Corpus Analysis and Its Applications in ELT

Budsaba Kanoksilapatham

Abstract

What is a corpus? What are its primary characteristics? Corpus analysis can shed light on research questions in linguistics and English language teaching. Corpus-based analyses, like move analysis and multidimensional analysis, are highlighted to demystify linguistic complexity in the target discourse.

Introduction

Corpus-based studies have become increasingly common and popular (e.g., Atkinson, 1999; Biber, 1988; Connor-Linton, 2001; Conrad & Biber, 2001; Hyland, 1998, 2001; Reppen, Fitzmaurice, & Biber, 2002). · With the help of computational tools, many corpora have been compiled or created, including Brown Corpus, Lancaster-Oslo/Bergen (LOB) Corpus, London-Lund Corpus or LLC, the British National Corpus or BNC, and the recent corpora like Michigan Corpus of Academic Spoken English or MICASE, the American National Corpus or ANC (see more details in Kennedy, 1998). Some corpora are specialized focusing on a specific genre (e.g., Biber & Finegan, 1994; Conrad, 1996; Cortes, 2002; Giannoni, 2002; Gledhill, 2000; Kanoksilapatham, 2003; Samraj, 2002). Be it specialized or general corpus, corpus analysis has shed light onto many new facts about language use.

Despite the enormous number of corpus-driven and corpus-based studies, many misconceptions or incomplete understandings about corpora and corpus analyses prevail, leading to detrimental consequences of obtaining unreliable and invalid findings and limited generalizations. To enhance the quality of future work along this line of research, this paper has two principal objectives. First, this paper addresses crucial characteristics of corpus analysis by exemplifying how to compile representative corpora. Second, the paper highlights move analysis and multidimensional analysis as analytical frameworks to elucidate certain facts about language use and linguistic complexity in the target discourse.

Corpus characteristics

This section addresses the two central issues: what a corpus is and what the primary characteristics of a corpus are. According to dictionaries, a corpus is

> *...body, collection, especially of writings on a specified subject or material for study* (Oxford Dictionary of English)
>
> *... a collection of all the writing of a particular kind by a particular person; a collection of information or materials to be studied*
> (Longman Dictionary of Contemporary English)

How adequate are these definitions in practice? Consider these examples:

- a large number of newspaper clippings collected to analyze language use in news reporting;
- a large volume of recorded conversations of two Japanese businessmen negotiating in a sound lab; and
- a series of 500 exchanges between a baby and the babysitter with the researcher monitoring the recording and sporadically initiating the conversational topics to the caretaker.

What is wrong with the above cases of corpus design? Yes, all of them are sizable, but other pertinent questions arise. Are they collected in a principled and systematic manner? Do they reflect natural or actual texts? For instance, a newspaper is a collection of a variety of writing genres, for example, editorials, movie review, news reporting, weather forecast, and classified ads. Are these linguistically and homogeneously comparable, representing the genre of news reporting? How can the recordings be natural when the two Japanese men were conducting conversations in a sound lab? How can the output from the child be natural or authentic when the researcher, unknown to the child, is there to observe and monitor the situation? What are the consequences if an analyzed corpus is not natural or authentic? The findings obtained are not going to be reliable or valid. In addition, generalizations can be limited.

Thus the misconceptions regarding the notion of corpus prevail. Size seems to be the first thing people think of in connection with corpora. A number of research works claim how big their corpora are. This criterion seems to out-prioritize other requirements. Giving priority to the size of the corpus would not mean much if the other required characteristics are ignored. By the same token, size does not indicate representativeness and vice versa. In fact, Kennedy (1998) deemed the issue of representativeness more important than size. How can representativeness be achieved? As Biber et al. (2001) contends, corpora should be large and principled collections of natural or actual texts or spoken language output. The definition implies that the corpus can be said to be representative if the means of collecting language output are principled and systematic.

