Stem Ov. Adj. Sen.	20.016***	0.026
Tense Aspect Rep.	19.283***	0.025
Arg. Ov.Adj. Sent.	17.818***	0.023
Tense Repetition	17.771***	0.023
Noun Ov.Adj. Sent.	17.385***	0.023
Con. Content Word	17.199***	0.022
–*to* Inf./1000 Word	11.391***	0.015
Word b4 MainVerb	10.742***	0.014
3rdP.-S./1000 words	10.066***	0.013
% Stem Ov.Adj. S	8.295***	0.011
Loc. Prep.& Nouns	7.930***	0.010
% Arg.Ov.Adj. S.	6.438**	0.009
% Con. Word Ov. .	5.624**	0.007
SD Word Fam.	5.266**	0.007
LD MTLD	4.873**	0.006
% Noun Ov.Ad.St.	4.860**	0.006
Wh-adv. /1000 Wo	2.073	0.003

Note. *** $p < .001$. ** $p < .01$. * $p < .05$.

Table 3. Mean (*M*) and standard deviation (*SD*) for the 3 groups: American abstracts published in American journals (AE), Korean abstracts published in American journals (AK), and Korean abstracts published in Korean journals (KK)

	M and (*SD*)		
Dependent Variable	AK	KK	AE
# Words/ Abstract	210.328	199.970	258.704
	(68.257)	(75.390)	(64.799)
Cont. Words / Sent.	1.241	1.366	1.451
	(0.236)	(0.236)	(0.235)
Int. Event/1000 Wo	13.005	13.709	7.951
	(9.525)	(9.965)	(5.923)
High-lConst./Word	0.626	0.628	0.650
	(0.038)	(0.038)	(0.039)
Hyper. Value Noun	4.080	4.421	4.404
	(0.654)	(0.612)	(0.503)
Card.#/1000 Words	48.375	48.666	67.528
	(36.904)	(45.388)	(40.477)
Loc. Prep./1000wor	47.281	38.227	42.286
	(18.824)	(17.213)	(16.070)
Stem Ov. Adj. Sen.	0.794	0.706	0.790
	(0.200)	(0.246)	(0.201)
Tense Aspect Rep.	0.737	0.713	0.806
	(0.277)	(0.247)	(0.185)
Arg. Ov.Adj. Sent.	0.738	0.662	0.742
	(0.218)	(0.253)	(0.213)
Tense Repetition	0.768	0.762	0.836
	(0.261)	(0.218)	(0.159)
Noun Ov.Adj. Sent.	0.703	0.623	0.698
	(0.224)	(0.260)	(0.226)
Con. Content Word	389.482	385.064	379.389
	(29.161)	(28.133)	(23.926)
–*to* Inf./1000 Word	11.261	9.681	13.238
	(9.085)	(8.877)	(10.756)
Word b4 MainVerb	6.578	5.936	6.812
	(2.963)	(2.625)	(3.384)
3rdP.-S./1000 words	9.357	10.591	7.730
	(9.630)	(10.440)	(9.261)
% Stem Ov.Adj. S	0.201	0.178	0.195
	(0.076)	(0.086)	(0.080)
Loc. Prep.& Nouns	0.460	0.433	0.435
	(0.114)	(0.133)	(0.103)
% Arg.Ov.Adj. S.	0.206	0.185	0.204
	(0.092)	(0.096)	(0.090)
% Con. Word Ov. .	0.164	0.148	0.160
	(0.069)	(0.075)	(0.073)
SD Word Fam.	47.132	47.291	44.987
	(11.842)	(13.943)	(10.370)
LD MTLD	71.707	76.533	74.302
	(22.316)	(27.344)	(20.530)
% Noun Ov.Ad.St.	0.188	0.171	0.185
	(0.088)	(0.094)	(0.088)
Wh-Adv./1000W	0.935	0.669	1.202
	(2.668)	(2.038)	(3.276)

McCarthy and colleagues' results described were used as a basis for comparison in the results described in Table 4 below.

Table 4. Summary of Findings in Relation to McCarthy and Colleagues (2007)

	Results	Significance
1. Co-reference	AE > AK	*
	AE > KK	***
2. Locational cues	AK > AE	***
	AE > KK	**
3. Intentional cues	Kor > US	***
4. Temporal cues	US < Kor	***
5. Concreteness	US & Kor. high	***
6. Syn. diversity	US > Korea	***
7. Low freq.words	US > Korea	***
8. 3rd p.-sin.- mor	Korea < US	**
9. Cardinal #	US > Korea	***
10. Word Familiar	Korea > US	**
11. Gerunds	No sig. dif.	*
12. Lexical Divers	No sig. dif.	*
13. *to* infinitives	US > Korea	**
14. *Wh*-adverbs	US > Korea	**

Note. *** $p < .001$. ** $p < .01$. * $p < .05$.

Discussion

The results suggest significant differences exist between KK abstracts and AE and AK abstracts. Fewer significant differences were discovered between AK and AE publications. Therefore, the results confirm the predicted hierarchy that the most prototypical representations of the genre would come from the AE group, followed by the AK group, and lastly the KK group. These results suggest that Korean authors accepted for publication in American biomedical journals have better learned to adapt and/or edit their language to meet the expected conventions of native English writers within the genre.

