Using a dedicated corpus to identify features of professional English usage: What do "we" do in science journal articles?

Judy Noguchi, Thomas Orr and Yukio Tono

Mukogawa Women's University
University of Aizu
Meikai University

Abstract

This chapter addresses the problem of inadequate educational materials for the effective training of non-native speakers in the professional English of the scientific community. We claim that one cause of this inadequacy is lack of proper linguistic research based on suitable linguistic research tools. The development of a major international corpus of professional English in the sciences and other fields is described as a significant resource for solving this problem, illustrated with a specific research project examining the use of "we" in professional English from the Corpus of Professional English. Results reveal the value of well-designed dedicated corpora for addressing the specific English instructional needs of non-native speakers of English in the professions.

1. Introduction

Although English is the language of preference for science in international contexts, the majority of the world's scientists are not native speakers of English and, therefore, they can be said to be considerably disadvantaged in professional English communication. Language teachers may have introduced non-native speakers (NNS) to the sounds, symbols, and structure of English for general purposes such as asking for directions, reading a menu, and writing classroom essays. However, few NNS of English have ever enjoyed the luxury of specialised training in the spoken and written discourse of their profession. In general, this means that they have had to pick this English up on their own, at considerable effort, with disappointingly limited results (Coates et al. 2002). This situation is terribly unfortunate, for it not only severely marginalises NNS in their scientific communities, but it also prevents the world from benefiting from the creative potential of those who cannot disseminate their ideas persuasively because of poor English.

Basically, there are two reasons for this unfortunate situation. One is the scarcity of language teachers who possess expertise in the English of science, and the second is the scarcity of instructional materials that adequately explain professional English text in scientific contexts. The English of research articles and technical specifications is very different from the English of storybooks and street signs, and mastery of this English does not come easily. Native-speaker

intuition or professional work experience alone are not enough to generate proper linguistic insight, and training materials based on "research" of this kind tend to be disappointingly vague and fraught with inaccuracies. Teachers and students of English in the sciences require more substantial materials based on far more rigorous research, enabled by far better research tools. The development of computer-based corpora, along with sophisticated software tools to analyse them properly, has now started to make this kind of research possible.

2. The value of specialised corpora

The advent of computers, computer-based corpora, and the whole new field of corpus linguistics is now beginning to provide some satisfactory solutions to the problems mentioned above. By collecting discourse samples in the linguistic domains requiring study, corpus linguists can now begin to identify features of language that were beyond the scope of thorough observation in the past. This development is welcome news for teachers and students of scientific English, for scientific genres, with all of their peculiarities, can now be studied on grander scales to generate far more objective and reliable data. General corpora, such as the British National Corpus (BNC), provide the means for studying how English is used in general contexts; while the development of specialised corpora provide the means for studying how English is used in special contexts.

For rigorous investigations of English in the theoretical and applied sciences, a very special corpus is required, dedicated to the language of these professional communities. Gathering a large enough collection of suitable discourse to develop a dedicated corpus of this kind, however, has not been an easy task. The value of corpora of scientific English on a small scale has been suggested in the past (Johns, 1991; Bondi, 2001), one good example being a corpus of physics research articles and their parallel academic conference presentations developed by Umesaki (2000). Using this corpus, Umesaki explored the variety of referents to the writer in academic papers, which she identifies as "one of the difficulties for non-native speakers of English in writing academic papers" requiring greater educational attention based on findings from corpus research.

Although work of this nature is being conducted in various disciplinary fields on small scales, there remains a need for a major dedicated corpus of English for studying professional English in the theoretical and applied sciences. This would provide a referent corpus against which these smaller corpora could be compared and also contribute to a better understanding of how English is used by professionals internationally when they communicate with each other.

3. The construction of the Corpus of Professional English, CPE

The creation of a major international corpus of English in the sciences and other professions to aid research in professional discourse is a serious undertaking, and a new organisation has recently been established to take on this challenge. Called the Professional English Research Consortium (PERC), this non-profit academic organisation headquartered in Tokyo, was established in April 2002 to create such a corpus and generate research in professional English that might be of particular benefit to NNS in the professions as well as to the educators and material developers that support them.

The corpus, named simply the Corpus of Professional English (CPE), aims to become a 100-million word balanced corpus of Professional English. It is designed for various research purposes, ranging from pure research on Professional English (e.g. variations of usage in lexis, syntax, semantics, and discourse across text types) to applied research such as lexicography, language testing, educational materials and program development. The CPE will be serviced by a sophisticated web-based query system so that those who are not familiar with corpora can extract linguistic features they need in a user-friendly manner. The design scheme of the CPE is shown in Table 1.