Corpus compilation

In this section, the two corpora compiled by the researcher are exemplified to illustrate how a corpus should be sizable and at the same time representative. The first corpus consists of 60 biochemistry research articles, whereas the second corpus consists of 60 microbiology research articles.* The two corpora are comparable in size (about 300,000 words after editing) and are compiled following the same principles. The factors taken into consideration when designing corpora are as follows:

(a) Specialized vs general corpus

People may wonder if we should use a specialized corpus as opposed to a general corpus (e.g., LOB, BNC, ANC, MICASE). Caveats are in order. The choice of specialized or general corpora is entirely up to the research questions being addressed, the scope of the research, etc. For instance, once the research question of "*How are research articles constructed?*" is formulated, we can proceed to the next step of designing the corpus.

(b) Academic journals

As we all know, academic journals have become one of the main channels of communication among scholars across disciplines in today's world. Particularly in science, the role of academic journals is more prominent, witnessed by their frequent publications (e.g., weekly, bi-weekly, monthly, etc.). Taking into consideration of our research question formulated as shown above, our choice of academic journals is justified. However, caution need to be exercised regarding the components of an academic journal. Usually, academic journals, in terms of their contents, are not homogeneous but mixed consisting of review articles and experimental research articles. Therefore, to answer our research question, only experimental research articles are focused on and included in the corpus.

(c) (Sub)Disciplines

Science is a big umbrella term comprised of many sub-disciplines like pure sciences, physical sciences, health sciences, applied sciences, etc. Moreover, move-based studies that uncovered a rhetorical pattern of each conventional section of research articles have shown that rhetorical structure varies according to academic disciplines (e.g., Brett, 1994; Posteguillo, 1999, Swales & Najjar, 1987; Thompson, 1993, Williams, 1999). Therefore, the author opts to focus on biochemistry/microbiology research articles.

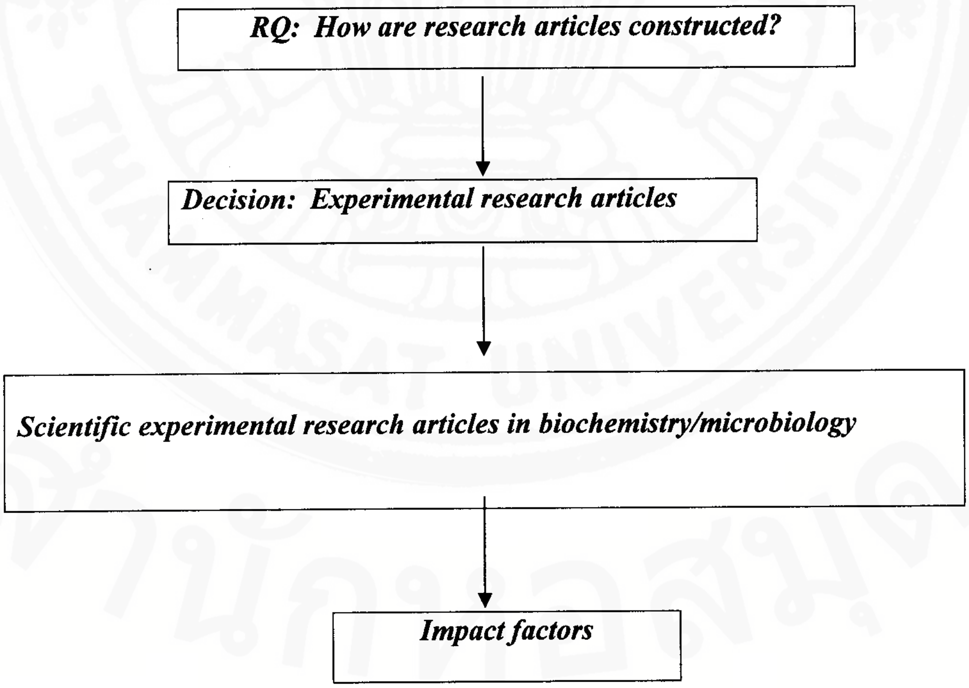
(d) Selected journals

Not all biochemistry/microbiology journals are equally recognized and respected. Some journals are deemed to be highly prestigious while others are not. How can one be assured that the selected journals in the corpus represent the most prestigious ones in the target discipline? Thanks to the availability of impact factors as indicators of how prestigeous a journal is, the selection of biochemistry/microbiology research articles can be objective. Based on the impact factors, the top five biochemistry/microbiology journals are selected.

Two comparable corpora: 60 articles or about 320,000 words (about 1000 pages) representing the articles from the most prestigious journals in the target disciplines of biochemistry and microbiology.

