European Journal of Business and Social Sciences, Vol. 1, No. 9, pp 55-71, **December 2012.**
URL: *http://www.ejbss.com/recent.aspx*
ISSN: 2235 -767X

LEXICAL FEATURES IN CORPORATE ANNUAL REPORTS: A CORPUS-BASED STUDY

Huili Wang
(Corresponding Author)
Dalian University of Technology, China

Lixin Li
Dalian University of Technology, China

Jingxiang Cao
Dalian University of Technology, China

ABSTRACT

Annual reports (ARs) are the most important external documents and the most used channels for communication between organizations and stakeholders. The study uses self-built one self-built large corpus and nine small self-built sub-corpora to investigate lexical features in ARs to help AR users understand written financial discourses. The corpus consists of 120 pieces of 2011 annual reports. The results showed that financial aspect is mostly concerned in CARs; vocabulary of chairman's statements is the richest while that of auditor's report is not rich; words in auditor's reports and corporate social responsibility (CSR) reports are longer and more complex than those in chairman's statements and financial statements; fuzzy words are noticeably more frequent in the section of chairman's statements and business overview; the personal pronoun "our" is the most frequently used personal pronouns in ARs especially in chairmen's statement; more positive words are used than negative words in CARs.

Keywords: annual reports, keywords analysis, lexical features, corpus methodology

Introduction

Nowadays, corporate annual reports (CARs) are mostly used channels to disclose information to stakeholders and are treated as a promotional tool of the company. CARs involve many variables such as stock exchanges and company industries. Organizations listed on different stock exchanges are subject to different reporting requirements. Under the regulatory framework of the financial reporting, a public listed company (PLC) should publish reports annually to its shareholders with audited financial statements and certain narrative documents.

ARs as a hybrid text are made up by two kinds of data, which is quantitative data and qualitative data. Each kind of data is legally required to two kinds of disclosure: compulsory disclosure and voluntary disclosure. In ARs, the compulsory disclosed quantitative elements are financial statements and notes while compulsory disclosed qualitative elements are chairman's statement, management's discussion and analysis, corporate governance (CG) and independent registered auditor's report. Other elements such as mission statement, business overview, CSR and limited comparative information related to the company's industry and competitors are voluntary disclosure.

Generally speaking, PLCs' ARs are available online in the format of either pdf or html with appealing photography to attract readers' attention. Although ARs are created by insiders, they are used by audiences of both insiders and non-experts. The primary audiences for ARs are shareholders and potential shareholders while targeted audiences are employees, customers, suppliers and the community.

However, the wide variety of audiences has different purposes. Firstly, shareholders and potential shareholders use ARs to make economic decisions for the purpose of continuous operation. Secondly, employees treat ARs as a source for learning about a company's products and various operating activities. Besides, ARs are also helpful for employees to compare their wages with others in the same industry. Thirdly, ARs is a vehicle for promoting its image and describing initiatives to customers as well as suppliers. Fourthly, lending institutions require ARs to assess the risk of fixed-asset investment loans as well as creditworthiness of the business to decide to extend credit. Furthermore, ARs are used by Government for the purpose of tax inspections. Lastly, other stakeholders, for instance, local communities are also interested in ARs for a variety of reasons.

Thus, for and AR's writers and the above variety of ARs' readers, a better understanding of the language used in CARs is helpful to measure the company's performances. Facing this situation, the study attempts to explore lexical features in the whole ARs and in each document respectively from five perspectives such as lexical richness, word length, keywords in CARs, first personal pronouns, hedges and evaluative words.

Literature review

2.1 Corpus and corpus analysis

In modern linguistics, a corpus was defined by many studies. A corpus was defined by Sinclair (1991) as a sample of language text collected on explicit linguistic criteria. According to Leech (1992), a corpus was a representative sample of particular language or texts and was used for particular purposes. Richards and Platt (2000) pointed out that one of the purposes of using a corpus was to analyze linguistic features of a particular language. Moreover, McEnery, Xiao and Tono's studies (2005) viewed a corpus as a sample of machine-readable texts in the form of written or tape recordings. From the above definitions, a corpus is a rich source of data collected from language texts for linguistic study.

Currently, major electronic corpora are Brown Corpus (BC), Lancaster-Oslo/Bergen (LOB) Corpus, London-Lund Corpus (LLC) and the British National Corpus (BNC). BC was the first computer-readable corpus consisting of one million words of American English collected by Brown University around 1963 (Kennedy, 2000). Similarly to the BC, LOB Corpus was also compiled for linguistic study in 1960s with one million British English words. Furthermore, LLC is a corpus of one million running words of British English from the sources of spoken English (Kennedy, 2000).

The BNC was a 100 million-word corpus of current British English compiled by a joint work of academic institutions, publishers and British Library in 1990s. BNC contained 4,124 pieces of written and spoken texts including daily conversations. Compared with BC, BNC was drawn from a wide variety of sources without limitation to particular subject and genre (Kennedy, 2000). In this paper, BNC was adopted as a reference corpus.

