Large-sample automated analysis of textual data in accounting research

Steven Young
Lancaster University

European Accounting Association PhD Forum

28 April 2015, Glasgow

Research funded by ESRC and ICAEW

TEXTUAL ANALYSIS IN ACCOUNTING RESEARCH

The power of analysing language

- Analyzing digital personas to provide investigators with clues to the identity of the individual or group hiding behind one or more personas
  - Digital personas play a key role in criminal tactics in online social media
    - **Child sex offenders masquerading as young people to gain their victims' trust**
    - **Radicalization of youth in online forums through persuasive messaging**
    - **Online dating sites that gain trust and the exploit users for financial gain**
  - Use textual analysis to predict a persona's key attributes (e.g., age and gender) even where the author is obfuscating the language (e.g., Rashid et al. 2013)

- Identifying dementia and Alzheimer's disease
  - Longitudinal changes in linguistic features (e.g., grammatical complexity and idea density) are associated with degenerative memory conditions
  - Use textual analysis to construct risk factors (e.g., Kemper et al. 2004; Le et el. 2011) to identify/predict dementia and Alzheimer's

TEXTUAL ANALYSIS IN ACCOUNTING RESEARCH

Text and accounting: Overview

- Text analytics is a growing area of academic and professional interest:
  - Large fraction of data produced as part of the financial reporting process (broadly defined) takes the form of text
  - Professional investors and regulators (e.g., SEC) using text as a screening tool

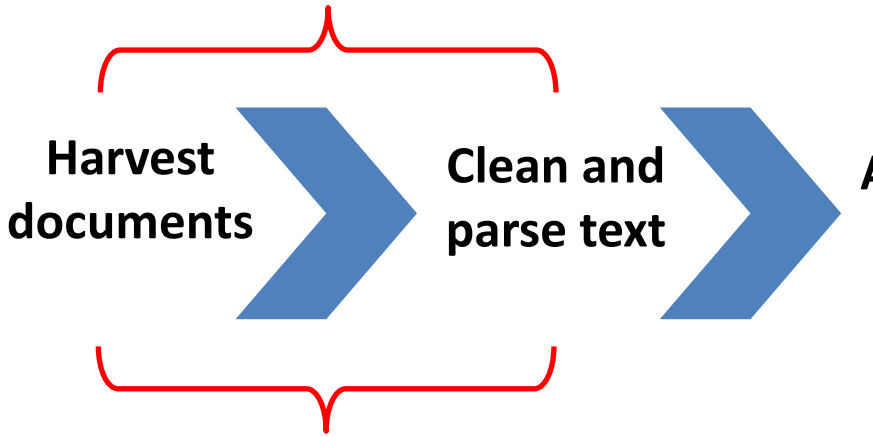
- Extensive textual resources exist for accounting researchers
  - Annual reports and other mandatory filings (e.g., 8-K, 14-A, including CD&A)
  - Earnings announcements
  - Conference call transcripts
  - Analyst reports and briefings
  - Media articles
  - Regulatory documents (e.g., accounting standards, exposure drafts, etc.)

- Large sample analysis of textual disclosure is all about data reduction → aggregate large amounts of text into simple numerical metrics (Li 2010a)

Textual Analysis in Accounting Research

Textual analysis

Text from some sources is much easier to scrape and parse than from other sources



- Bag-of-words analyses: dictionaries & readability
- Measure constructs using off-the-shelf software (e.g., Diction)
- Machine learning (training software to recognise patterns)
- Semantic-level analysis to capture meaning
- Text mining to identify patterns

Requires a level of programming competence:

- **Pearl widely used in accounting literature**
- **Other languages as good if not better (e.g., Python, Java, etc.)**
- **Most statistical packages offer text processing capabilities (e.g., SAS, R, SPSS, etc.)**

TEXTUAL ANALYSIS IN ACCOUNTING RESEARCH

Textual analysis: Methods

- Bag-of-words approaches: Dictionaries
  - Word lists for measuring positivity, negativity, uncertainty, forward-looking, etc. (General Inquirer, Diction, etc.)
  - Optimizing the dictionary for financial context is key (e.g. Henry 2008, Loughran and McDonald 2011)
  - Ignores context and meaning (semantic level)

- Bag-of-words approaches: Fog Index and document length
  - Proxies for document readability and complexity (Li 2008) → hard to disentangle the two constructs
  - Naïve summary measures at best → e.g., long documents aren't necessarily more complex
  - Naïve summary measures at best → Fog Index is independent of word order and complex words are defined in the context of high school children

TEXTUAL ANALYSIS IN ACCOUNTING RESEARCH

Textual analysis: Methods *cont.*

- Machine learning (e.g., Naïve Bayes, Li 2010b)
  - Involves training a computer algorithm to classify text into distinct categories based on a sample of manually coded text
  - Appears sophisticated but involves hidden problems including replicability (Loughran and McDonald 2015)
  - Most accounting papers use Naïve Bayes despite evidence from NLP literature that other methods may work better (e.g., Logistic or Random Forest)

- Semantic analysis and parts-of-speech
  - Understand meaning $\rightarrow$ disambiguation
  - Assign properties (e.g., verb, noun, adjective), and hence meaning, to words based on the context in which they are used
  - Central issue for researchers in linguistics but not currently applied in accounting!