Conclusion: Sizable? Yes. Representative? Yes Authentic or natural? Yes

The following diagram recaps how the two comparable corpora are designed:


RQ: How are research articles constructed?
Decision: Experimental research articles
Scientific experimental research articles in biochemistry/microbiology
Impact factors

Corpus analysis: Move analysis and multidimensional analysis

This section illustrates how one of the corpora is analyzed using the two principal discourse approaches of move analysis and multidimensional analysis and what insights they provide.

Move analysis

Move analysis was created by Swales (1990) as an analytical framework to analyze the Introduction section of research articles from multiple disciplines. In a nutshell, as a qualitative approach to discourse analysis, texts are analyzed into moves, a portion of a text that has a communicative function. The framework has been extended to analyze other sections of research articles (e.g., Yang & Allison, 2003; Brett, 1994; Kanoksilapatham, 2003, Nwogu, 1997; Posteguillo, 1999; Samraj, 2002; Williams, 1999) and to other genres such as textbooks (e.g., Nwogu, 1991) and university lectures (e.g.,Thompson, 1994).

Move analysis conducted on the biochemistry research article corpus reveals a rhetorical structure of, for instance, the Introduction section. The section typically consists of three moves namely, Move 1: Announcing the importance of the field/topic (by claiming the centrality of the topic, making topic generalizations, and reviewing previous research), Move 2: Preparing for the present study (by identifying a gap in previous research work or raising a question), and Move 3: Introducing the present study. Each move has its variations or so-called steps. The rhetorical pattern commonly found in the biochemistry corpus is delineated in the following:

Introduction

Move 1: Announcing the importance of the field

By Step 1: Claiming the centrality of the topic
Step 2: Making topic generalizations
Step 3: Reviewing previous research

Move 2: Preparing for the present study

By Step 1: Indicating a gap
Step 2: Raising a question

Move 3: Introducing the present study

By Step 1: Stating purposes
Step 2: Describing procedures
Step 3: Presenting findings

Figure 1: Rhetorical pattern of biochemistry Introductions (Kanoksilapatham, 2004)

The pattern identified provides a template of what elements are usually included when biochemistry introductions are constructed.

Multidimensional analysis

Multidimensional analysis, another corpus-based analysis, is both qualitative and quantitative in nature (e.g., Biber, 1988, Biber & Finegan, 1994, Connor-Linton, 2001). The analytical framework was invented and fully developed by Biber in 1988. The framework involves many succinct steps requiring computational tools to process a large volume of linguistic data. In a nutshell, the framework involves the following major steps.

In the first and most crucial step, linguistic features are selected. The framework has been criticized for focusing on what the researcher wants to focus on, and this means that some linguistic features that might play a prominent role in the corpus might be overlooked. One set of linguistic features selected might be appropriate for one corpus but not the other. For example, the feature of contractions (e.g., *isn't, can't*) and sentence fillers (e.g., *well, er...*) might be salient in spoken corpus but not in written corpus. Therefore, as an initial and crucial step, it is recommended that the researcher know his/her corpus well so that only salient linguistic features of the corpus in focus are selected for further analysis. In addition, since the analysis involves a sizable corpus and relies on a computational tool, the decision on what linguistic features to be analyzed is partly determined by the availability of the computer programs to analyze selected linguistic features.

Next, after the selection of the linguistic features, the corpus is tagged accordingly. The corpus is marked up with the identifiable lexical or syntactic categories in such a way to be recognized by a computer, and the computer can count how frequently the linguistic features occur in the corpus. The frequencies of the occurrence of these linguistic features serve as variables to be analyzed by factor analysis.

Factor analysis is performed to identify the patterns of linguistic features that tend to occur frequently in the corpus. Based on the assumption that linguistic features co-occur because they help perform a communicative function, the co-occurrence pattern of linguistic features is subsequently interpreted functionally and called a dimension. The dimension score for each text is calculated. Finally, the mean dimension score of each dimension is calculated and plotted to reveal textual variation and relationships for each dimension. For more details regarding multidimensional analysis, please refer to Biber (1988, 2001).

How can multidimensional analysis demystify linguistic complexity in scientific discourse?