Corpus linguistics is the study of language in corpora of natural text (Kennedy, 2000). The history of corpus linguistics dated back to 1970s when Kucera and Francis (1967) published Computation Analysis of Present-Day American English in which computational analysis was used to analyze the BC from perspectives of linguistics, statistics, psychology and sociology (Kennedy, 2000).

Currently, corpus-based approach has been widely applied to many fields such as stylistic analysis, language teaching and discourse analysis (Kennedy, 2000). In Biber, Conrad and Reppen's study (2000), features of corpus-based studies were summarized as an empirical approach, on the basis of a sample of natural texts, computer assisted analysis and combination of quantitative and qualitative analysis. In the study of Kennedy (2000), researches on corpus were mainly based on four aspects as lexical level, syntactic level, discourse analysis and genre analysis. This paper focused the corpus-based study on lexical level.

2.2 Studies on lexical features

Vocabulary is of central importance in second language proficiency. According to David Wilkins (1972), little information can be conveyed without grammar, but nothing can be conveyed without vocabulary. Moreover, based on the study of Enkvist (1969), the study of lexical characteristics was one of the aspects to analyze linguistic features of a text.

Lexical richness is a measure of the diversity of the vocabulary to reveal the difficulty of a given text. According to Biber (2009), news had the highest lexical richness while conversation had the lowest lexical richness. The type-token ratio (TTR) also known as lexical density was used to measure lexical richness of a text through the calculation of the number of types divided by the number of tokens (Nation, 1990). A high TTR index indicated that a text was enriched with vocabulary variety while a low TTR showed that a text used the same words repeatedly. However, TTR was influenced by the text length as the longer the text was, the lower the number of new word types were (Sichel, 1986). For this reason, standard TTR was adopted to measure lexical richness (Scott, 2004).

Furthermore, word length was also an important indicator to reflect the stylistic features of a text because the longer average word length, the more difficult the text was (Butler, 1985). Mendenhall (1887) studied the frequency of words within a limited length to identify the authorship. Kelih, Antic, Grzybek and Stadloker (2005) studied the impact of word length and concluded that the scientific way to measure word length was based on the number of syllables rather than the number of letters. Besides, word length was used in stylistic analysis. Deng (2007) used word length to study stylistic features of straight news. Huang (2010) studied lexical features of commercial English from perspectives of word length, lexical density and frequency distribution.

Besides studies on commercial English, studies of lexical features applied to various subjects such as Legal English, Journal English, Forestry English, Auditing English and English for civil engineering. Hao (2006) researched on lexical characteristics of legal English from ten areas such as technical terms, archaisms, big lengthy words, adoption of pronouns and vague words and concluded that the language of law meet the criteria of formality, accuracy and precision.

Tian and Liang (2011) researched on stylistic analysis of auditing reports in a corpus-based approach. Tian and Liang compared the self-built Auditor's Reports Corpus with BC and LOB by WordSmith and analyzed stylistic features from three perspectives of layouts, lexical features and grammatical features. This paper was inspired by Tian and Liang's study in which lexical features was discussed from aspects of lexical richness, word length and frequency.

2.3 Studies on annual reports

Annual reports (ARs) are one of the most important external documents prepared by organizations as a traditional vehicle for communication between organizations and stakeholders (Courtis, 1998). ARs assess organizations' operations and performance during the past financial year and discuss the management's view of the upcoming year (Kloptchenko, Back, Vanharanta, Eklund, Karlsson and Visa, 2002). Besides, ARs are also used as a tool for organizations to promote their values, objectives, strategies and other information which convey a positive image to the public as well as to increase employee's morale, improve relations with the community (Kendall, 1993).

Over the years, many studies have been conducted on ARs. Hildebrandt and Snyder (1981) studied on the chairman's statements in ARs aiming to find what impacts of corporate performance will have on language used in chairman statements. Results shows that languages are predominantly positive no matter what financial status an organization has, which was known as Pollyanna Hypothesis. Additionally, the results were confirmed by Thomas (1997). Thomas analyzed chairman's statements in the same company's ARs in five years and found that more words were used to present positive information.

Moreover, Judd and Tims (1991) pointed out that chairman's statements as the second most-read part after "financial highlights" in readership contributes to build customer loyalty. Jacobson (1998) researched on evaluative words in chairman's statements and concluded that more positive words were used instead of negative words in describing bad news which influenced readers' judgments. Donatella Malavasi (2005) concentrated on lexical meaning in expressing evaluation and suggested that ARs should be researched as a promotional genre instead of simply financial documents.

Furthermore, Bowman (1994), Rogers and Grant's studies (1997) and Krippendorff (2004) turned attention to the contents of ARs while Cambell (2002) focused on communicative efficiency in ARs. Rogers and Grant demonstrated that apart from financial statements in ARs, the language used in narrative documents can influence financial analysts as well.

Swales (2004) and Hyland (2005) studied chairman's statements in ARs and concentrate on the text and context. Goodman (2000) concluded that CARs were used as a communicative tool by organizations to gain competitive advantages, persuade and motive potential investors.

