Kwansei Gakuin University
Humanities Review
Vol.19, 2014
Nishinomiya, Japan

An Analysis of Lexicogrammar in Samples From a Corpus of Science Texts

Daniel PARSONS*[1]

Abstract

The sampling methodology for building a corpus depends on the purpose of the corpus and the community from which the sample texts are a product (Sinclair, 2005). A corpus of science texts currently being built at the Faculty of Science and Technology, Kwansei Gakuen Unversity, contains samples of undergraduate and graduate textbooks as well as research articles. This paper recognises that textbooks and research articles are products of similar academic communities but have different purposes. A genre analysis, employing a systemic functional linguistics framework, is carried out to compare the introduction sections of one research article and one textbook chapter from the corpus. The comparison shows that there is a significant different between the use of process types, and no significant difference in the use of Theme and complexity of Thematic Heads. The implications for sampling methodology is discussed.

Introduction

Sampling for the Corpus of Science Texts

When building a corpus for a specific variety of language and for a specific purpose or research question, issues of corpus size, diversity of texts, length and number of samples, and, of course, representativeness and balance of the samples all need to be taken into consideration. (Biber, *et al.* 1998; McEnery *et al.* 2006; O'Keeffe *et al.* 2007; McEnery & Hardy 2012, Clancy 2010). McEnery *et al.* (2006) relate representativeness and balance to the sampling methodology employed and the principled way in which choices regarding text size and diversity are made. This is being considered for the corpus of science texts currently being built at Kwansei Gakuin University's Faculty of Science and Technology.

[1] Instructor of English, Kwansei Gakuin Universi

Taking recommendations from McEnery *et al.* (2006), each scientific research article is being sampled within each stage of the IMRAD (Introduction, Method, Results and Discussion) genre. This choice of sampling methodology is based on two criteria for the corpus. The first is that the purpose of the corpus is to aid graduate students and professionals when writing scientific dissertations and papers, and preparing presentations. The second criteria leads from the first, and is based on the necessity that the texts represent the communicative functions within the community from which they are sampled (Sinclair, 2005). These two criteria, then, ensure that the corpus is both representative of its users and relevant to graduate students and professionals.

The corpus of science texts is also being constructed with samples from undergraduate and graduate textbooks. This decision is premised on the notion that basic technical collocations would be represented within these texts alongside lexical bundles which exhibit discourse functions such as imprecise reference, text deixis and topic elaboration (Biber, 2006). Again, the criteria for sampling from textbooks is based on Sinclair's (2005) recommendation that the corpus is as representative of the language as possible, in this case, a generalised scientific language. However, this criteria is at odds with the genre based stratified random sampling criteria described above. Specifically, it is not clear when sampling from the textbooks what communicative function each text has. It is difficult to define what community a textbook sample represents and thus it is necessary, as recommended by Reppen (2010), to resolve this issue before compiling this part of the corpus.

This issue can be resolved through a comparative genre analysis of research articles and textbooks. Genre analysis can reveal the connections between particular lexicogrammatical features and the purpose and context of a text (Martin, 2001). The texts - textbooks and research articles - are the products of two different though overlapping communities of practice (Wenger, 1998), one being concerned with education, the other with research. A comparative genre analysis can make explicit the differences in purpose within these two communities of practice, which will not only have implications for the sampling methodology, but also for the annotations and retrieval methods. Ultimately, it will provide an insight into the communities of practice which allow for more principled sampling decisions.

Genre, Sampling and Purpose

In Martin's (1997:13) definition of genre as context, genre is "the system of staged goal oriented social processes through which social subjects live their lives". Bhatia (2002) defines genre in linguistic terms, referring to the constraints imposed by a conventional setting on the lexicogrammatical choices available for that setting. Coffin (2001:110) describes how the structure of a genre may contain beginnings, middles and ends, but that these stages have distinct functions which vary depending on the overall social purpose of the text. In fact, Painter (2001) highlights the variation in linguistic features between a procedural text and an analytical exposition and thus shows explicitly how specific genres are associated with specific linguistic features. Painter reminds us of the Neo-Firthian interest in language and context, that "the context is 'created' by the language of the text" and that "the relationship between context and text is systematic and … two way" (Painter, 2001:178).

In this article, the introduction section of a research article will be compared with the introduction section of a chapter from a textbook. The analysis is based on the comparative analysis of newspaper reports and newspaper commentaries carried out by Lavid *et al.* (2012). Though the newspaper texts belong to related communities of practice, the analysis revealed that the different purposes of the texts are construed through significant differences in the lexicogrammatical features within the two texts. Following Lavid *et al.* (2012), I turn to systemic functional linguistics as the framework for analysis.