Language, be it spoken or written, is complex, reflecting an interaction and manifestation of linguistic features conveying a message. How linguistic features interact with each other is subtle and implicit. As mentioned earlier, often times, we rely too much on our intuition or anecdotal evidence. Unfortunately, our intuition can be unreliable and probably based on a few speakers' idiosyncrasies. As English has the status of a foreign language in Thailand, Thai learners' intuition about English does not help much. Moreover, our understanding of how linguistic features interact with each other is limited by rigid and heavy reliance on our grammatical knowledge of discrete linguistic features through many years of formal education of English and through commercial grammar textbooks. The following section reveals how multidimensional analysis conducted on the biochemistry research article corpus can shed some light on the interplay of linguistic features in scientific discourse.

Forty-one aggregated linguistic features (lexical features, grammatical features, and syntactic constructions) were selected, tagged, and counted. The objective of multidimensional analysis is to identify clusters or sets of linguistic features that tend to occur in a particular text by using factor analysis, taking the frequencies of these features as variables. Based on the assumption that linguistic features co-occur because they share the same communicative function, the co-occurrence of linguistic features is interpreted for a communicative function in discourse. The following sections present selected sets of linguistic features identified in biochemistry corpus, their interpreted communicative functions, and representative excerpts taken from the corpus to illustrate such co-occurrences. The highlighted features here include passive constructions, past tense verbs, extraposed 'it', *that* complement clauses controlled by predicative adjectives, *to* complement clauses controlled by adjectives, present tense verbs, references, and pointers.

Passive constructions and past tense verbs

The analysis reveals that passives and past tense verbs tend to occur quite frequently in biochemistry corpus. The functions of individual features are considered, then, the shared communicative function of these features is interpreted.

Passive constructions, both agentless passives and *by-* passives, are crucial elements in scientific discourse (e.g., Bazerman, 1988; Hanania & Akhtar, 1985; Riley, 1991; Swales, 1990, Tarone *et al.*, 1981; Trimble & Trimble, 1982; Wilkinson, 1992; Wingard, 1981). According to these scholars, this feature is typically and most effectively used when the emphasis is on the actions, and the role of the agent is downgraded (in *by*-passives) or omitted (in agentless passives) in the discourse.

Past tense verbs index another common rhetorical strategy used in ESP, describing research activities or procedures performed. The co-occurrence of passives and past tense verbs suggests their primary pragmatic function is describing scientific research activities.

The following text sample taken from the corpus illustrates the set of co-occurring features: passives (bolded) and past tense verbs (underlined).

TEXT SAMPLE 1
<Jbc04Jul01>
Unlabeled acetyl-CoA **was purchased** from Amersham Pharmacia Biotech; desulfo-CoA, acetylated BSA, and trypsin **were obtained** from Sigma. Other reagents **were obtained** from commercial sources unless otherwise described.

In conclusion, the features of passive and past tense verbs are used by the author to recount scientific activities.

Past tense verbs and pointers

Pointers are a linguistic feature uniquely selected for multidimensional analysis due to their prevalence in the corpus. Pointers refer to metatexual devices directing readers to the source of information (e.g., *see Table 1, Data not available, see Figure 8A).* The co-occurrence of passive verbs and pointers is functionally explicit. It reflects a focus on current findings produced by the study being reported.

Text sample 2 is a typical example illustrating the use of pointers (underlined) and past tense verbs (bolded).

TEXT SAMPLE 2
<cel10Jan32>
Moreover, these troughs **were** labeled with antibodies against -catenin (Pointer) and **were** flanked by desmosomes associated with thick bundles of keratin intermediate filaments (Pointer). At late times, the undulating cell–cell border **had** flattened, and the epithelium **appeared** as a sheet, with continuous contacts of alternating desmosomes and Adherens junctions (Pointer).

Text sample 2 has instances of two pointers and many past tense verbs. Pointers are tied to the results produced by the study being reported, providing visual accompaniments to the statement of results. As observed by Oster (1981), past tense usage in scientific discourse reports completed actions at the particular time frame. That is, the past tense is used to express newly achieved evidentiality or knowledge in science, reflecting limited degree of certainty and reliability of the new information. Since no generalization is assumed, this new information or finding remains to be substantiated or validated (as opposed to present tense verbs that indicate the established status of the proposition).