Yuthas, Rogers and Dillard (2002) pointed out that organization revealed operational and financial information strategically in CARs. Santema and Rijt (2001) found that the format of ARs did not change over time while the quantitative data vary. Based on the quantitative and qualitative data in ARs, Back, Toivonen, Vanharanta, and Visa (2001) studied the difference between each type's disclosures.

Through investigating quarterly reports of companies in telecommunications sector, Kloptchenko et al. (2002) found that both past performance and future perspectives were discussed in reports and the written style is influenced by organization’s performance. Based on the same sample of Kloptchenko et al. (2002), Magnusson, Arppe, Eklund, Back, Vanharanta, and Visa (2005) confirmed the results in different methodology.

The above studies on ARs demonstrated that few studies were conducted on lexical features of ARs. Furthermore, the study of Tian and Liang (2011) focused on one particular document of ARs. This paper took further study of lexical features in all documents of ARs based on Tian and Liang’s study.

Methodology

The present study is based on one large corpus and nine small sub-corpora related to CARs designing to find out lexical features in the whole CARs and that in each section of CARs. In this way, the corpus-based study focuses on frequency and distribution of the words in CARs, both in the whole reports and across the sections.

3.1 Data collection

The corpus consists of 120 PLCs’ 2011 annual reports with 8,102,762 running words. The companies come from 28 industries ranking within 2011 Fortune Global 500. Since PLCs are required to disclose annual reports publicly, all the reports of 2011 Fortune Global 500 companies can be downloaded from the official websites. However, only 120 annual reports were available among the 2011 Fortune Global 500 companies in the process of data collection. Accordingly, 120 pieces of 2011 annual reports were downloaded in the format of pdf and html with each ranging from 23 pages to 476 pages.

3.2 The corpora

In building the corpus, the first step is to convert ARs from pdf and html format to txt format. Next, based on different functions, move analysis was conducted through discussion between the author and professional accountants from the Association of Chartered Certified Accountants (ACCA) (Mohammad, 2011). After the thorough discussion, nine sub-corpora were split as follows:

Table 1

Contents and Tokens of Sub-corpora

Sub-corpus	Section of CARs	No of Tokens
1	Company Description	46,487
2	Chairman’s Statements	203,802
3	Business Overview	533,796
4	Management’s Discussion and Analysis	2,618,353
5	CSR	88,010
6	CG	913,991
7	Auditor’s Report	68,655
8	Financial Statements and Notes	3,409,620
9	Investor Information	154,645
Total Corpus		8,037,359

Among the above sub-corpora, corpus 1 to 7 is in narrative format while corpus 8 mainly contains accounting data and notes providing contextualizing information that makes accounting entries easier to understand. Corpus 9 includes supplementing corporate and stock information.

3.3 The analytical approach

After the corpus was split into nine sub-corpora, all the corpora were carefully read for the sake of getting a "feel" for the data in order to discover linguistic features of the discourse. Then, the data were processed with the wordlist of Wordsmith Tools 4.0 (Scott, 2004) to obtain frequency data. The wordlists obtained from sub-corpora are employed to compare lexical richness and word length between each sub-corpus.

Furthermore, keyword lists are generated based on the obtained wordlists. In this study, two kinds of keyword analysis were conducted. The first keyword analysis adopts BNC as a reference corpus for a comparison, which compares the whole CARs corpus with BNC aiming to find out general characteristics in CARs. Besides, the second keyword analysis compares each sub-corpus of CARs with the whole CARs corpus which acted as the reference file to compare linguistic features in each section of CARs.

Results and Discussion

The study discusses lexical features in CARs from five perspectives such as lexical richness, word length, keywords in CARs, first personal pronouns, hedges and evaluative words.

4.1 Lexical richness

The Types/token ratio (TTR) measures how rich the vocabulary in a given text is. In Wordsmith 4.0, TTR is achieved through the calculation of the number of types divided by the number of tokens. However, it is difficult to compare the TTR of smaller texts against larger texts since the bigger the texts the lower number of new word types. For this reason, the study adopts standard type/token ratio to compare lexical richness among nine sub-corpora. The standard type/token ratio is calculated by Wordsmith 4.0 based on every 1000 words and the results are shown in Table 2.

Table 2

Standard Type/Token Ratio of Nine Sub-corpora

Sub-corpus	1	2	3	4	5	6	7	8	9
Standard Type/Token Ratio	39.56	42.49	39.01	34.31	42.30	33.45	27.42	30.61	32.99

Obviously, vocabulary of chairman's statements (sub-corpus 2) is the richest in CARs. Chairman's statements, also called letter to shareholders, provide an overview of company's past performance and future perspectives taking the form of letter signed by the president, chairman or CEO. This section is believed to be the second most-read part after "financial highlights" in readership. Usually, changes in board of directors and dividend distribution are announced in this section. Thus, various kinds of vocabulary are adopted to express central information and attract reader's attention.

In comparison, the report of CSR in sub-corpus 5 uses extensive vocabulary ranking second in the table of lexical richness. Although CSR report is voluntary disclosure in CARs, many companies choose to disclose this information to encourage a positive impact on the environment and stakeholders. In this sense, various types of words are used in this section.