TEXTUAL ANALYSIS IN ACCOUNTING RESEARCH

Textual analysis: Methods *cont.*

- Text mining
  - Methods developed in computer science → similar to neutral networks and AI
  - Data-driven text classification models based on statistical relations (e.g., Balakrishnan et al. 2010, Lee et al. 2014)
  - No attempt to understand properties of the text → classifier is pure black-box

TEXTUAL ANALYSIS IN ACCOUNTING RESEARCH

Textual analysis: Evidence (Li 2010a)

- Information content and accounting quality
  - MD&A disclosures are value relevant and predict future earnings (Bryan 1997)
  - But (some) disclosures are designed to obfuscate (Huang et al. 2014, Li 2012)
- Market efficiency
  - Evidence that the stock market may not understand fully the implications of textual disclosure (Feldman et al. 2010, Li 2010b)
- Information environment
  - 10-K readability linked with small investor trading (Miller 2010, Loughran & McDonald 2014) and analyst research (Lehavy et al. 2011, Brown & Tucker 2011)
- Litigation
  - Plaintiffs target optimistic statements in lawsuits and sued firms' earnings announcements are abnormally optimistic (Rogers et al. 2011)

TEXTUAL ANALYSIS IN ACCOUNTING RESEARCH

Annual reports

- Most firms required to publish an annual report document combining financial statement data with textual commentary

- Outside the U.S., these reports tend to be "glossy", **unstructured** documents (provided as **PDF**s) that include:
  - Photographs, charts, graphics
  - Key narratives including a letter to shareholders from the chairman, a review of strategy, information on governance, compensation, and CSR policies, etc.

- U.S. registrants also required to file their glossy annual report to the SEC using a **structured template** comprising 15 reporting items (Form 10-K)
  - 10-K filed through **EDGAR** as a **plain text document** with no pictures, graphics, etc. $\rightarrow$ facilitates document retrieval and processing on a large scale
  - Primary focus of the extant literature on annual reporting

TEXTUAL ANALYSIS IN ACCOUNTING RESEARCH

Annual report software tool

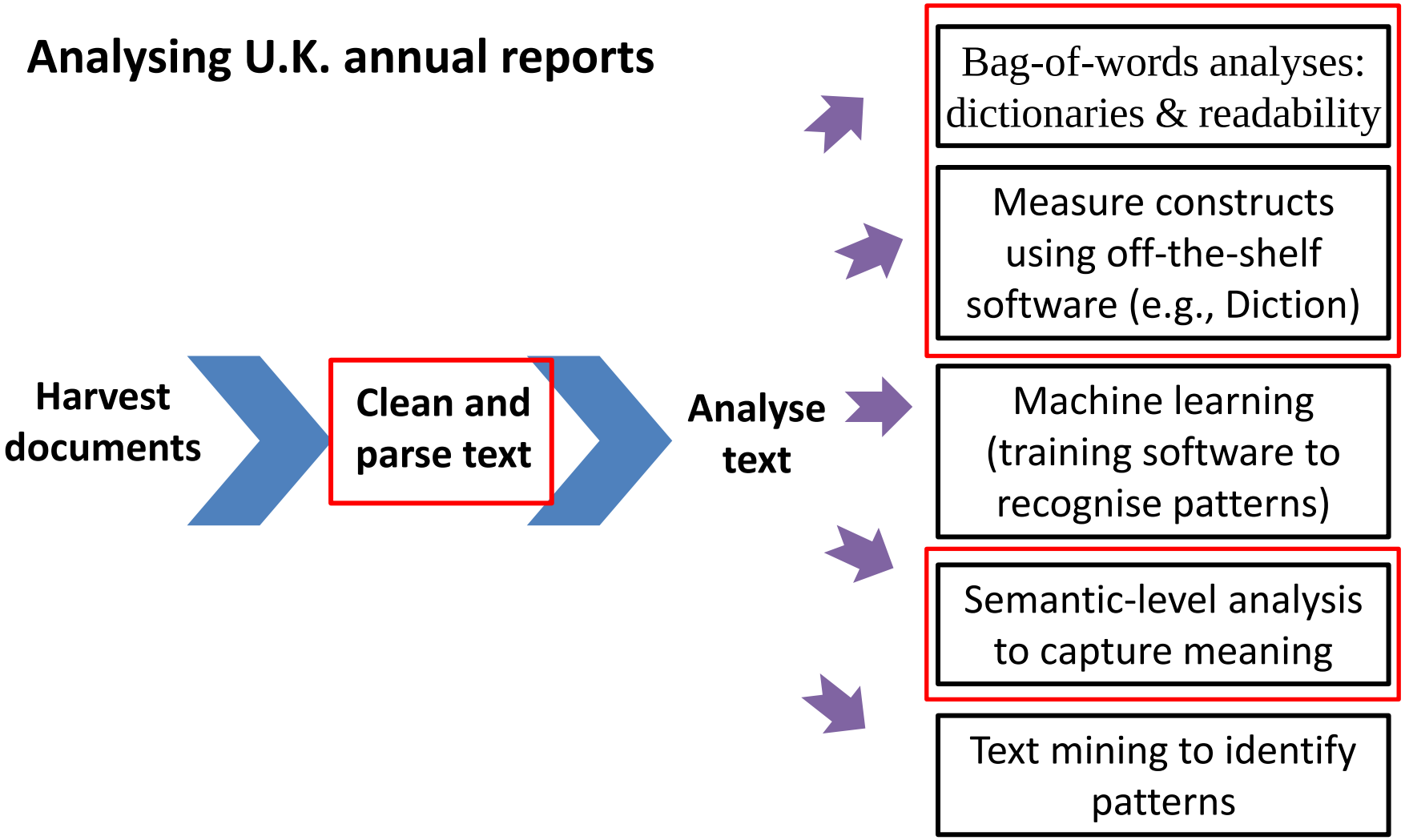
- Develop a software tool to extract and process narrative disclosures from U.K. annual reports