Systemic Functional Linguistics

SFL offers a description of how the context of a particular genre is construed through specific lexicogrammatical features (Halliday and Matthiessen, 2004). SFL describes three layers of meaning present in the clause: the ideational, realized by the grammatical system of transitivity, concerns human experience; the interpersonal, realized by the system of mood, concerns interaction; the textual, realized by the system of theme, concerns the message that runs through a text.

In the system of transitivity, a clause consists of the following: a process which usually includes a lexical verb; some participants within the process; circumstances associated with the process. An example is given below:

[We]	[will meet]	[my wife's parents]	[at nine o'clock].	[1]
Participant	Process	Participant	Circumstance	

Processes can be analysed into 6 categories which depend on the nature of our experience. These categories are "material, behavioural, mental, verbal, relational, existential", and the definitions of these are outlined in Halliday and Matthiessen (2004:170-71). Halliday (2004:185) identifies what he calls the "prototypical clause of modern scientific English", the features of which are as follows: (i) the structure is simple: nominal group + verbal group + nominal group; (ii) a relational process is construed by the clause to explain logical relationships between the nominal expressions; (iii) the nominal groups are often nominalisations of processes through grammatical metaphor.

The message of a text unfolds in the discourse through the system of theme. Textual meaning is realized in the message structure of Theme + Rheme in a clause. Theme is always put first in English. I take the model of Theme from Lavid *et al.* (2012) as the framework for analysis. In this model, Thematic Head is "the first nuclear experiential constituent within the main clause". Interpersonal Themes are the elements which express attitude and evaluation in clause-initial position, and Textual Themes are logical connections and textual markers in clause-initial position. The PreHead is any circumstantial or finite element preceding the Thematic Head. Anything after the Thematic Head is Rheme. This is a refined model from Halliday and Matthiessen (2004:79) who define the Theme of a clause as that part which "ends with the first constituent that is either participant, circumstance or process".

The fine-grained model of Theme defined by Lavid *et al.* (2012) allows us to quantify more specifically the resources that writers employ to control the discursive flow of texts. Therefore it is possible to more accurately compare an introduction from a textbook and an introduction from a research article. Furthermore, as Painter (2001) demonstrated with a simple example, comparison of process types can further elucidate the purpose of a text. In turn, this can demonstrate how writers deploy the resources of ideational meaning differently or similarly depending on the genre. Finally, Thematic Head is central to the unfolding of a text. The nominal group of the Thematic Head can be investigated for differences in levels of complexity between the two genres. Complexity here is taken from Lavid *et al.* (2012), and refers to the degree of pre- and post-modification of a nominal Head.

The three types of analyses outlined above, namely the analysis of the Thematic elements, the analysis of process types and the analysis of the complexity of Thematic Head, offer a detailed overview of the two genres. This detailed series of analyses should be able to sift out some of the lexicogrammatical differences between the two genres and thus begin to shed light on what constitutes a sample text from a textbook. With these in mind, three research questions are formed:

i) Is there a significant difference in the distribution of Thematic elements between the introduction in a textbook and the introduction in a research article?
ii) Is there a significant difference in the distribution of process types between the introduction in a textbook and introduction in a research article?
iii) Is there a significant difference in the complexity of nominal elements in Thematic Heads between the introduction in a textbook and introduction in a research article?

Methodology

Data

Two introduction texts were taken from the corpus, both in the field of chemistry and one from a research article, and another from a textbook. The textbook's Japanese translation is used by undergraduates and graduates in the faculty. Altogether, the sample consisted of 495 clauses and 15,807 tokens.

Procedure

The texts were segmented into clauses similarly to Halliday and Matthiessen (2004:101). Each main clause is an instance of the clause variable in this study. Subordinate clauses were also treated as instances since subordinate clauses contain thematic structure [2]. However, embedded bound clauses function within nominal groups, and so their thematic structure is ranked downwards and the contribution to the discourse is minimal [2]. Anaphoric elliptical clauses were included in coordinate clauses and the thematic structure assumed from the previous clause [3]. Following segmentation, Thematic elements were labelled and counted.

[Clause 59] Nuclear reactions are very much more energetic than normal chemical reactions
[Clause 60] *because* the strong force is much stronger than the electromagnetic force **that** binds electrons to nuclei. [2]

In [2] the subordinate clause 60 is bound to the main clause 59, but they are treated as separate clauses with their own Thematic Heads. The Thematic Heads are underlined and the textual Themes *italicised.* Notice how the relative clause in

[2] is not treated separately. This is because it functions on the nominal group with "force" as nominal Head.