The ultimate goal of the CPE is to achieve a 100-million word written corpus, composed from a reasonable balance of text types. The balance of each text type will be further examined in the future, but tentatively it was decided that the following ratio of text types seemed most suitable for representing professional English: academic journals (30%), legal/workplace documents (20%), trade journals (10%), reference books (10%), websites (10%), newsletters (5%), correspondence (5%), manuals (5%) and ephemera (5%). At present, the project team is focusing on the collection of academic journal articles, for this is the text that frequently proves most difficult when obtaining copyright permission, and yet is needed most by researchers in PE.

Table 1: Tentative balance of text types for the CPE

Academic journals	30%
Legal/workplace docs	20%
Trade journals	10%
Reference books	10%
Websites	10%
Newsletters	5%
Correspondence	5%
Manuals	5%
Ephemera	5%

In order to select the journal texts in an objective way, the project team decided to base content decisions on data obtained from Journal Citation Reports (JCR), which presents quantifiable statistical data for an objective and systematic

approach to determining the relative importance of journals within their subject categories. At present, the JCR contains 5,700 journals in the Science Edition and has a unique measure called "Impact Factor", which provides a way to evaluate or compare a journal's relative importance to others in the same field. Employing this data, the top 20% of the journals with the highest impact factor in each field were selected for inclusion in the CPE. JCR classifications were also used to define the subject fields.

Acquiring texts for a corpus is always difficult but not impossible. Following procedures established by the creators of the BNC, the CPE project team sent letters to major journal publishers in order to obtain copyright permission for more than 1,500 journals. As is often the case, the first letter did not yield a sufficient response; however, much more favourable responses were obtained after the launching of the PERC website which explained the CPE in greater detail than the letters. At present, copyright clearance has been obtained for almost 300 journals from over 50 publishers.

To facilitate research on the CPE, a web-based corpus query system with a 3-level interface has also been developed in collaboration with the major PERC member institution, Shogakukan, Inc. For elementary users, the tool provides simple word search and KWIC displays only. For intermediate users, independent word/POS/lemma queries can be made with detailed collocation statistics (raw frequency/T-score/MI-score/log-log). And for experienced researchers, complex search (word/POS/lemma and combinations of the three) is made available with more sophisticated output. Shogakukan plans to launch a portal site for mega-corpora such as the BNC, the ANC, and COBUILD-Direct in the future so that ordinary language teachers can access these large corpora without worrying about the installation of a complicated interface.

Currently, the CPE project team is in the process of acquiring the targeted texts as well as cleaning and formatting the existing texts that have already been obtained. A prototype version of the CPE (c.2 million words of copyright-cleared academic journal text) has been set up on this new query system and is beginning to provide an impressive wealth of data that contrasts significantly with that which can be obtained from general corpora such as the BNC.

4. The research project and its results

To demonstrate the value of a corpus specifically dedicated to professional English, we have chosen for this chapter a simple analysis of current usage of the pronoun *we* as it appears specifically in professional scientific texts in comparison with similar data gathered from corpora devoted to English on a broader, more general scale. We will first explain the methods we employed in this research, then follow this with specific results and a discussion of their significance.

4.1 Methodology

As stated above, the CPE is currently under construction. The research for this paper was conducted on a prototype corpus from the CPE of 260 journal article texts from 18 journals in fields ranging from multidisciplinary agriculture, biochemistry and molecular biology, cell biology, developmental biology, plant science and forestry to multidisciplinary materials science, mineralogy, oceanography, general and internal medicine, health care sciences and service, orthopaedics medicine, pharmacology and pharmacy, and psychiatry. The prototype corpus included 1,787,484 tokens of almost 60,000 types.

The corpus was examined for verbs used after *we* using Wordsmith Ver. 2 (Oxford University Press). The concordance lines were rearranged with the word to the right of *we* in alphabetical order, and then the verbs were classified according to the seven major semantic domains described in the *Longman grammar of spoken and written English* (Biber et al., 1999): "activity verbs, communication verbs, mental verbs, causative verbs, verbs of simple occurrence, verbs of existence or relationship, and aspectual verbs." A summary of these verb features and some examples are presented in Table 2.