- The tool:
  - Web-based system based on Java script integrating iText open source libraries
  - Generic in the sense it is designed to extract all text from annual reports → focus is not restricted to particular type of content
  - Can be tweaked to process annual reports in other jurisdictions and languages
  - Free access for non-commercial purposes (https://cfie.lancaster.ac.uk:8443/)
  - Interfaces with **W-matrix** to provide a full suite of corpus linguistics tools (e.g., semantic tagging, keyness, concordance, collocation, n-grams, etc.)

- Tool is described and evaluated (and empirical constructs validated) in a companion paper based on a sample of approx. 12,000 annual reports

Textual Analysis in Accounting Research

Analysing U.K. annual reports



Analysing Annual Reports: Method

Extraction method

- Use annual report table of contents to extract text from digital pdf

- Summary of extraction process:
  - Detect table of contents
  - Parse contents page
  - Synchronize page numbering in table of contents with pages in pdf
  - Use synchronized page numbers to determine start/end of sections
  - Extract text by section

Contents

Analysing Annual Reports: Method

Parsing the contents page

What's in our report

Features

Sir Ian Gibson
We are committed to making food shopping fresh, friendly and affordable.
Page 2

Dalton Philips
A clear strategy is in place that is delivering our objectives.
Page 4

Richard Pennycook
Our strong financial performance positions us well for sustainable long term growth.
Page 8

Johanna Waterous
A strong performance culture, long term shareholder value and competitive positioning remain key principles.
Page 46

Also see...

Corporate responsibility review 2011/12

Annual review 2011/12

Investor relations website
www.morrisons.co.uk/corporate

Directors' report and business review

Introduction
2 Chairman's statement

Business and strategy review
4 Chief Executive's business and strategy review
8 Group Finance Director's financial review
12 Our strategic objectives

Performance review
26 Key performance indicators
28 Risks and uncertainties
30 Corporate responsibility
33 Our people

Governance
36 Board of Directors and Management Board
40 Corporate governance report
46 Directors' remuneration report
56 General information
59 Statement of Directors' responsibilities

Financial statements

60 Group financial statements
60 Independent auditor's report
61 Consolidated statement of comprehensive income
62 Consolidated balance sheet
63 Consolidated cash flow statement
64 Consolidated statement of changes in equity
65 Group accounting policies
70 Notes to the Group financial statements

95 Company financial statements
95 Company balance sheet
96 Company accounting policies
99 Notes to the Company financial statements

Investor information
108 Five year summary of results
109 Supplementary information
110 Investor relations and financial calendar

Analysing Annual Reports: Method

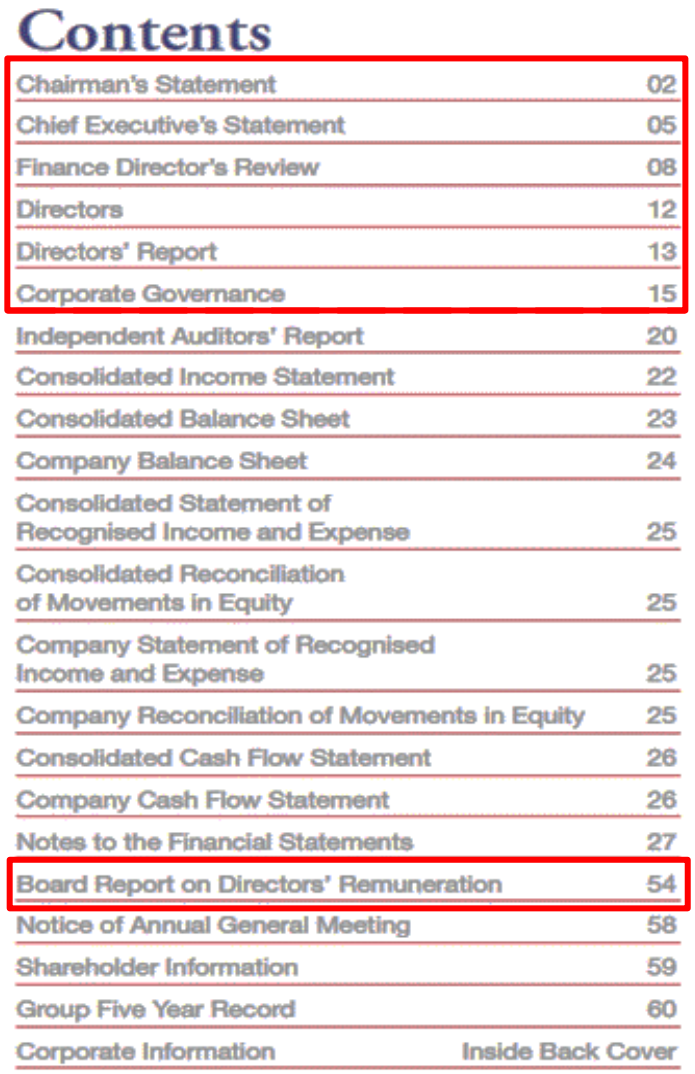
Parsing the table of contents

Contents

ANALYSING ANNUAL REPORTS: METHOD

Extraction method

- Use annual report table of contents to extract text from digital pdf
- Summary of extraction process:
  - Detect table of contents
  - Parse contents page
  - Synchronize page numbering in table of contents with pages in pdf
  - Use synchronized page numbers to determine start/end of sections
  - Extract text by section
- Analyse entire narrative component and by section
  - Problem → unstructured format