Extraposed 'it', *that* complement clauses controlled by predicative adjectives, and *to* complement clauses controlled by adjectives

Another set of linguistic features co-occurring quite frequently include extraposed '*it*', *that* complement clauses controlled by predicative adjectives, and *to* complement clauses controlled by adjectives. In scientific discourse, the extraposed '*it*' provides a means for scientists to express their comments or attitudes without making their identification explicit (e.g., Biber *et al*., 1999; Hewings & Hewings, 2002; Rodman, 1991). The preference of extraposed '*it*' over the first person *('I'* or

'We') in academic writing, according to Craswell (2005), can persuade the readers to believe that the content of *that* clause is objectively presented.

Likewise, predicative adjectives provide the authors with a means to express their stance (e.g., Biber *et al.*, 1999; Soler, 2002). *That* complement clauses controlled by adjectives and *to* complement clauses controlled by adjectives indicate clearly that predicative adjectives are used as heads of *that* or *to* complement clauses, indexing an expression of the authors' stance. *That* complement clauses are generally known to index information integration to expand the idea-unit (e.g., Biber, 1988). Winter (1982) notes that *that* complements provide a means to talk about the information in the dependent clause. That is, the authors' stance is given in the main clause, and the propositional information is given in the *that* complement clause (e.g., *it is possible that we don't detect...*). The stance towards propositions can be characterized as interpretation, attitude, argumentation or generalization.

The adjectives that control *that* complement clauses are particularly likelihood adjectives (e.g., *likely, possible, probable*), attitudinal adjectives (e.g., *interesting, acceptable, necessary*), and factual/certainty adjectives (e.g., *impossible, evident, obvious*). This indicates that these co-occurring features index the authors' expression of their agreement, opposition, evaluation, and interpretation of propositions.

Similarly, *to* complement clauses controlled by predicative adjectives are another feature. The semantic class of controlling predicative adjectives are evaluative adjectives (e.g., *appropriate, important, essential, necessary*) and ease/difficulty adjectives (e.g., *difficult, easy, impossible*). Again, the co-occurrence of these predicative adjectives and *to* complement clauses represents the authors' appraisal of, and the authors' ease or difficulty with, propositions in complement clauses.

Text samples 3 and 4 from the corpus illustrate the set of co-occurring features.

TEXT SAMPLE 3

<Moc14Nov01>

It is *interesting* that the experiments in this paper were all carried out using assays for genetic interference in somatic tissues of the animal in the first generation after injection. **It** is *conceivable* that distinct mechanisms might operate in longer term RNAi (Ref.) or in specific tissues, such as the germline.

TEXT SAMPLE 4

<Mbc14Feb01>

In the absence of an atomic structure, **it** is not *possible* TO determine which residues are solvent exposed and thus likely to make physical contact with the microtubule and which ones contribute to the domain's structural organization.

The two text samples illustrate the use of extraposed '*it*' (bolded), *that* clauses controlled by adjectives (underlined), predicative adjectives (italicized), and *to* clauses controlled by adjectives (capitalized). These features work together to create a text that conveys the authors' evaluative stance.

Taken together, the co-occurrence of these linguistic features (extraposed 'it', *that* complement clauses controlled by adjectives, predicative adjectives, and *to* complement clauses controlled by adjectives) indexes the scientists' personal stance towards the propositions in the *that/to* complement clauses in an impersonal way. That is, their personal stance is backgrounded and not directly attributed to specific individuals. Therefore, the interpretive label '*Evaluative Stance*' is proposed for the functional dimension underlying this co-occurrence.

Present tense verbs and references

Present tense verbs, as summarized by Swales (1990), as used in a professional genre have two main pragmatic functions: 1) to situate a particular event in the present tense, and 2) to mark a particular proposition as a generalization. In the latter case, the use of the present tense indicates that the propositional information is valid regardless of time. References or citations, a device essential for academic attribution, have been investigated by many scholars (e.g., Dong, 1996; Hyland, 1999; 2001; Salager-Meyer, 1999). The co-occurrence of reference with present tense verbs suggests that the latter of the two pragmatic functions of present tense is more relevant to the functional interpretation of this cluster of features--to index generalized background knowledge established by previous research in the field.