[Clause 72] A neutrino is electrically neutral
[Clause 73] *and* has a very small (possibly zero) mass. [3]

The Thematic Head in clause 72 is counted again in clause 73. Although elliptical in clause 73, it is functioning in a different process from clause 72 and is thus discursively prominent. However, in the case when a subordinate clause acted as circumstantial PreHead to the Thematic Head, this subordinate clause was not treated separately [4]. This is in line with Lavid *et al*'s (2012) model of Theme outlined in section 1.3.

[Clause 68] **When it is emitted**, the mass number of the nuclide is unchanged. [4]

Clause 68 shows a subordinate clause in bold type acting as a thematic element, specifically a circumstantial PreHead. Other circumstantial PreHeads are also included in the analysis, and an example is shown in [5].

[Clause 89] **Under these extreme conditions**, helium burning becomes viable. [5]

Following the identification of Thematic elements, the process types were next identified, labelled and counted. The majority of the process types were relational. Relational clauses characterise and identify participants (Halliday and Matthiessen, 2004: 210), something which is necessary for elaborating technical taxonomies and making logical connections between nominalised processes. The modes of a relational clause are either "attributive" whereby the process attributes a characteristic or class membership to a participant; and "identifying", whereby the process connects an identity with a participant. There are further three types of relation: intensive, possessive and circumstantial. When combined, this gives six possibilities for relational processes as shown in Table 1, adapted from Halliday and Matthiessen (2004:216).

Table 1

Examples of Relational Processes

	Attributive "a is an attribute of x"	Identifying "a is the identity of x"
Intensive "x is a"	Daniel is wonderful.	Daniel is the leader; the leader is Daniel.
Possessive "x has a"	Daniel has a guitar.	The guitar is Daniel's; Daniel's is the guitar.
Circumstantial "x is at a"	The ceremony is on Wednesday.	Yesterday was the 18th; The 18th was yesterday.

Great care was needed to identify which type of relational process a clause construed since, as Halliday (2004) noted, the metafunction of a process can be difficult to determine. Take example [6].

[Clause 5] **Through minimising the surface energy**, the molecules at the surface may assume a preferred orientation. [6]

The process in example [6] is labelled "relational-circumstantial-attributive". The word "assume" here is a case where a circumstance is acting as a process. In other words, "assume" can be reconstrued as "be in", and so the state of a preferred orientation is attributed to the molecules at the surface. The six relational processes, along with other less frequent material, verbal, mental and existential processes were labelled and counted.

Finally, the complexity of the Thematic Head was interrogated. Given the long nature of nominal elements in technical scientific texts, it was decided to make the measure of complexity a little more fine grained than that used in Lavid *et al.* (2012). Here the complexity is defined as "simple", "complex" and "very complex". Simple refers to a single nominal element or a collocation. Collocation here refers to recurring word groups, usually noun groups, which reference a single concept or object. Examples include "diffusion constant", "surface analytical tools", "nuclear equation", "atomic numbers". Complex refers to a single nominal element or a collocation which has a single modification. Examples include "the mass number of the nuclide", "the collapse of the star's core". Any further modification was labelled as very complex, and examples include "the high abundance of iron and nickel in the universe" and "low concentrations of Li and B". Simple, complex and very complex Thematic Heads were then counted.

After all the counts were taken, Chi-square statistics were applied to examine any differences in distributions between the two genres.

Results

Distribution of Thematic Elements between the Two Genres

Table 2

Observed Counts of Thematic Elements in the Introductions of the Research Article and the Textbooks

Genre	Thematic Head	Circumstantial PreHead	Textual
Research Article	260	38	98
Textbook	220	26	87

Table 2 shows the observed values of Thematic elements in the introductions of the research article and the textbook. Calculating Chi-squared reveals that p(χ^2 > 0.7989 *df* = 2) = 0.67. This result is not significant at the p < 0.05 level.

Distribution of Process Types between the Two Genres

Table 3

Observed Counts of Process Types in the Introductions of the Research Article and the Textbook.

Genre	Material	Other	Rel-Int-Attr	Rel-Int-Identity	Rel-Cir-Attr	Rel-Cir-Identity	Rel-Poss-Attr	Rel-Poss-Iden
Research Article	52	18	42	22	14	24	15	5
Textbook	24	15	30	61	7	27	18	8

Table 3 shows the observed values of process types in the introductions of research articles and textbooks. Calculating Chi-squared reveals that p(χ^2 > 33.84, *df* = 7) = 0.000018. This result is significant at the p < 0.05 level.