Contents

Analysing Annual Reports: Method

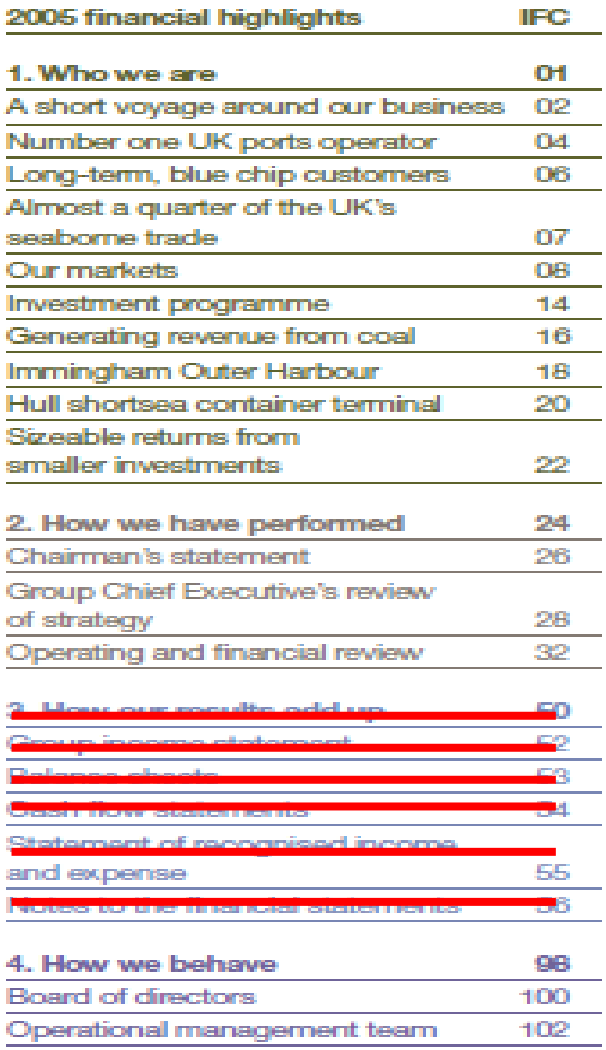
Partitioning method

- Partition annual report into front and back components based on median annual report structure, with adjustments for idiosyncratic deviations

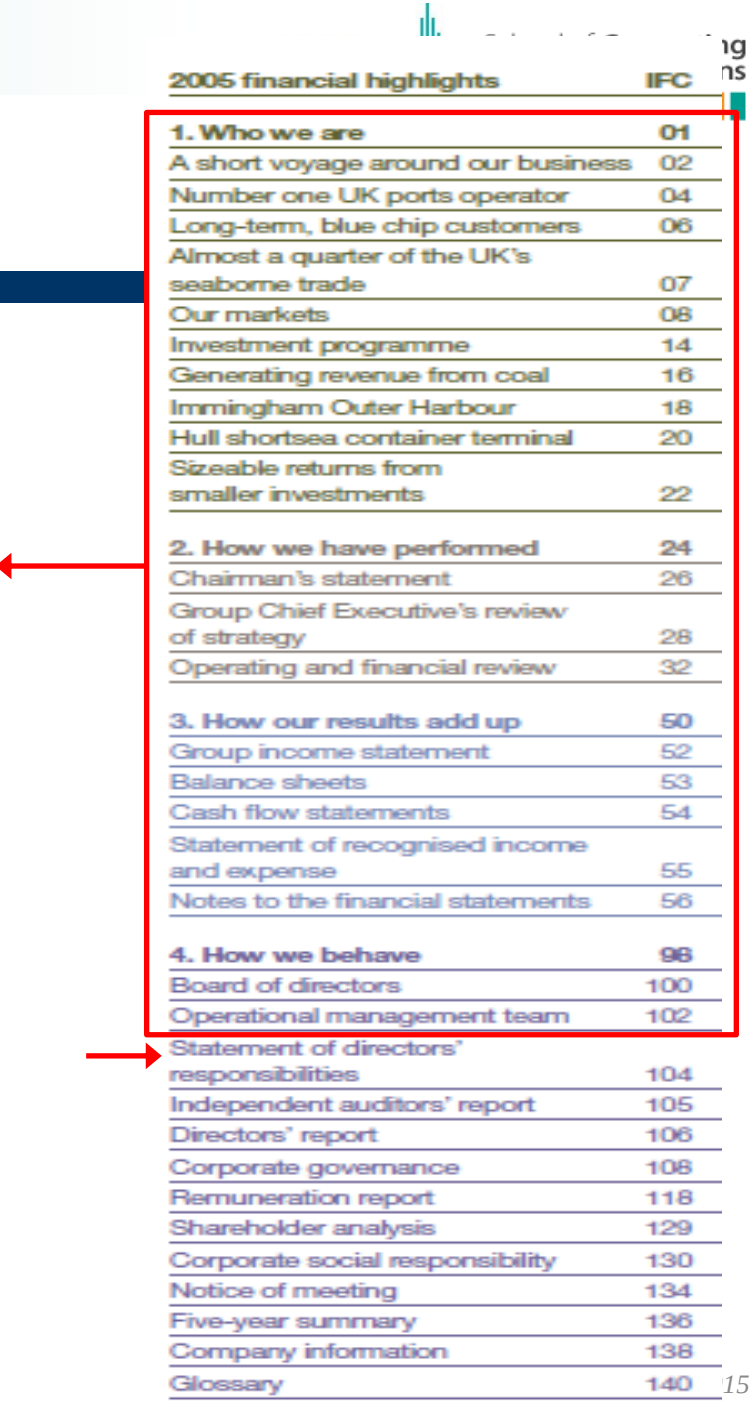
	Section	Comments
1.	Section	Comments
2.	Overview	(including highlights)
3.	Chairman's statement	
4.	Performance commentary	(including one or more of the following sections: chief executive's review, review of operations, business review, strategic review, financial review)
5.	Other sections	(various, common examples of which include risk review and corporate social responsibility report)
6.	Director's biographies	
7.	Directors' report	
8.	Governance statement	
9.	Remuneration Report	
10.	Statement of directors' responsibilities	
11.	Auditor's report	
12.	Primary financial statements	(as required by IAS 1)
13.	Notes to the accounts	(as required by IAS 1)
14.	Other disclosures	(various, common examples of which include notice of annual general meeting, three- or five-year review, subsidiaries and operating locations, etc.)