The following text sample illustrates the set of co-occurring features. The sample illustrates the use of present tense verbs (bolded) and references (underlined). These features work together to create a text that conveys attributed information.

TEXT SAMPLE 5
<Mbc01Jul01>
Interest in prenylation has stemmed from the discovery that key proteins in multiple signal transduction cascades **contain** covalently attached isoprenoids (Ref.). Perhaps the most notable examples **are** the Ras proteins. Mutated forms of Ras proteins **are** found in 30% of all human tumors (Ref.). However, these mutant Ras proteins are not oncogenic if they **cannot** be prenylated (Ref.). Prevention of Ras prenylation thus **holds** promise as a new tactic for cancer chemotherapy (Ref.). To this end, many prenylation inhibitors have been developed, several of which **appear** to be effective anticancer agents in animal studies and **are** undergoing clinical trials (Ref.).

Text sample 5 contains many instances of present tense verbs (*contain, be, hold, appear*). According to Biber *et al*. (1999), these verbs belong to the semantic domains of relationship verbs (e.g., *contain*), existence verbs (e.g., *be*, *appear*), and aspectual verbs (e.g., *hold* promise). The present tense correlates with these categorizations of verbs indicating general time or implying a lack of time restriction in scientific writing. That is, propositional information presented (such as generic background information or general characterization of *key proteins in multiple signal transduction cascades*, examples of *key proteins*) is true or valid regardless of time.

Text sample 5 also contains many instances of "references," demonstrating citation practices in biochemistry. References or citations reflect the intellectual tribute to previous researchers for having provided information that can be utilized in a productive way. This device helps ensure that the knowledge-manufacturing of science is efficient and productive. Although present tense verbs help perform more than one underlying communicative function (e.g., Swales, 1990), the co-occurrence of present tense verbs with reference features helps indicate that a propositional information has been established by previous research studies. References, as used in text sample 5, provide sources of generic background information or general truths that have temporal neutrality. Taken together, the co-occurrence of present tense verbs and citations represent attributed knowledge, a crucial requirement in scientific discourse to situate and contextualize the study being reported, giving credence to the source of information.

Overall, multidimensional analysis applied to this specialized corpus reveals certain sets of co-occurrence of linguistic features that contribute differently in biochemistry discourse.

Pedagogical implications

The analysis of the corpus using two approaches in discourse analysis is pedagogically beneficial in language teaching in general and in particular in the instruction of reading and writing academic research articles in a number of ways. First, the analyses can help inform strategies in teaching reading and writing research articles. The study demonstrates that teaching advanced learners academic English can be bi-directional. That is, the top-down instructional approach can be adopted, relying on the results generated from move analysis. The rhetorical structure of each section provides the learners with the template to follow and to anticipate when reading scientific research articles resulting in facilitated access to scientific information presented. In writing, the template helps the learners know what content to include, enabling them to write research articles in a manner conforming to their respective discourse community's expectation.

Meanwhile, the linguistic characterization yielded from multidimensional analysis also has a significant pedagogical impact. It allows teachers to adopt bottom-up instructional approach focusing on linguistic features that co-occur to convey information in academic discourse. Given the fact that move analysis provides the template of what to be included, multidimensional analysis provides insights onto how to express those ideas linguistically. Both approaches to discourse analysis are thus complementary, giving learners a clearer portrait of scientific writing both at the macro level of the rhetorical pattern and the micro level of clusters of linguistic features characterizing such rhetorical moves.

This study also informs decisions concerning the design of academic programs and other related pedagogical issues such as the development of teaching materials and test materials. The results generated from multidimensional analysis enable language teachers to decide what linguistic features to teach and how to teach them. Obviously, prominent and salient features identified by multidimensional analysis receive our immediate attention. The analysis also informs us that the function of linguistic features at the discourse level can be different and deviate from what grammar books typically prescribe. For instance, the choice of tense usually represents the time line. However, as shown earlier, tense usage in academic discourse depends on the generality of the proposition. Likewise, extraposed '*it*' constructions is not simply the transformation of a sentence. In fact, the construction allows scientists to express their opinions without feeling too explicit or too committed to the proposition presented. It should be pointed out as well that each individual linguistic feature has its own meaning; however, when it is used in co-occurrence with other features, its original function might shift to accommodate the member set. Therefore, in order to accurately understand academic discourse, grammatical knowledge beyond the sentence level is imperative.