Furthermore, words in sub-corpus 1 and 3 are relatively richer. Sub-corpus 1 is the company description with mission statement and corporate profile to give a general picture of what the company does with words being hardly repeated. Additionally, sub-corpus 3 contains performance highlights and chief executive's review. Financial highlights usually take up the front document and are the mostly read section of an annual report. It summarizes high financial points such as one to three year's sales and earnings to give readers a quick review. Chief executive's review is usually in the format of narrative feature article. This part conveys a wide range of topics related to the organization's operations, such as new products, employee relations, sales trends, public relations, tax and legislation concerns. Thus, the rich contents in this part result in a relatively high standard type/token ratio.

In contrast, auditor's report (sub-corpus 7) is relatively not enriched with vocabulary variety among all sections of CARs. The result confirmed Tian and Liang's study (2011) that auditor's report is not enriched with vocabulary. In the study of Tian and Liang (2011), standard type/token ration of Auditor's Report Corpus is 33.93 which are lower than that of LOB and BROWN. Because auditor's report following the financial statement reports whether the financial statements have been complied with corresponding accounting requirements, it employs a lot of technical terms in accounting and auditing. For this reason, words in different reports from independent auditors are barely the same leading to the lowest TTR.

4.2 Word length

Table 3

Average Word Length of Nine Sub-corpora

Sub-corpus	1	2	3	4	5	6	7	8	9
Average Word Length	5.28	4.99	5.25	5.05	5.31	5.13	5.51	4.89	5.05

As shown in Table 3, words in auditor's reports (sub-corpus 7) and CSR Reports (sub-corpus 5) are more complex than those in chairman's statements (sub-corpus 2) and financial statements (sub-corpus 8). The result confirmed Tian and Liang's study (2011) that auditor's reports use longer words with an average of 5.47 than LOB and BROWN with 4.03 and 4.29 respectively.

Firstly, auditor's reports have the highest average word length followed by CSR reports, reflecting the highest degree of formality in these two types of reports. According to Tian and Liang (2011), longer words are used in auditor's reports to attract readers' attention. Secondly, the result shows that company description in CARs involves complex vocabulary. In this section, a mission statement is commonly used to briefly describe the purpose of an organization. Furthermore, a mission statement provides the framework within which organization's strategies are formulated and it guides the actions of the organization and the overall goal. Therefore, complex words are used to attract the readers' attention and conclude corporate profile concisely.

Thirdly, vocabulary used in CG (sub-corpus 6) is of moderate difficulty. This section gives a profile of company's management and its board of directors, and discloses CG report as well as director's remuneration report. Information such as the number of directors, directors' positions and experience are provided for readers to evaluate the quality of company's management.

Fourthly, words used in investor information (sub-corpus 9) and management's discussion and analysis (sub-corpus 4) are relatively easier in CARs with the same average word length of 5.05. Investor information at the end of CARs provides information concerning contact address and phone number, web sites and stock ownership and other additional information of the company. Management's discussion and analysis can be written at all different levels of comprehension to give a detailed review of what has happened with the company over the past three years from the perspectives of operational review, financial review, risk review and capital review.

Lastly, the result indicates that financial statements (sub-corpus 8) use the simplest words with the average of 4.89. This may be largely due to the restricted formats and accounting terms in financial statements which typically consisting of four basic statements such as statement of financial position, statement of comprehensive income, statement of changes in equity and statement of cash flows. They usually appear at the latter parts of CARs followed by notes which supply information such as accounting policies, dividends paid, and other calculations. Besides, in this section, most companies provide N year summary of financial data, generally 5, 6, 10, or 11- year summary.

4.3 Keywords in CARs

Keywords are those whose frequency is unusually in a particular text compared to a larger corpus (Scott, 1997). Keyword analysis is a useful way to characterize a text. In the study, the keyword list is generated by comparing the frequency wordlist of whole CARs corpus with the frequency of the same word in the wordlist of British National Corpus (BNC). The following table reviews the results of the first 200 keywords run by WordSmith 4.0.

Table 4

First 200 Keywords in CARs

1	FINANCIAL
2	ASSETS
3	MILLION
4	VALUE
5	NET
6	INCOME
7	COMPANY
8	RISK
9	DECEMBER
10	OUR
11	LIABILITIES
12	FAIR
13	SECURITIES
14	GROUP
15	CONSOLIDATED
16	CASH
17	TOTAL
18	EQUITY
19	CREDIT
20	LOANS
21	BILLION
22	STATEMENTS
23	MANAGEMENT
24	BUSINESS
25	TAX
26	INTEREST
27	LOSSES
28	OPERATING
29	ASSET
30	SHARES
31	IMPAIRMENT
32	INVESTMENT
33	RELATED
34	OPERATIONS
35	CAPITAL
36	DIRECTORS
37	ANNUAL
38	INVESTMENTS
39	DEBT
40	RATE
41	LOSS
42	SHARE
43	SALES
44	CORPORATE
45	ACCOUNTING
46	BANK
47	DEFERRED
48	INSURANCE