Analysing Annual Reports: Method

Example



2005 financial highlights	IFC
1. Who we are	**01**
A short voyage around our business	02
Number one UK ports operator	04
Long-term, blue chip customers	06
Almost a quarter of the UK's seaborne trade	07
Our markets	08
Investment programme	14
Generating revenue from coal	16
Immingham Outer Harbour	18
Hull shortsea container terminal	20
Sizeable returns from smaller investments	22
2. How we have performed	**24**
Chairman's statement	26
Group Chief Executive's review of strategy	28
Operating and financial review	32
3. How our results add up	**50**
Group income statement	52
Balance sheets	53
Cash flow statements	54
Statement of recognised income and expense	55
Notes to the financial statements	56
4. How we behave	**98**
Board of directors	100
Operational management team	102



2005 financial highlights	IFC
1. Who we are	**01**
A short voyage around our business	02
Number one UK ports operator	04
Long-term, blue chip customers	06
Almost a quarter of the UK's seaborne trade	07
Our markets	08
Investment programme	14
Generating revenue from coal	16
Immingham Outer Harbour	18
Hull shortsea container terminal	20
Sizeable returns from smaller investments	22
2. How we have performed	**24**
Chairman's statement	26
Group Chief Executive's review of strategy	28
Operating and financial review	32
3. How our results add up	**50**
Group income statement	52
Balance sheets	53
Cash flow statements	54
Statement of recognised income and expense	55
Notes to the financial statements	56
4. How we behave	**98**
Board of directors	100
Operational management team	102
Statement of directors' responsibilities	104
Independent auditors' report	105
Directors' report	106
Corporate governance	108
Remuneration report	118
Shareholder analysis	129
Corporate social responsibility	130
Notice of meeting	134
Five-year summary	136
Company information	138
Glossary	140

Analysing Annual Reports: Method

Example

2005 financial highlights	IFC
1. Who we are	**01**
A short voyage around our business	02
Number one UK ports operator	04
Long-term, blue chip customers	06
Almost a quarter of the UK's seaborne trade	07
Our markets	08
Investment programme	14
Generating revenue from coal	16
Immingham Outer Harbour	18
Hull shortsea container terminal	20
Sizeable returns from smaller investments	22
2. How we have performed	**24**
Chairman's statement	26
Group Chief Executive's review of strategy	28
Operating and financial review	32
3. How our results add up	**50**
Group income statement	52
Balance sheets	53
Cash flow statements	54
Statement of recognised income and expense	55
Notes to the financial statements	56
4. How we behave	**98**
Board of directors	100
Operational management team	102
Statement of directors' responsibilities	104
Independent auditors' report	105
Directors' report	106
Corporate governance	108
Remuneration report	118
Shareholder analysis	129
Corporate social responsibility	130
Notice of meeting	134
Five-year summary	136
Company information	138
Glossary	140

Analysing Annual Reports: Method

Partitioning method *continued*

- Partition narratives (front) component into the following generic categories to facilitate cross-sectional and temporal analysis:
  - Chairman's statement
  - CEO review
  - Strategic commentary
  - Performance commentary (incl. CEO review, strategic review, finance director's review, operating review, business review, etc.)
  - Governance statement (incl. chairman's introduction, separate committee statements, statement on internal control, etc.)
  - Remuneration report
  - Residual commentary (incl. overview, highlights, CRS report, principal risks and uncertainties, directors' report, etc.)