Table 2: Verb semantic domains based on the *LGSWE* (Biber et al., 1999: 360-364)

Semantic Domain	Features	Examples
Activity	Actions and events associated with choice; subject is semantic role of agent	bring, buy, carry, come, give, go, leave, work
Communication	Subcategory of activity verbs; associated with activities for communication	ask, explain, say, suggest
Mental	Activities and states experienced by humans; including cognitive, emotional, and perception	think, know, love, hate, see, taste, read
Facilitation or causation	A new state of affairs is brought about by a person or an inanimate entity	allow, cause, enable, require, permit
Occurrence	Events happening without volitional activity	become, change, happen, develop
Existence or relationship	State of relationship between entities	be, seem, appear
Aspectual	State of progress of an event or activity	begin, continue, keep, stop

According to the *LGSWE* (Biber et al., 1999), the most frequently used verb type is the activity verb. The *LGSWE* bases its findings on corpus studies which identified the four registers of conversation, fiction, news, and academic prose. Examination of the distribution of commonly-used verbs according to the four registers revealed that activity verbs ranked the highest in three of the four registers. It was only in academic prose that existence verbs display almost

equivalent frequency. Another feature of academic prose overall is the use of more causative verbs and occurrence verbs, compared to that of other registers.

With respect to the usage of the personal pronoun, the *LGSWE* (Biber et al. 1999: 329-330) recognises its common use "to refer to a single author, a group of authors, to the author and the reader, or to people in general." Biber et al. also state that "In some cases, academic authors seem to become confused themselves, switching indiscriminately among the different uses of *we*." This highlights the problem pointed out by Umesaki (2002) faced by the NNS scientist who often finds author-referent conventions confusing.

In this work, we focused on identifying the type of verb following *we* in the prototype CPE corpus. Knowing what types of verbs are commonly used in academic papers should offer help to the NNS scientist in deciding when and how to use *we* when writing up research.

4.2 Research data from the CPE

A total of 3,401 instances of *we* followed by a verb were identified in the CPE prototype corpus. The main verb was identified and classified according to the semantic domain classifications in the *LGSWE*. Its tense and aspect were also noted. If the verb occurred at least twice, it was included in the count for the distribution of verb types presented in Tables 3a and 3b.

Table 3a: Distribution of verb types used with *we* according to semantic domain and tense/aspect

Verb type	Total	%	past	%	pres	%	prep	%
Men	1433	45.13	668	46.62	472	32.94	119	8.30
Act	1038	32.69	612	58.96	187	18.02	160	15.41
Com	440	13.86	71	16.14	290	65.91	45	10.23
Exi	168	4.66	63	31.76	70	44.59	6	4.05
Cau	40	1.89	23	65.00	3	11.67	11	18.33
Asp	32	1.01	8	25.00	29	90.63	6	18.75
Occ	24	0.76	14	58.33	3	12.50	5	20.83
	3175		1459	45.95	1054	33.20	352	11.09

Table 3b: Distribution of verb types used with *we* according to semantic domain and tense/aspect

Verb type	mod	%	prec	%	pstp	%
Men	122	8.51	11	0.77	16	1.12
Act	49	4.72	18	1.73	10	0.96
Com	33	7.50	3	0.68	0	0.00
Exi	20	13.51	5	3.38	0	0.00
Cau	3	5.00	0	0.00	0	0.00
Asp	7	21.88	2	6.25	0	0.00
Occ	1	4.17	0	0.00	1	4.17
	235	7.40	39	1.23	27	0.85

Verb semantic domains: men = mental, exi = existence, act = activity, com = communication, exi = existence, cau = causative, asp = aspective, occ = occurrence.

Verb tense and aspect: past = past tense, pres = present tense, prep = present tense perfect aspect, mod = modal auxiliary, prec = present tense continuous aspect, pstp = past tense perfect aspect

The present tense continuous aspect occurred in 5 instances, two with mental verbs and three with activity verbs, but these data are not included in the table.

As can be seen from Tables 3a and 3b, the most frequently used verb type after the personal pronoun *we* was the mental verb accounting for 45.13% of the total. This was followed by activity verbs at 32.69% and communicative verbs at 13.86%. The most frequently-used verb tense was the past tense, accounting for 45.95% of all instances catalogued. However, this tense was the predominant one only for the mental, activity, causative, and occurrence verbs. The present tense form was more commonly used for the communicative, existence and aspectual verbs.