Conclusion

In conclusion, this paper demonstrates the characteristics of representative corpora focusing on research articles, the canonical form of communication of scientific findings. This paper also shows that corpus analyses like move analysis and multidimensional analysis can enhance our understanding of how a representative corpus of biochemistry research articles is constructed both at the macro (rhetorical moves) and micro (linguistic features) levels. This study's findings offer practical implications to advanced language learners and teachers interested in pedagogy in reading and writing instruction.

Note

* The corpus was a part of the author's dissertation work supported by the National Science Foundation (NSF), USA, under the Grant No. 0213948 and TOEFL Grant for Doctoral Research from the ETS. The second corpus is a part of the author's on-going research work supported by the Thailand Research Fund (TRF), Thailand. The two comparable corpora were compiled and analyzed primarily by move analysis, a discourse analysis proposed by Swales. The qualitative nature of move analysis precludes the enormous size of the corpus. However, the corpus of 60 research articles was considered to be the biggest corpus ever analyzed by move analysis.

References

Atkinson, D. 1999. Scientific discourse in sociohistorical context: The Philosophical Transactions of the Royal Society of London, 1675-1975. Hillsdale, New Jersey: Lawrence Erlbaum.

Bazerman, C. 1988. Shaping written knowledge: Studies in the genre and activity of the experimental article in science. Madison: University of Wisconsin Press.

Biber, D. 1988. Variation across speech and writing. Cambridge: Cambridge University Press.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. 1999. Longman grammar of spoken and written English. Harlow: Longman.

Biber, D. 1995. Dimensions of register variation: A cross-linguistic comparison. Cambridge, New York: Cambridge University Press.

Biber, D 1993. Representativeness in corpus design. Literary and Linguistic Computing, 8, 4, 243-257.

Biber, D. & Finegan, E. 1994. Intra-textual variation within medical research articles. In N. Oostidijk & P. de Haan (Eds.), Corpus-based research into language (pp. 201-221). Amsterdam, Netherlands: Rodopi.

Brett, P. 1994. A genre analysis of the results section of sociology articles. English for Specific Purposes, 13, 1, 47-59.

Connor-Linton, J. 2001. Author's style and worldview: A comparison of texts about American nuclear arms policy. In D. Biber & S. Conrad (Eds.), Variation in English: Multi-dimensional studies (pp.84-93). Harlow: Pearson.

Conrad, S. 1996. Investigating academic texts with corpus based technique. Linguistics and Education, 8, 299-326.

Conrad, S & Biber, D. 2001. Variation in English: Multidimensional studies. Pearson Education. English.

Cortes, V. 2002. Lexical bundles in freshman composition. In R. Reppen, Fitzmaurice, & D. Douglas (Eds.), Using corpora to explore linguistic variation (pp. 131-146). Philadelphia: Johns Benjamin.

Craswell, G. 2005. Writing for academic success. Sage Publications: London.

Dong, Yu-Ren. 1996. Learning how to use citations for knowledge transformations: Non native doctor students' dissertation writing in science. Research in the Teaching of English, 30, 4, 428-457.

Giannoni, D. 2002. Worlds of gratitude: A contrastive study of acknowledgement texts in English and Italian research articles. Applied Linguistics, 23, 1, 1-31.

Gledhill, C. 2000. The discourse functions of collocation in research article introductions. English for Specific Purposes, 19, 115-135.

Hanania, E. & Akhtar, K. 1985. Verb form and rhetorical function in science writing: A study of MS theses in biology, chemistry, and physics. English for Specific Purposes, 4, 49-58.

Hewings, M. & Hewings, A. 2002. "It is interesting to note that ...": A comparative study of anticipatory 'it' in student and published writing. English for Specific Purposes, 21, 367-383.

Hyland, K. 2001. Humble servants of the discipline? Self mention in research articles. English for Specific Purposes, 20, 207-226.

Hyland, K. 1999. Academic attribution: Citation and the construction of disciplinary knowledge. Applied Linguistics, 20, 3, 341-356.