- Methods described in the paper → inconsistent labelling makes more granular analysis difficult

ANALYSING ANNUAL REPORTS: METHOD

U.K. Annual report software tool and data

- In addition to performing text extraction, the tool provides a range of text analysis options:
  - Readability metrics
  - Word counts for pre-defined dictionaries (e.g., Henry 2008, Loughran & McDonald 2011) and user-defined dictionaries
  - Comparison with reference corpus (word level and semantic level)
  - Concordance and collocates
  - N-grams
  - Upload and analyse user-defined text file
- Tool able to process approx. 84% annual reports between 2003 and 2013
  - Majority of non-processed reports are image-based (scanned) files
- Use matching algorithm based on firm name to merge textual metrics with financial and market data from Datastream

U.K. Annual Report Narratives: Evidence

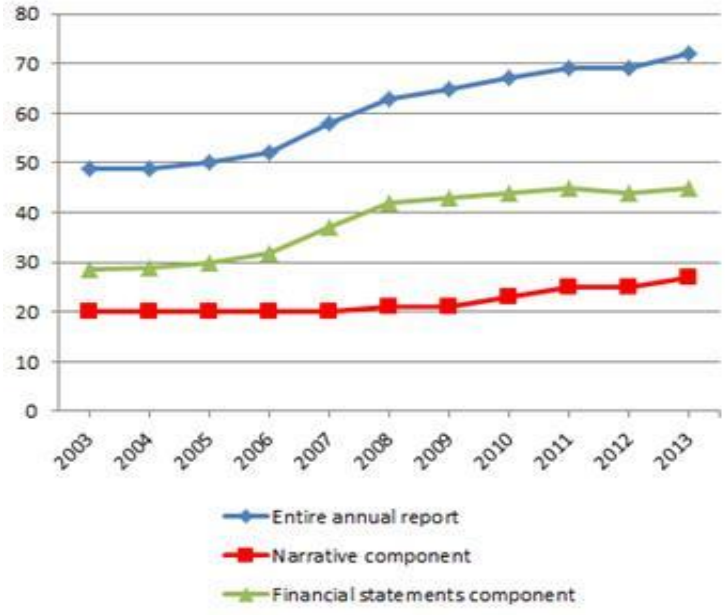
Summary statistics

	N	Mean	Stdev	Median
Page count (annual report)	10,228	70.55	39.04	60.00
Page count (narratives)	10,228	28.25	22.17	22.00
Page count (financial statements)	10,228	42.31	21.99	38.00
N sections (narratives)	10,228	9.00	5.99	8.00
N sections (financial statements)	10,228	10.21	4.03	10.00
Word count (narratives)	10,228	20956.38	20687.30	14567.00
Word count (financial statements)	10,228	25255.64	24221.90	19959.00
Net tone (narratives)	10,228	0.12	0.17	0.13
Forward looking content (narratives)	10,228	0.01	0.01	0.01
Uncertainty content (narratives)	10,228	0.03	0.01	0.03
Causal reasoning content (narratives)	10,228	0.01	0.01	0.01
Readability (narratives)	10,228	24.69	36.30	21.66

