Computer-based Analysis of UK Annual Report Narratives

FINANCIAL REPORTING & BUSINESS COMMUNICATION CONFERENCE

3 & 4 July 2014, Bristol

Steven Young

Mahmoud El-Haj

LANCASTER UNIVERSITY
Management School

School of **Computing** and **Communications**

ICAEW

E·S·R·C
ECONOMIC & SOCIAL RESEARCH COUNCIL

Session outline

- Background to project on narrative reporting

- Overview of software tool for analysing textual content in UK annual reports

- Demo of software tool

- Questions

- Example of applying software tool to study strategy-related disclosures in a large sample of UK annual reports
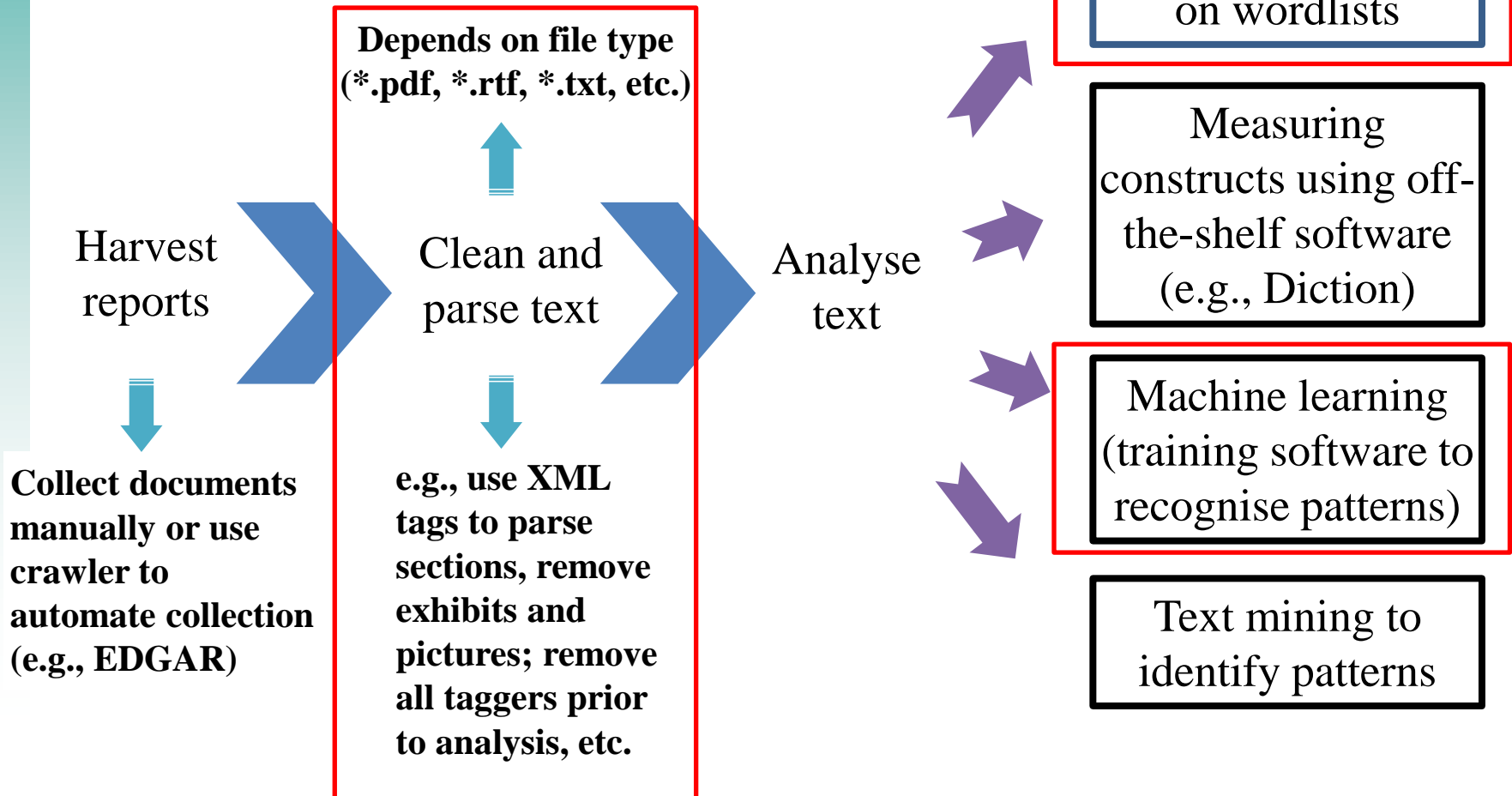
Background & objectives

- Part of an ESRC- and ICAEW-funded project examining the *Corporate Financial Information Environment*
  - Martin Walker, *Manchester Business School*
  - Steven Young, *Lancaster University Management School*
  - Paul Rayson, *Lancaster University School of Computing & Communications*
  - Mahmoud El-Haj, *Lancaster University School of Computing & Communications*
  - Vasiliki Athanasakou, *London School of Economics*

- Project seeks to analyse UK financial narratives, their association with financial statement information, and their informativeness for investors

- Automated, large sample analysis of UK annual report narratives represents a cornerstone of the project
  - Develop software for general use by academics

Analysing annual report narratives

Harvest reports

Collect documents manually or use crawler to automate collection (e.g., EDGAR)

Depends on file type (*.pdf, *.rtf, *.txt, etc.)

Clean and parse text

e.g., use XML tags to parse sections, remove exhibits and pictures; remove all taggers prior to analysis, etc.