A closer examination of the verbs in their semantic domain classifications reveals a more complex picture. As can be seen from Tables 5a and 5b, while the mental verbs *find, observe* and *examine* overwhelmingly occur in the past tense, *conclude* occurs more than 80% as the present tense form. In the case of activity verbs with the highest frequencies, *use, analyse* and *test* are predominantly used in the past form, but *show* is used in the past form in only 14.29% of the instances observed while it appears more frequently and almost equally as the present tense perfect aspect (41.27%) and the present tense (40.48%).

Table 4: Number of verbs in each semantic domain and some examples

Verb semantic domain	No. of verbs observed	Examples (No. of instances)
Mental	97	find (163), observe (108), examine (103), conclude (58), identify (54), know (47), compare (46) , see (45), determine (41), investigate (39)
Activity	110	use (183), show (126), analyse (51), test (48), demonstrate (46), perform (46), measure (33), calculate (24), obtain (23), do (21)
Communicative	34	thank (101), report (49), present (34), describe (30), propose (29), note (19), ask (18), suggest (18), acknowledge (14), discuss (11)
Existence	12	have (71), be (28), exclude (14), include (11), stand (5), have to (4)
Causative	5	be able to (28), be unable to (20), allow (6), require (3), subject (3)
Aspect	8	begin (10), continue (5), start (4), undertake (4), initiate (3), achieve (2), enter (2), keep (2)
Occurrence	6	develop (12), fail (4), modify (2), increase (2), change (2), become (2)

Verbs which appeared after *we* two or more times were counted.

Table 5a: Verb tense and aspect distribution for most frequently used mental and activity verbs

Verb	Total	past	%	pres	%	prep	%
Mental							
find	163	129	79.14	11	6.75	19	11.66
observe	108	83	76.85	16	14.81	7	6.48
examine	103	78	75.73	3	2.91	17	16.50
conclude	58	6	10.34	47	81.03		0.00
Activity							
use	183	131	71.58	24	13.11	17	9.29
show	126	18	14.29	51	40.48	52	41.27
analyse	51	43	84.31	2	3.92	5	9.80
test	48	39	81.25	1	2.08	6	12.50

Table 5b: Verb tense and aspect distribution for the most frequently used mental and activity verbs

Verb	mod	%	prec	%	pstp	%
Mental						
find	3	1.84		0.00	1	0.61
observe		0.00		0.00	1	0.93
examine	3	2.91		0.00		0.00
conclude	5	8.62		0.00		0.00
Activity						
use	9	4.92	1	0.55	1	0.55
show	4	3.17		0.00	1	0.79
analyse	1	1.96		0.00		0.00
test	1	2.08		0.00		0.00

Not including one instance of the past perfect for *observe*

Verb tense and aspect: past = past tense, pres = present tense, prep = present tense perfect aspect, mod = modal auxiliary, prec = present tense continuous aspect, pstp = past tense perfect aspect

Table 6: Top twenty clusters in the vicinity of *we*

N	Cluster	Freq.
1	we found that	76
2	in this study	50
3	we conclude that	39
4	we examined the	39
5	we have shown	35
6	acknowledgments we thank	32
7	we used the	32
8	have shown that	28
9	in the present	28
10	we did not	28
11	we do not	25
12	found that the	23
13	we show that	22
14	we have previously	21
15	the effect of	20
16	the present study	19
17	in this paper	18
18	this study we	18
19	we have found	18
20	we used a	18

The clusters presented in Table 6 show that the top three clusters involve mental verbs, *found, conclude,* and *examined.* This suggests that cognitive activities of dynamic nature (*find* and *examine*) are expressed using the past tense, while the more stative *conclude* appears most frequently in the present tense. The sixth cluster is from the acknowledgements section and indicates an almost formulaic usage of *we thank.* The repeated references to the work being presented in the paper (*in this study, in the present, the present study, in this paper)* in the vicinity of *we* indicate that direct reference to the author(s) appears particularly when attention is being drawn to the study under discussion.

Interestingly, of the 260 texts examined, *we* was used at least once in 223 texts. The highest number per 1,000 instances was 13.94, with the total overall average being 2.12 per 1,000. The average for the top ten texts was 9.74 instances per 1,000 words.