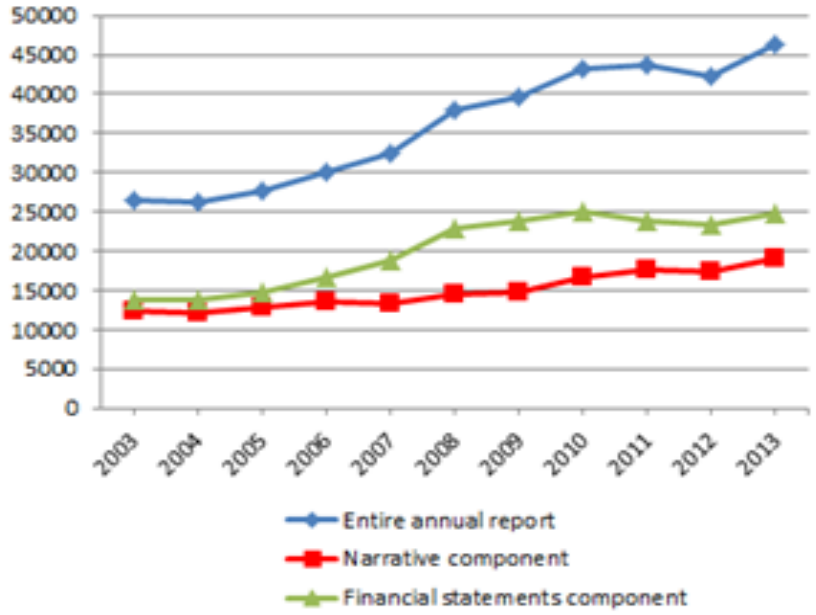
U.K. Annual Report Narratives: Evidence

Page count and word count by time

Page count



Word count



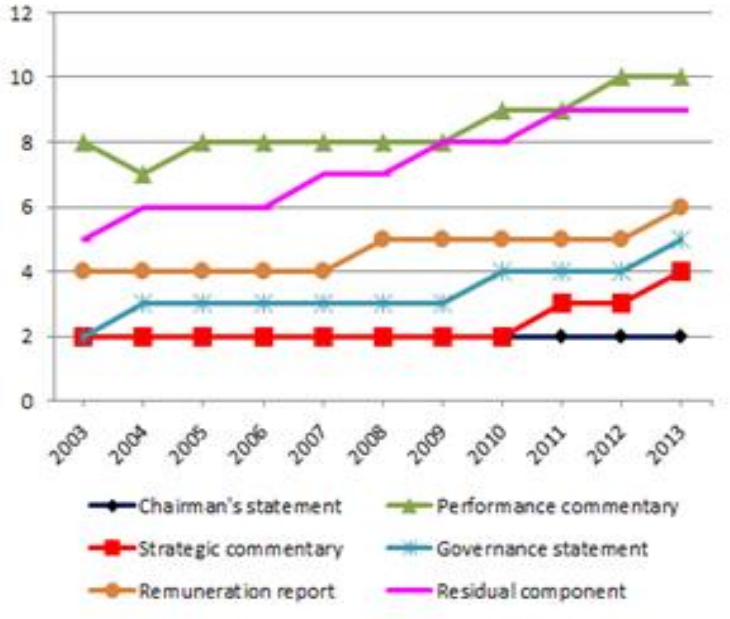
LANCASTER UNIVERSITY Management School

School of Computing and Communications

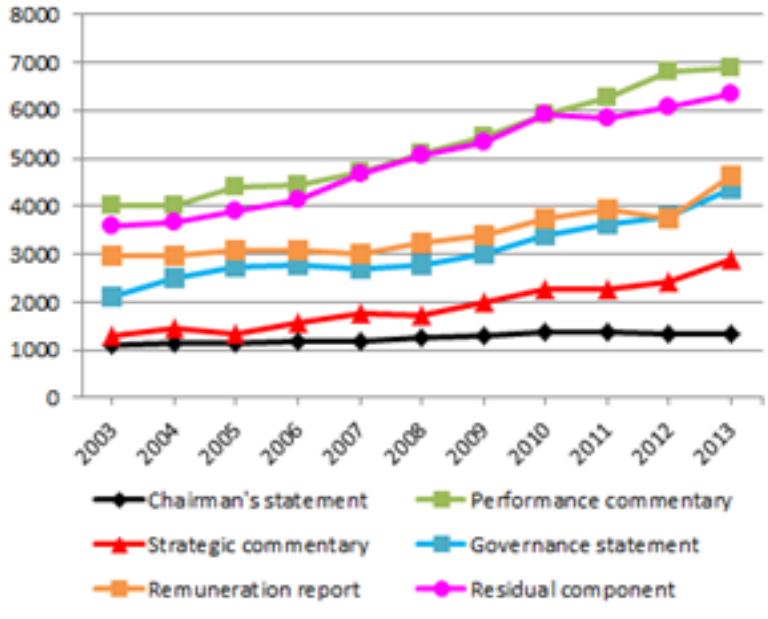
U.K. Annual Report Narratives: Evidence

Page count and word count by time

Page count



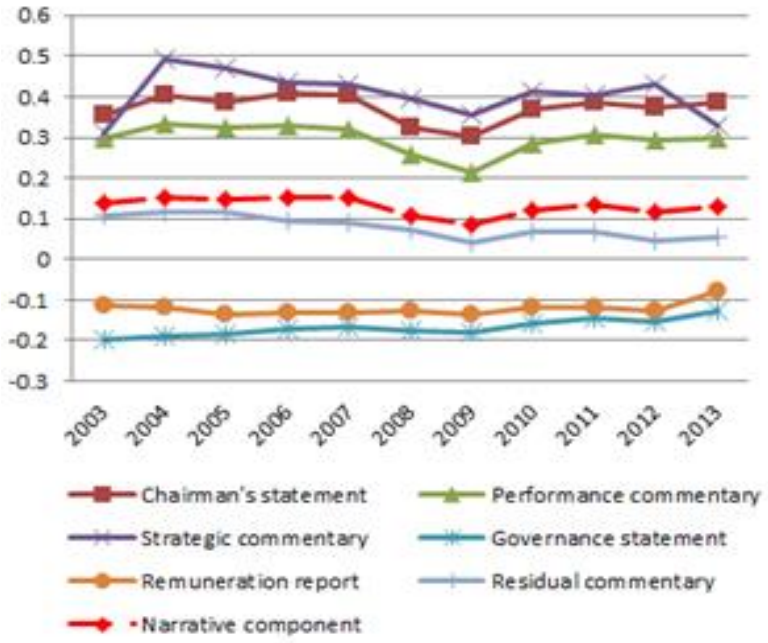
Word count



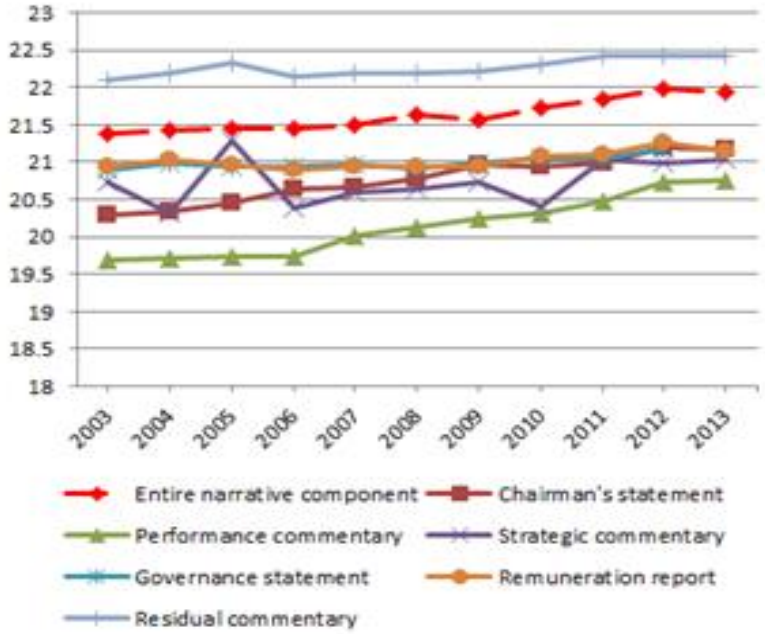
U.K. Annual Report Narratives: Evidence

Properties by section

Net tone



Readability



TEXTUAL ANALYSIS IN ACCOUNTING: RESEARCH

Textual analysis: Caveats

- Recognize limitations of automated analysis
  - Language is complex → serious dangers in trying to reduce this complexity to a few simple summary dimensions such as reabability and tone
  - Many important questions are hard to address using automated methods
- Big is not always better
  - Don't be seduced by "quasi rigor" → devil is in the detail
- Results can be extremely sensitive to choice of methods and parameters (Loughran & McDonald 2015) → transparency and replicability are crucial
- Avoid "discipline arrogance" → engage with specialist fields rather than relying on prior work in accounting and finance
- Avoid the "street light fallacy" → easy access doesn't make it interesting

TEXTUAL ANALYSIS IN ACCOUNTING: RESEARCH

Summary & conclusions

- Text analytics is a growing area of academic and professional interest in accounting and finance
- Extensive textual resources exist for accounting researchers
- Most analysis of annual report text has focused on U.S. disclosures (10-K) due to (relative) ease of document access and processing → EDGAR
- New software tool available for analysing "glossy" annual reports where management have much more discretion over content and presentation
  - http://ucrel.lancs.ac.uk/cfie/index.php
- Large sample automated analysis of accounting still in its infancy
  - Much to learn from foundation disciplines such as linguistics and computing

TEXTUAL ANALYSIS IN ACCOUNTING: RESEARCH

References

- Balakrishnan, R., X. Y. Qiu & P. Srinivasan, 2010. On the predictive ability of narrative disclosure in annual reports. *European Journal of Operational Research* 202: 780-801
- Brown, S. & J. Tucker, 2011. Large sample evidence on firms' year-over-year MD&A modifications. *Journal of Accounting Research* 49(2): 309-348
- Bryan, S. H., 1997. Incremental information content of required disclosures contained in Management Discussion and Analysis. *The Accounting Review* **77**(2): 285-301