Distribution of Complexity in Thematic Head between the Two Genres

Table 4

Observed Counts of Complexity in the Introductions of the Research Article and the Textbook

Genre	Simple	Complex	Very Complex
Research Article	92	59	37
Textbook	112	44	32

Table 4 shows the observed values of complexity in Thematic Heads in the introductions of research articles and textbooks. Calculating Chi-squared reveals that p($\chi^2 > 4.5076$, $df = 2$) = 0.105. This result is not significant at the
$p < 0.05$ level.

Discussion

Discussion of Findings

Comparing the distribution of Thematic elements, process types and complexity of Thematic Head can reveal how the particular linguistic features serve the contextual purpose of the text. The first finding is that there is no significant difference between the distribution of Thematic elements between introductions in research articles and textbooks. This implies that within both communities of practice, background information is presented in similar ways. A cursory analysis of the Circumstantial PreHeads shows that they play a role in setting up a context for the development of a Thematic Head [7], [8] in the proceeding clauses.

[Textbook, clause 63] **In a balanced nuclear equation**, the sum of the mass numbers of the reactions is equal to the sum of the mass numbers of the products (12 + 4 = 16). [7]

[Research, clause 133] **In this setup**, the liquid was forced through a 10-20 μm slit by use of low-soluble helium to apply a backing pressure. [8]

A further role for circumstantial PreHeads is the unfolding of a story. In the research article, the story offers background information about previous experimental techniques related to the current research. In the textbook introduction, the authors explain how the elements were first created in the universe. This use of circumstantial PreHeads to set contexts for Thematic Head development and giving background information is something which can inform the annotation of a corpus.

Regarding Textual themes, the use of conditional, additive, appositive and causal conjunctive adjuncts is common to both texts and shows how the writers in both texts are involved in three main discursive processes. These are i) creating logical connections between abstract ideas, ii) demonstrating cause and effect between nominalised processes, iii) elucidating concepts through examples. Again, this is another way in which the texts can be annotated for retrieval by an end user.

The significant difference found between the texts in the distribution of process types is interesting as it shows that the way the writers construe the experience of their respective contexts is fundamentally different, in spite of the similarity between the texts' thematic distribution. In a textbook, the writer is more likely to be concerned with showing equivalence, exemplifying, symbolising, equating, defining and demonstrating for the purposes of educating the reader on basic concepts, equations and meanings in diagrams. A writer for a research article might be less concerned with this since the background knowledge is already assumed. This contrast in purpose is manifested through the high count of relational - intensive - identifying processes in the textbook.

In contrast, intensive and circumstantial attributions were found to be more common in the research article than in the textbook. This is probably due to the fact that the research article is concerned with attributing particular qualities to experimental setups [9] and relating outcomes as results of particular experimental inputs in a cause-effect relationship. This concern with experimental background is not present in the textbook, and the use of these process types in textbooks seems mostly concerned with attributing qualities to particular Thematic elements [10].

[Research, clause 181] after evacuation of the sample chamber, the inside of the cell would quickly become saturated at the vapor pressure of the solvent, glycerol. [9]

[Textbook, clause 59] Nuclear reactions are very much more energetic than normal chemical reactions [10]

Again, the relatively high count for material processes in the research article compared with the textbook can be accounted for by the research writer's concern with describing background experimental information.

Finally, the distribution in the complexity of Thematic Heads was found not to be different between the two texts. This is not surprising, since scientific texts will be concerned with a core experiential domain which can be represented through particular collocations. Furthermore, as the text unfolds and logical connections are made and processes are nominalised, Thematic Heads will fluctuate in complexity. Halliday (2004) theorises that grammatical metaphor is a common feature of all scientific English, which explains why the complexity of Thematic Heads would be similar between any scientific text.

Implications and Conclusions

This small study has shown that there are similarities and differences between introductions from scientific research articles and introductions from science textbooks. The similarities can be attributed to the similarity between the two communities of practice - research and education, both members of the wider academic community. Academic language in general uses Circumstantial PreHeads and Textual themes to set context and structure logical connections between ideas. However, the difference between the two texts highlights the different concerns of the two communities of practice. In research, the concern is with situating a paper within the work of the wider research community, and this is construed through relational-intensive-attributive type processes. In education, on the other hand, the concern is with informing the student about basic concepts and definitions, and this is construed through relational-intensive-identifying type processes.

This finding, though not surprising, has implications for how texts are sampled and annotated for the corpus. As Sinclair (2005) has recommended, representing the community in the corpus is a must. For graduate students writing dissertations, the ability to define key concepts in a similar style to a textbook is a necessary skill which justifies using graduate and undergraduate textbooks within the corpus. However, in order to delineate the lexicogrammatical features of textbooks and research articles, it seems necessary to annotate the corpus in terms of text genre and the purpose of each stage within the genre. This will allow the corpus access interface to be designed to meet the needs of the intended users of the corpus. Such annotation will allow users to define their searches with greater accuracy, and generate concordance results and collocations information that are relevant to their needs.