49	CONTRACTS
50	MARKET
51	OTHER
52	INSTRUMENTS
53	SUBSIDIARIES
54	DERIVATIVE
55	DUE
56	NOTES
57	YEAR
58	BOARD
59	AMOUNTS
60	EARNINGS
61	PERFORMANCE
62	SEGMENT
63	EXPENSES
64	CORPORATION
65	DERIVATIVES
66	REPORTING
67	AMOUNT
68	BASED
69	STOCK
70	EXPENSE
71	PRODUCTS
72	PRIMARILY
73	TERM
74	CURRENCY
75	TRANSACTIONS
76	COSTS
77	BALANCE
78	RECEIVABLES
79	PROFIT
80	MILLIONS
81	CUSTOMERS
82	SHAREHOLDERS
83	FISCAL
84	GAINS
85	NOTE
86	EXECUTIVE
87	CONTINUED
88	IMPACT
89	REPORT
90	INCLUDING
91	PORTFOLIO
92	EXCHANGE
93	REVENUE
94	REMUNERATION
95	PLAN
96	COST
97	PENSION
98	LIQUIDITY
99	GOODWILL
100	LOAN
101	OBLIGATIONS
102	BANKING
103	RISKS
104	FLOWS
105	CHANGES
106	AUDIT
107	GLOBAL
108	REGULATORY
109	RATES
110	INCREASE
111	ENDED
112	SERVICES
113	APPROXIMATELY
114	PERCENT
115	ACQUISITION
116	COMMITTEE
117	GOVERNANCE
118	CERTAIN
119	MARKETS
120	BUSINESSES
121	LIABILITY
122	CURRENT
123	AND
124	INCREASED
125	ISSUED
126	INTERNAL
127	OPTIONS
128	LIMITED
129	FOREIGN
130	TRADING
131	ASSUMPTIONS
132	GROWTH
133	BASIS
134	ATTRIBUTABLE
135	BENEFIT
136	INCLUDED
137	OFFSET
138	PLANS
139	BENEFITS
140	VALUATION
141	RECOGNIZED
142	ENTITIES

143	SALE
144	SIGNIFICANT
145	COLLATERAL
146	PERIOD
147	HEDGE
148	REVENUES
149	EMPLOYEES
150	FUTURE
151	DECREASE
152	RESULTS
153	INTANGIBLE
154	EXPOSURE
155	RESPECTIVELY
156	MORTGAGE
157	AVERAGE
158	COMPENSATION
159	CUSTOMER
160	WEIGHTED
161	PAYMENTS
162	OUTSTANDING
163	REPURCHASE
164	HEDGING
165	ACCORDANCE
166	FIXED
167	DIVIDENDS
168	RECOGNISED
169	HSBC
170	GROSS
171	ADJUSTMENTS
172	ACTIVITIES
173	BORROWINGS
174	COMPARED
175	DEPOSITS
176	PRIOR
177	AGREEMENTS
178	SWAPS
179	ESTIMATED
180	EXPECTED
181	AMORTIZATION
182	CARRYING
183	MATURITY
184	DIVIDEND
185	ADJUSTED
186	RECORDED
187	FUNDS
188	DATE
189	SHEET
190	BONDS
191	HEDGES
192	REQUIREMENTS
193	FINANCING
194	UNDER
195	EMPLOYEE
196	AWARDS
197	ESTIMATES
198	BALANCES
199	UNREALIZED
200	COMMITMENTS

The above keywords have key relation in CARs and take up key position in the reader's minds. In this case, these keywords will be discussed from five perspectives to give a general understanding of what the CAR is.

Firstly, the result shows that the financial aspect is mostly concerned in CARs. In finance description such words as "Financial" , "Assets/Asset", "Value", "Income", "Liabilities", "Cash", "Total", "Equity", "Credit", "Loans", "Tax", "Losses/Loss", "Interest", "Impairment", "Capital", "Debt", "Rate" ,"Sales" , "Expense/ Expenses" , "Derivatives", "Amount", "Reporting", "Costs/Cost", "Receivables", "Profit", "Fiscal", "Gains" "Revenue" "Pension", "Liquidity", "Goodwill", "Loan" and "Obligations" are most frequently used indicating that CARs are external financial reports with necessary financial data to disclose and forecast organizations' performance. Besides, the writers of CARs are based on facts with true and fair view in preparing the reports.

Secondly, the frequency of "December", "Annual", and "Year" is particularly high in expressing time. This is because CARs are prepared annually for the year ended on the 31st December. Thirdly, the results also indicate that such words as "Million", "Millions" and "Billion" are the common monetary terms in CARs. Fourthly, from the view of the parties, CARs mainly include "Company", "Corporate/Corporation", "Group", "Consolidated", "Directors", "Bank", "Subsidiaries", "Customers" and "Shareholders". However, the high frequency of the word "Bank" may be due to unbalanced distribution of CARs among industries since 21 pieces of ARs in the corpus come from companies in the bank industry.