- Feldman, R., S. Govindaraj, J. Livnat & B. Segal, 2010. Management's tone change, post-earnings announcement drift and accruals. *Review of Accounting Studies* 15(4): 915-953
- Henry, E., 2008. Are investors influenced by how earnings press releases are written. *Journal of Business Communication* 45: 363-407
- Huang, X., S. H. Teoh & Y. Zhang, 2014. Tone management. *The Accounting Review* 89(3): 1083-1113
- Kemper, S., L. Greiner, Marquis, J., K. Prenovost, T. Mitzner (2004). Language decline across the life span: Findings from the nun study. *Psychology and Aging* 16(2): 227-239
- Larcker, D. & A. Zakolyukina (2012). Detecting deceptive discussions in conference calls. *Journal of Accounting Research* 50(2): 495-540
- Le,, X., I. Lancaster, G. Hirst, R. Jokel (2011) Longitudinal detection of dementia through lexical and syntactic changes in writing: A case study of three British novelists. *Literary and Linguistic Computing*: 1-27
- Lee, H., M. Surdeanu, B. MacCurtney, D. Jurafsky (2014). On the importance of text analysis for stock price prediction. Unpublished working paper http://www.stanford.edu/~jurafsky/pubs/lrec2014_stocks.pdf

TEXTUAL ANALYSIS IN ACCOUNTING: RESEARCH

References

- Lehavy, R., F. Li & K. Merkley, 2011. The effect of annual report readability on analyst following and the properties of their earnings forecasts. *The Accounting Review* 86: 1087-1115
- Li, F. (2008). Annual report readability, current earnings, and earnings persistence. *Journal of Accounting and Economics* 45: 221–47
- Li, F., 2012. Managers' self-serving attribution bias and corporate financial policies. Available for download at: http://ssrn.com/abstract=1639005
- Li, F., 2010b. The information content of forward-looking statements in corporate filings: A naïve Bayesian machine learning approach. *Journal of Accounting Research* 48(5): 1049-1102. Li 2010a JAL
- Loughran, T., B. McDonald (2011). When is a liability not a liability? *Journal of Finance* 66, 35-65
- Loughran, T. & B. McDonald, 2014. Measuring readability in financial documents. *Journal of Finance* 69: 1643-1671
- Loughran, T. & B. McDonald, 2015. Textual analysis in accounting and finance: A survey. Available for download at: http://ssrn.com/abstract=XXXX
- Purda, L., D. Skillicorn (2014). Accounting variables, deception, and a bag of words: Assessing the tools of fraud detection. *Contemporary Accounting Research* forthcoming DOI: 10.1111/1911-3846.12089
- Rashid, A., A Baron, P. Rayson, C. May-Chahal, P. Greenwood, J. Walkerdine (2013). Who am I? Analyzing digital personas in cybercrime investigations. *Computer* (April): 54-61
- Rogers, J., A. Van Buskirk & S. Zechman, 2011 Disclosure tone and shareholder litigation. *The Accounting Review* 86: 2155-2183