This paper has demonstrated how a comparative analysis can shed light on the language resources used to achieve a purpose in a text. However this study is limited in a number of ways. First, the sample used was very small and restricted only to one stage in the two respective genres. Including more stages in the genres, and widening the genres to include fields such as biology and physics would further shed light on the differences and similarities. Second, the study was decidedly uni-dimensional. Biber and Conrad (2001) have pointed out that analysing single linguistic features does not shed light on the systematic variation of clusters of features. This study could be extended to include a multi-dimensional analysis of a wide range of linguistic features over a larger corpus. A multi-dimensional analysis could also reveal fallacies in the assumptions made about what counts as a particular stage in the genres of textbooks and research articles. Finally, the classification of the linguistic features in this study were based on the researcher's own understanding. Collaboration with other researchers would reduce bias and error and produce a more accurate and valid sets of results. In spite of these limitations, this paper demonstrates that careful annotation of samples from specialised genres can ensure that the assumptions of the corpus builder are fully documented. These assumptions include what samples we believe to be sufficiently representative of the target genres.

References

Biber, D. (2006). *University Language. A corpus based study of spoken and written registers.* Amsterdam: John Benjamins Publishing Company.

Biber, D. & Conrad, S. (2001). Multi-dimensional analysis and the study of register variation. In S. Conrad & D. Biber (Eds.). *Variation in English; Multi-dimensional studies* (pp.3-12). London, England: Longman.

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

Bhatia, V.K. (2002). 'A generic view of academic discourse', in Flowerdew, J. (ed.) *Academic Discourse.* pp21-39. Harlow – London – New York: Longman.

Clancy, B. (2010). 'Building a corpus to represent a variety of language', in O'Keeffe, A. and McCarthy, M (eds.) *The Routledge handbook of corpus linguistics.* pp. 80-92. USA, Canada: Routledge.

Coffin, C. (2001). 'Approaches to written language – a TESOL perspective', in Burns, A. and Coffin, C. (eds.) *Analysing English in a global context.* pp. 93-122. USA and Canada: Routledge.

Halliday, M.A.K. (2004). 'On the grammar of scientific English', in Webster, J. (ed.) *The language of science.* London: Continuum.

Halliday, M.A.K. and Matthiessen, C. (2004). *An introduction to functional grammar* (3rd ed.). Great Britain: Hodder Education.

Lavid, J., Arus, J., and Moraton, L. (2012). Genre realized in theme: The case of news reports and commentaries. *Discours: Revue de linguistique, psycholinguistique et informatique.* Retrieved from http://discours.revues.org/8623.

Martin, J.R. (1997). 'Analysing genre: functional parameters', in Christie, F. and Martin, J.R. (eds.) *Genre and Institution.* London: Cassell.

McEnery, T., Hardy, A. (2012). *Corpus Linguistics.* Cambridge, New York: Cambridge University Press.

McEnery, T., Xiao, R. and Tono, Y. (2006). *Corpus Based Language Studies. An advanced resource book.* New York: Routledge.

O'Keeffe, A., McCarthy, M. and Carter, R. (2007). *From corpus to classroom. Language use and language teaching.* Cambridge, New York: Cambridge University Press.

Painter, C. (2001). 'Understanding genre and register: implications for language teaching', in Burns, A. and Coffin, C. (eds.) *Analysing English in a global context.* pp. 167-80. USA and Canada: Routledge.

Reppen, R. (2010). 'Building a corpus. Key considerations', in O'Keeffe, A. and McCarthy, M (eds.) *The Routledge handbook of corpus linguistics.* pp. 31-7. USA, Canada: Routledge.

Sinclair, J. (2005). 'Corpus and text-basic principles', in Wynne, M. (ed.) *Developing linguistic corpora: A guide to good practice.* pp. 1-16. Oxford: Oxbow Books.

Wenger, E. (1998). *Communities of practice. Learning, meaning and Identity.* Cambridge and New York: Cambridge University Press.

Secondary References

Atkins, P.W., Overton, T.L., Rourke, J.P., Weller, M.T., and Armstrong, F.A. (2010). *Inorganic Chemistry.* Oxford: Oxford University Press.

Andersson, G., and Ridings, C. (2014). Ion Scattering Studies of Molecular Structure at Liquid Surfaces with Applications in Industrial and Biological Systems. *Chemical Reviews, 114; 8361-8389.*