Lastly, from the perspective of operation-related and management category, operational and risks are also highly concerned in CARs. In describing corporate operational performance and management, the more frequently used words are as follows: "Risk", "Fair" , "Management", "Business", "Operating/Operations", "Investment/Investments", "Market", "Performance", "Reporting/Report", "Products", "Transactions", "Executive", "Continued" , "Remuneration", "Plan".

4.4 First personal pronouns

In the keyword list of CARs in Table 4, the first personal pronoun "our" tanks the10th while other first personal pronouns such as "we", "I" and "my" are out of the keyword list. To explore the reasons and the distributions of first personal pronouns, the author compared each sub-corpus with the whole corpus as a reference file to obtain the keyword list of each sub-corpus.

Results show that the first personal pronoun "our" ranks first in the sub-corpus 3 and 4 and second in sub-corpus 2 and 5, but takes up 29th position in keywords in the sub-corpus 7. Furthermore, although the word "we" is not a first 400 keyword in the whole corpus, it is the first keyword in sub-corpus 2, the second keyword in sub-corpus 3 and the third keyword in sub-corpus 1. However, "I" and "my" only appear in the keyword list of sub-corpus 2, with each ranking the 5th and the 9th position. Besides, all of the identified four first personal pronouns are not keywords in the first 400 keyword lists of sub-corpus 6, 8, and 9.

The frequent occurrence of the word "our" in the section of business overview and management's discussion and analysis is a helpful way to shorten the distance between the organizations and readers as well as to get more approvals by readers in the given financial and performance analysis. Moreover, the frequency of "our", "we", "I", "my" is particularly high in the section of chairmen's statement, because using first personal pronouns in the letter crosses the blurry line among owner, management and employees, which contributes to create affinity, build corporate image and resonate with readers.

In contrast, sections of CG, financial statements and investor information provide objective information instead of financial analysis and performance judgments. Therefore, first personal pronouns are hardly used in these parts.

4.5 Fuzzy words

Fuzzy words refer to the words that make things fuzzier or less fuzzy such as sort of, some, about, more or less, believe, approximately. Since Lakoff published “Hedges：A Study in Meaning Criteria and the Logic of Fuzzy Concepts”(1972), hedges became a linguistic concept. According to Channell (2000), the use of vague words not only contributes to give the right amount of information but also is a way of self-protection. On the contrary, vague words are used when lacking specific information or deliberately to withhold information Channell (2000).

Since ARs emphasis on accuracy rather than precision, words used in ARs can be fuzzy to ensure accurate information is conveyed. In order to find out lexical features in describing different information in ARs, the distribution of hedges is studied. In analyzing the first 400 keywords in nine sub-corpora, the use of hedges are shown in Table 5.

Table 5

Hedges from First 400 Keywords in Sub-corpora

Sub-corpus	Hedging
1	MANY
2	MANY, BELIEVE, ABOUT, FEW, SOME, THINK, WILL, CAN
3	MANY, SOME, BELIEVE,SEVERAL, COULD, MAY, WILL, CAN
4	APPROXIMATELY, POTENTIAL,RELATIVELY, COULD
5	-
6	SHALL, MUST
7	MAY, BELIEVE
8	-
9	SHOULD, MAY

Hedges are noticeably more frequent in sub-corpus 2 chairman’s statements and sub-corpus 3 business overview than in any other sub-corpus. The main reason is that these sections contain forward looking statements. In describing organization’s forward performance, using such words as “believe”, “think”, “could” can make expressions more accurate and reliable.

Furthermore, quantifier “many” is the most frequently used word in the sections of corporate profile, letters to shareholder and business overview to describe financial conditions, which reflects the strategy of self-protection and preserves the information that companies are not willing to disclose publicly.

In management’s discussion and analysis, such words as “approximately”, “potential”, “relatively” and “could” are frequently used to forecast financial and operational performance. Such fussy words are used to avoid risks of predicting organization’s performance inaccurately.

Last, hedges are barely used in CSR and financial statements and notes. Unlike financial and operational performance, CSR uses self-regulating mechanism and states the facts of impactions on environment, employees, customer, the community and other stakeholders. The words used in CSR should be honest and objective. Similarly, words used in financial statements and note should be precise and not subjective.

4.6 Evaluative words

Through analyzing first 400 keywords in nine sub-corpora, 21 positive words and 5 negative words were obtained in table 6.

Table 6

Evaluate Words from First 400 Keywords in Sub-corpora

Positive Words	Negative Words
BENEFIT/BENEFITS/BENEFITED	DECREASE/DECREASED
BEST	LOSS/LOSSES
EFFECTIVE/EFFECTIVENESS	UNREALISED
GAIN	
GROWTH	
INCREASE /INCREASES/INCREASED	
LEADING	
PROFIT/PROFITS/PROFITABLE/PROBITABILITY	
SIGNIFICANT/SIGNIFICNATLY	
SUCCESS/SUCCESSFUL/SUCCESSFULLY	

In general, ARs use more positive words than negative words. The result confirms and expends Tomas's study that more words were used to present positive information in chairman's statement. Using positive words can make a good impression on the readers that the company has potential to improve and achieve in the future. In detail, the number of evaluative keywords used in each sub-corpus is counted as follows.