Hyland, K. 1998. Hedging in scientific research articles. Amsterdam: Johns Benjamins.

Kanoksilapatham, 2004. Rhetorical structure of biochemistry research articles. English for Specific Purposes. (in print).

Kanoksilapatham, 2003. Corpus-based investigation of biochemistry research articles: Linking move analysis with multidimensional analysis. Unpublished dissertation. Georgetown University: Washington, D.C.

Kennedy, G. 1998. Corpus linguistics. Oxford University Press. New York.

Longman dictionary of contemporary English (3rd Edition). 1995. Oxford University Press. New York.

Nwogu, K. 1997. The medical research paper: Structure and functions. English for Specific Purposes, 16, 2, 119-138.

Nwogu, K. 1991. Structure of science popularizations: A genre-analysis approach to the schema of popularized medical texts. English for Specific Purposes, 10, 2, 111-123.

Oster, S. 1981. The use of tenses in "reporting past literature" in EST. In L. Selinker, E. Tarone, & V. Hanzali (Eds.), English for academic and technical purposes (pp. 76-90). Rowley, MA: Newbury House.

Oxford Dictionary of English. 1995. Oxford University Press. New York.

Posteguillo, S. 1999. The schematic structure of computer science research articles. English for Specific Purposes, 18, 2, 139-158.

Reppen, R. 2001. Register variation in student and adult speech and writing. In D. Biber & S. Conrad (Eds.), Variation in English: Multi-dimensional studies. New York: Longman.

Riley, K. 1991. Passive voice and rhetorical role in scientific writing. Technical Writing and Communication, 21, 3, 239-257.

Rodman, L. 1991. Anticipatory it in scientific discourse. Journal of Technical Writing and Communication, 21,1, 17-27.

Salager-Meyer, F. 1999. Referential behavior in scientific writing: A diachronic study (1810-1995). English for Specific Purposes, 18, 3, 279-306.

Samraj, B. 2002. Introductions in research articles: Variations across disciplines. English for Specific Purposes, 21, 1-17.

Soler, V. 2002. Analysing adjectives in scientific discourse: An exploratory study with educational applications for Spanish speakers at advanced university level. English for Specific Purposes, 21, 145-165.

Swales, J. 1990. Genre analysis: English is academic and research settings. Cambridge: Cambridge University Press.

Swales, J. & Najjar, H. 1987. The writing of research article introductions. Written Communication, 4, 175-192.

Tarone, E., Dwyer, S., Gilette, S., & Icke, V. 1981. On the use of the passive in two astrophysics journal papers. English for Specific Purposes, 1, 2, 123-140. Reprinted in J. Swales, 1985. Episodes in ESP. Oxford: Pergamon.

Thomas, S. & Hawes, T. 1994. Reporting verbs in medical journal articles. English for Specific Purposes, 13, 2, 129-148.

Thompson, D. 1993. Arguing for experimental facts in science. Written Communication, 10, 1, 106-128.

Trimble, M. & Trimble, L. 1982. Rhetorical-grammatical features of scientific and technical texts as a major factor in written ESP communication. In J. Hoedt et al. (Eds.), Proceedings of the third European symposium on LSP, Copenhagen 1981: Pragmatics and LSP (pp. 199-216). Copenhagen: Villadsen &Christensen.

Yang, R. & Allison, D. (2003). Research articles in applied linguistics: Moving from results to conclusions. English for Specific Purposes, 22, 365-385.

Wilkinson, A. 1992. Jargon and the passive voice: Prescriptions and proscriptions for scientific writing. Technical Writing and Communication, 22, 3, 319-325.

Williams, A. 1999. Results section of medical research articles: An analysis of rhetorical categories for pedagogical purpose. English for Specific Purposes, 18, 4, 347-366.

Winter, E. 1982. Towards a contextual grammar of English: The class and its place in the definition of sentence. London: George Allen and Unwin.

Wingard, P. 1981. Some verb forms and functions in six medical texts. In L. Selinker, E. Tarone, & V. Hanzali (Eds.), English for academic and technical purposes (pp. 53-64). Rowley, MA: Newbury House.

Yang, R. & Allison, D. (2003). Research articles in applied linguistics: Moving from results to conclusions. English for Specific Purposes, 22, 365-385.
385.