Table 7: Texts with high frequency usage of *we*

No.	File	Words	Hits	per 1,000	Field
1	079.txt	1,865	26	13.94	Health services
2	078.txt	5,549	70	12.61	Health services
3	072.txt	2,956	34	11.5	Health services
4	070.txt	6,292	68	10.81	Health services
5	047.txt	4,821	50	10.37	Psychiatry
6	035.txt	7,748	73	9.42	Plant sciences
7	195.txt	4,668	38	8.14	Cell biology
8	076.txt	2,693	21	7.8	Health services
9	058.txt	7,696	50	6.5	Plant sciences
10	314.txt	7,921	50	6.31	Oceanography
			Ave	9.74	

Table 7 suggests that some journals or fields may display a higher frequency of first-person pronoun usage than others. Also revealing were the negative data of texts in which *we* was not used even once. Such texts occurred across all fields, but all sixteen texts from one journal in multidisciplinary materials science had no instances of *we* at all. The only instance that was detected was the abbreviation of *WE* for *working electrode.*

5. Present applications and future research

The above analyses on the usage of *we* in the 260 texts in the prototype CPE corpus of scientific academic journal articles from a range of fields reveal the following:

i) The use of *we* is rather common, occurring at least once in 85.77% of the texts examined.

ii) Some journals and fields tend to display more *we* usage that others. On the other hand, all texts coming from one of the journals had no instances of *we*. Thus, there seems to be a need for even further specialization of corpora to illuminate differences among journals and/or fields of study.
iii) The verb type most commonly used with *we* is the mental verb (45.13%) followed by the activity verb (32.69%).
iv) The most frequently used tense is the past tense, but the distribution of past and present tense usage is reversed for some verb types and even within verb type categories.
v) High-frequency clusters in the vicinity of *we* tend to be related to mental activities and references to the work under discussion.

The findings overall point to the need for even further refining of corpus studies of specialised texts in order to reveal features which can be used when planning course materials for English for specific purposes classes. Such comparative studies of texts from different research fields and genres must await the completion of the CPE corpus; however, this research conducted from a small sample of dedicated texts already reveals some interesting things that differ from earlier findings based on general English corpora.

Even in its present state, the CPE corpus prototype dedicated to text in scientific disciplines can prove very helpful as reference material for postgrad students or NNS scientists who are at the stage of writing up their research. If they are given background instruction in the genre-analysis approach to understanding the framework of moves and steps that compose the research journal article (Swales, 1990; Weissberg and Buker, 1990), a data-driven learning approach to concordancing (Johns, 1991a and b) can serve as a valuable tool for aiding NNS with their professional writing, when it comes to word choice or other writing issues (Hunston, 2002). As the CPE continues to develop, we envisage a system by which NNS scientists can access a website for online support when creating professional documents, which would include access to selections of dedicated corpora in the specific fields for which they are writing.

References

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

Bondi, M. (2001), 'Small corpora and language variation', in M. Ghadessy, A. Henry, and R. L. Roeberry (eds.), *Small corpus studies and ELT*. Amsterdam: John Benjamins.

Coates, R., B. Sturgeon, J. Bohannan, and E. Pasini (2002), 'Language and publication in *Cardiovascular research* articles', *Cardiovascular research*, 53: 279-285.

Hunston, S. (2002), *Corpora in applied linguistics.* Cambridge: Cambridge University Press.

Johns, T. (1991a), 'From printout to handout: Grammar and vocabulary teaching in the context of data-driven learning', in T. Johns and P. King (eds.), *Classroom concordancing.* Birmingham, UK: Centre for English Language Studies, The University of Birmingham, 27-46.

Johns, T. (1991b), 'Should you be persuaded: Two examples of data-driven learning', in T. Johns and P. King (eds.), *Classroom concordancing.* Birmingham, UK: Centre for English Language Studies, The University of Birmingham, 1-16.

Johns, T. (2002), homepage http://web.bham.ac.uk/johnstf/timeap3.htm

Swales, J. (1990) *Genre analysis: English in academic and research settings.* Cambridge, UK: Cambridge University Press.

Umesaki, A. (2000), 'Syntactic differences in the discourse of oral and written papers', *English corpus studies,* 7: 39-59.

Umesaki, A. (2002), 'Reference to the presenter in academic papers'. Paper presented at *AILA 2002*, Singapore International Convention and Exhibition Centre, Dec. 21, 2002.

Weissberg, R. and S. Buker (1990). *Writing up research: Experimental research report writing for students of English.* Englewood Cliffs, NJ: Prentice-Hall.