Table 7

The Number of Evaluate Keywords within First-400 Keywords

Sub-corpus	1	2	3	4	5	6	7	8	9
Positive Keyword Number	2	7	7	6	3	1	2	7	-
Negative Keyword Number	-	-	-	3	-	-	-	3	-

The table shows that positive words are noticeably more frequent in the section of chairman's statement, business overview, management's discussion and analysis and financial statements than in the section of company description, CSR and auditor's report such as "growth", "success", "best", "profitable", "leading" and "successfully". Furthermore, in describing the corporate profile, two positive words "leading" and "best" are frequently used. In comparison, CSRs use only three evaluative words such as "best", "leading" and "success" in reporting environmental issues.

However, only two positive words "effectiveness" and "effective" are shown in the first 400 keywords in auditor's reports because these two words are used frequently in expressing auditor's opinions that whether financial statements have been complied with corresponding accounting requirements. In the section of CG, the only positive word used is "effectiveness" since this section focus on management efficiency rather than financial performance.

In contrast, such negative words as "decreased", "decrease" and "decreases" are used frequently in management's discussion and analysis while "loss", "losses" and "unrealized" are used frequently in financial statements. However, negative words are not frequently used in other sections of ARs.

Conclusion

To sum up, this study analyzes lexical features in CARs and compares the words used in different sections of CARs based on one self-built large corpus and nine self-built small sub-corpora.

Firstly, financial aspect is mostly concerned which indicates that CARs are external financial reporting with necessary financial data to disclose and forecast organizations' performance.

Secondly, vocabulary of chairman's statements is the richest following by that of CSR report. Company description and business overview are relatively enriched with vocabulary while auditor's report has the least vocabulary density. The result confirmed Tian and Liang's study (2011) that auditor's report is not enriched with vocabulary.

Moreover, vocabulary used in auditor's reports is of the most difficulty. The result proved Tian and Liang's study (2011) that auditor's reports use longer words to enhance the degree of formality and attract reader's attention. Words in CSR reports are more complex than those in chairman's statements. However, words used in investor information and management's discussion and analysis are relatively easier. Conversely, financial statements use the simplest words due to restricted formats and a large number of accounting terms.

Furthermore, first personal pronouns "our", "we", "I", "my" are noticeably more frequent in the section of chairmen's statement to create affinity, build corporate image and resonate with readers. The word "our" occurs frequently in the section of business overview and management's discussion and analysis contributing to narrow the distance between the organizations and readers and get more approvals by readers. However, first personal pronouns are hardly used in CG report, financial statements and investor information.

The results also showed that hedges are more frequently used in the section of chairman's statements and business overview for the purpose to make expressions more flexible, efficient and avoid risks. Moreover, quantifier "many" is the most frequently used word in the sections of corporate profile, letters to shareholder and business overview reflecting organization's strategy of self-protection.

Finally, ARs use more positive words than negative words to make a good impression on readers that the company has potential to improve and achieve the success in the future, especially in the section of chairman's statement, business overview, management's discussion and analysis and financial statements.

It is hoped that this study has contributed to a better understanding of the words used in written financial discourse. Further studies will be conducted on the same self-built corpora in genre based approach to help writers and readers use CARs more efficiently.

About Authors

First Author

Huili Wang, School of Foreign Languages, Dalian University of Technology, China.

Prof. Wang Huili works in the School of Foreign Languages at Dalian University of Technology, China, and has been teaching English as a foreign language since 1989. She has written over 20 textbooks, and has published over 20 articles in journals and presented two papers at the IEEE international conferences. She has been responsible for about five projects concerning EFL teaching and three projects involving psycholinguistics and cognitive linguistics. She is a professor and an MA adviser in applied linguistics. As the first participant of the teaching reform projects, she was awarded second prize in Liaoning Province.

Address: Dalian University of Technology, Dalian, 116024, P.R. China. Contact: hilarydut@gmail.com

Second Author

Jingxiang Cao, School of Foreign Languages, Dalian University of Technology, China

Jingxiang Cao is an associate professor at the School of Foreign Languages, Dalian University of Technology, China. She was born in Changting, China in 1973. She finished her MA in Linguistics and Applied Linguistics in Dalian Maritime University in 2000 and is doing her PhD in Natural Language Processing in the Department of Computer Science, Dalian University of Technology. She has been teaching Statistics in Linguistics to MA students and supervising MA theses in Corpus Linguistics and Language Acquisition.

Address: School of Foreign Languages, Dalian University of Technology, Dalian, 116024, P.R. China. Contact: alicia1973@yahoo.cn

Third Author

Lixin Li, School of Foreign Languages, Dalian University of Technology, China

Li Lixin was born in January 30, 1988 in China. She holds a Bachelor of Science in Management from Liaoning University, Shenyang, China. She has been a full-time postgraduate student majoring in foreign linguistics and applied linguistics at China's Dalian University of Technology since September 1, 2010. She was awarded the ACCA (Association of Certified Chartered Accountants) Advanced Diploma in Accounting and Business in August 2011.