With regard to the comparison between texts published in *American journals* versus *Korean journals*, the results provide evidence that abstracts accepted for publication in American biomedical journals use significantly (1) more *total number of words,* (2) greater *syntactic diversity,* more (3) *cardinal numbers,* (4) *locational items,* (5) *temporal items,* (6) *argument overlap,* (7) *noun overlap,* (8) *content word overlap,*(9) *stem overlap*, (10) *-to infinitives,* and (11) *Wh-adverbs* than the those in Korean journals. In a comparison of *American L1* and *Korean L2* biomedical authors, American authors use (1) more *words per abstract,* (2) more *low frequency words,* (3) fewer *intentional events,* greater *syntactic diversity,* (4) more *cardinal numbers*, more *temporal cohesive devices,* (5) a smaller range of *word familiarity,* (6) fewer *3rd person verb morphemes*, more *–to infinitives*, and (7) more *Wh-adverbs* than Koreans.

Given that the differences between American biomedical abstracts published in American journals and Korean biomedical abstracts published in Korean journals (and possibly denied publication in American journals) differ significantly at each level (*word*, *sentence*, and *discourse*), it is reasonable to conclude that a NS-reviewer, editor, or reader may interpret Korean abstracts as lacking in key areas of the expected linguistic and rhetorical conventions of the genre. Furthermore, failure to meet the expectations of the genre may signal to readers within the community that an author does not belong and should not be granted central membership. As a result of the prominence of the abstract, native English speaker reviewers may correlate the research paper and its results to their reading of the abstract and review the entire paper with more skepticism. It is therefore reasonable to assume that the differences in linguistic and discourse styles limit Koreans' chances of gaining optimal acceptance into American biomedical journals.

Benefits of the Study

While international biomedical editors may believe that their journals actively encourage publication from L2 speakers, promote cultural diversity and do not consider English language difficulties as a reason for exclusion, the findings of this study show that significant differences at the word-, sentence-, and discourse-level appear between the research published in Korean and American journals. Therefore, what biomedical editors view as English language difficulties may in fact be limited to superficial matters of spelling, vocabulary, grammar, and punctuation. Larger issues, such as textual cohesion, syntactic diversity, and other strategic appeals used to situate one's research within the conventional expectations of the biomedical discourse community may be so ingrained in editors' subconscious as to be overlooked and unexamined in the determination of L2 speakers' submissions. Simply put, the line between language and content, form and function is likely not as clear as L1 reviewers wish it to be.

Tacit linguistic and discourse generic conventions at the word-, sentence- and discourse-levels distinguish what is published and what is not by top-tier biomedical journals. The differences outlined in this study may assist Korean biomedical researchers in matching their writing to the generic conventions, but it should also be used to help L1 editors to assess L2 speakers' submissions. An expanded and more explicit understanding of the genre's conventions is required on both sides.

An additional benefit of the current study is the direction it suggests for materials development for L2 and English for Specific Purposes (ESP) teachers. As demonstrated in the current study, a large corpus of biomedical abstracts is freely available on databases such as Medline. Thus, one immediate pedagogical implication

for this study is that educators (as well as researchers themselves) may collect corpora and use the techniques highlighted in this study to better assess the degree to which their text corresponds to a prototypical or desired target text type. Indeed, publishers, such as Cambridge University Press, have begun to encourage the use of corpus examination in the preparation of commercial materials (Moore 2005), but the use of corpus study can and should be much more widely applied.

Acknowledgments

This research was supported in part by the Department of English at the University of Memphis and the Institute for Education Sciences (IES R305G020018-02). Opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the university or IES. The authors would also like to acknowledge Drs. Philip McCarthy, Danielle McNamara, Emily Thrush, Reginald Martin, and Teresa Dalle for their contributions.

References

Biber, D. 1988. *Variations Across Speech and Writing*. Cambridge: Cambridge University Press.

Brill, E. 1992. A Simple Rule-based Part of Speech Tagger. In Proceedings of the Workshop on Speech and Natural Language, 112-116. Morristown, NJ: Association for Computational Linguistics.

Coltheart, M. 1981. The MRC Psycholinguistic Database Quarterly. *Journal of Experimental Psychology*, 33A: 497-505.

Graesser, A.C. 1994. Question Asking and Answering in *Handbook of Psycholinguistics.* San Diego, CA: Academic Press.

Graesser, A.C., McNamara, D.S., Louwerse, M.M., and Cai, Z. 2004. Coh-Metrix: Analysis of Text on Cohesion and Language. *Behavior Research Methods, Instruments and Computers, 36*: 193-202.

Jurafsky, D.S. and Martin, J. H. 2000. *Speech and Language Processing*. Englewood, NJ: Prentice Hall.

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

Landauer, T.K., McNamara, D., Dennis, S., and Kintsch, W. 2007. *LSA: A Road to Meaning*. Mahwah, NJ: Erlbaum.

Lillis, T. and Curry, M. 2006. Re-Framing Notions of 'Competence' in Multilingual Scholarly Writing. *Revista Canaria de Estudios Ingleses, 53*: 63 – 78.

McCarthy, P.M., Graesser, A.C., and McNamara, D.S. 2006. Distinguishing Genre Using Coh-Metrix Indices of Cohesion. Paper presented at the *Society for Text and Discourse Conference*, Minneapolis, MN.

McCarthy, P.M., Lehenbauer, B.M., Hall, C., Duran, N., Fujiwara, Y., and McNamara, D.S. 2007. A Coh-Metrix Analysis of Discourse Variation in the Texts of Japanese, American, and British Scientists.' *Foreign Languages for Specific Purposes*, *6*: 46-77.

Miller, G.A. 1990. WordNet: An On-line Lexical Database. *International Journal of Lexicography, 3*: 235-312.

Sekine, S. and Grishman, R. 1995. A Corpus-based Probabilistic Grammar with Only Two Non-terminals. In Proceedings of the Fourth International Workshop on Parsing Technology, 216-223. Prague, the Czech Republic: ACL/SIGPARSE.