Address: School of Foreign Languages, Dalian University of Technology, Dalian, 116024, P.R. China. Contact: phoebe.xuanyuan@gmail.com

References

1. Back, B., Toivonen, J., Vanharanta, H., & Visa, A. (2001). Comparing numerical data and text information from annual reports using self-organizing maps. *International Journal of Accounting Information Systems*, 249-269.

2. Biber, D., Conrad, S. and Johansson, S. (2009). *Longman grammar of spoken and written English*. Beijing: Foreign Language teaching and research Press

3. Bowman, E. H. (1984). Content analysis of annual reports for corporate strategy and risk. *Interfaces 14*, 61-71

4. Bums, L.C. (2000). *Vagueness language*. Shanghai: Shanghai Foreign Language Teaching Press.

5. Courtis, J. K. (1998). Annual report readability variability: Tests of the obfuscation hypothesis. *Accounting, Auditing & Accountability Journal*, 459-471.

6. Channel, J. (2000). *Vague language*. Shanghai: Shanghai Foreign Language Teaching Press.

7. Enkvist, N.E. (1969). *On defining style: an essay in applied linguistics*. London: Oxford University Press

8. Frazier, K. B., Ingram, R. W., & Tennyson, B. M. (1984). A methodology for the analysis of narrative accounting disclosures, *Journal of Accounting Research*, *21*(1), 318-331.

9. Hildebrandt, H. W., & Snyder, R. D. (1981). The Pollyanna hypothesis in business writing: Initial results, Suggestions for Research, *Journal of Business Communication*, *18*(1), 5-15.

10. Hynes, G. E., & Bexley, J. B. (2004). The contribution of banks' annual report writing quality to investor decision-making, *Journal of Commercial Banking and Finance*, *3*(2), 113-122.

11. Kelih, E., Antic, G., Grzybek, P. and Stadloker, E. (2005). Classification of author and genre: The impact of word length. In: Weihs, Claus; Gaul, Wolfgang (Eds.), *Classification the Ubiquitous Challenge* (pp. 488-505). Heidelberg: Springer.

12. Kendall, J.E. (1993). Good and evil in the chairmen’s ‘boiler plate’: An analysis of corporate visions of the 1970s. *Organizations Studies*, 271-592

13. Kennedy, G. (2000). *An introduction to corpus linguistics*. Foreign Language Teaching and Research Press.

14. Kloptchenko, A., Back, B., Vanharanta, H., Eklund, T., Karlsson, J., & Visa, A. (2002). Combining data and text mining techniques for analyzing financial reports. *Eighth Americas Conference on Information Systems*, 20-28.

15. Kohut, G. F., & Segars, A. H. (1992). The president's letter to stockholders: An examination of corporate communication strategy, *Journal of Business Communication*, *29*(1), 7-21.

16. Lakoff, G. (1973). Hedges: A study in meaning criteria and logic of fuzzy concepts. *Journal of Philosophical Logic*, *2*, 458-508

17. Leech, G. (1992). Corpora and theories of linguistic performance, In J. Svartvik, (Ed.). *Directions in Corpus Linguistics* (pp. 105-122). Berlin: Mouton de Gruyter.

18. Magnusson, C., Arppe, A., Eklund, T., Back, B., Vanharanta, H., & Visa, A. (2005). The language of quarterly reports as an indicator of change in the company’s financial status, *Information & Management*, *42*(4), 561-574.

19. Mendenhall, C. (1887). The characteristic curves of composition. *Science, 9*(214), 237-249

20. Mohammad, A. S. N. (2011). The place of genre analysis in international communication, *International Journal of Language Studies*, *5*(1), 63-74

21. McEnery, T., R. Xiao and Y. Tono (2005). *Corpus-based language studies: An advanced resource book.* Routledge Press

22. Nation, L.S.P. (1990). *Teaching and Learning Vocabulary*. New York: Newbury House Publishers Press.

23. Richards, J.C., J. Platt & H. Platt. (2000). *Longman dictionary of language teaching and applied linguistics.* Foreign Language Teaching and Research Press.

24. Santema, S., & Rijt, J. V. (2001). Strategy disclosure in Dutch annual reports, *European Management Journal*, *19*(1), 101-108.

25. Scott, M. (2004). *Wordsmith Tools 4.0.* Oxford: Oxford University Press.

26. Sichel, H (1986). Word frequency distribution and type-token characteristics. *Mathematical Scientist*, *11*, 45-72

27. Sinclair, J.(1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press,

28. Subramanian, R., Insley, R. G., & Blackwell, R. D. (1993). Performance and readability: A comparison of annual reports of profitable and unprofitable corporations, *Journal of Business Communication*, *30*(1), 49-61.

29. Tian, X. and Liang, H. (2011), A corpus based study on stylistic analysis of auditing reports. *Market Modernization 642*, 136-138

30. Thomas, J. (1997). Discourse in the marketplace: The making of meaning in annual reports, *Journal of Business Communication*, *34*(1), 47-66.