Analyse text

Word counts based on wordlists

Measuring constructs using off-the-shelf software (e.g., Diction)

Machine learning (training software to recognise patterns)

Text mining to identify patterns

Extant research

- Majority of <u>large sample</u> analysis of annual report narratives has been conducted on US filings (10-Ks) available via EDGAR
  - Management Discussion and Analysis (MD&A) section (Item 7)
  - Risk-related disclosures (Item 1A and Item 7A)
  - Entire 10-K filing

- Analysis of 10-K filings in EDGAR is relatively straightforward
  - Plain text files with consistent structure
  - Use HTML/XML taggers to identify section(s) and extract text

EDGAR 10-K format extract

PART I

ITEM 1   Description of Business

ITEM 1A. Risk Factor

ITEM 1B. Unresolved Staff Comments

ITEM 2.   Description of Properties

ITEM 3.   Legal Proceedings

ITEM 4.   Mine Safety Disclosures

PART II

ITEM 5.   Market for Registrant's Common Equity, Related Stockholder Matters and Issuer Purchases of Equity Securities

ITEM 6.   Selected Financial Data

ITEM 7.   Management's Discussion and Analysis of Financial Condition and Results of Operations

ITEM 7A. Quantitative and Qualitative Disclosures About Market Risk

ITEM 8.   Financial Statements and Supplementary Data

ITEM 9.   Changes in and Disagreements With Accountants on Accounting and Financial Disclosure

ITEM 9A. Controls and Procedures

ITEM 9B. Other Information

EDGAR 10-K example

PART III

ITEM 10. Directors, Executive Officers and Corporate Governance

ITEM 11. Executive Compensation

ITEM 12. Security Ownership of Certain Beneficial Owners and Management and Related Stockholder Matters

ITEM 13. Certain Relationships and Related Transactions, and Director Independence

ITEM 14. Principal Accounting Fees and Services

PART IV

ITEM 15. Exhibits, Financial Statement Schedules Signatures

EDGAR 10-K example

<P style="font-family:times;text-align:justify"><FONT SIZE=2><A
NAME="dc47401_item_1._business"> </A>
<A NAME="toc_dc47401_1"> </A></FONT> <FONT SIZE=2><B> ITEM 1.   BUSINESS
  </B></FONT></P>
<P ALIGN="CENTER" style="font-family:times;"><FONT SIZE=2><B> GENERAL DEVELOPMENT OF BUSINESS </B></FONT></P>

<P style="font-family:times;text-align:justify"><FONT SIZE=2>   Abbott Laboratories is an Illinois corporation, incorporated in 1900. Abbott's* principal business is the discovery, development, manufacture, and sale of a broad and diversified line of health care products. </FONT></P>
<P ALIGN="CENTER" style="font-family:times;"><FONT SIZE=2><B> FINANCIAL INFORMATION RELATING TO INDUSTRY SEGMENTS, GEOGRAPHIC AREAS, AND CLASSES OF SIMILAR PRODUCTS </B></FONT></P>

<P style="font-family:times;text-align:justify"><FONT SIZE=2>   Incorporated herein by reference is Note 6 entitled "Segment and Geographic Area Information" of the Notes to Consolidated
Financial Statements included under Item 8, "Financial Statements and Supplementary Data" and the sales information related to Humira® included in "Financial Review." </FONT></P>

<P ALIGN="CENTER" style="font-family:times;"><FONT SIZE=2><B> NARRATIVE DESCRIPTION OF BUSINESS </B></FONT></P>

<P style="font-family:times;text-align:justify"><FONT SIZE=2>   Abbott has five reportable revenue segments: Proprietary Pharmaceutical Products, Established Pharmaceutical Products, Diagnostic Products, Nutritional Products, and Vascular Products. </FONT></P>

<P style="font-family:times;text-align:justify"><FONT SIZE=2>   On

October 19, 2011, Abbott announced that it plans to separate into two publicly traded companies, one in diversified medical products and the other in research-based pharmaceuticals. The diversified medical products company will consist of Abbott's existing diversified medical products portfolio, including its branded generic pharmaceutical, devices, diagnostic and nutritional businesses, and will retain the Abbott name. The research-based pharmaceutical company will include Abbott's current portfolio of proprietary pharmaceuticals and biologics and will be named later. </FONT></P>

Extant research

- Majority of <u>large sample</u> analysis of annual report narratives has been conducted on US filings (10-Ks) available via EDGAR
    - Management Discussion and Analysis (MD&A) section (Item 7)
    - Risk-related disclosures (Item 1A and Item 7A)
    - Entire 10-K filing

- Analysis of 10-K filings in EDGAR is relatively straightforward
    - Plain text files with consistent structure
    - Use HTML/XML taggers to identify section(s) and extract text

- UK annual reports pose more significant challenges to researchers
    - Normally supplied as *.pdf
    - Unstructured format → no consistent template

UK Annual report tool: Extraction

- Use contents page to extract text by section from digital pdf

Contents

UK Annual reports: Unstructured format

Contents

Financial highlights:

> Continuing profit before taxation* up 25.3% at £50.1m (2006/07: £40.0m)
> Profit for the period £34.5m (2006/07: loss £22.5m)
> Adjusted EPS[1] 20.9% ahead at 7.88p (2006/07: 6.52p)
> Strong cash management continues to drive low net debt[2] at £200.2m (2006/07: £174.2m)
> Robust balance sheet – continuation of share buy-back programme in 2008/09
> Proposed dividend up 5.9% at 4.50p (2006/07: 4.25p)

UK Annual reports: Unstructured format (cont'd)

UK Annual report tool: Extraction

- Use contents page to extract text by section from digital pdf

- Steps in extraction process:
  - Detect contents page
  - Parse contents page
  - Detect page numbering to determine section start/end
  - Add headers as bookmarks to pdf
  - Extract text for each section

- Analyse extracted text by section and for entire document

Contents

Chairman's Statement	02
Chief Executive's Statement	05
Finance Director's Review	08
Directors	12
Directors' Report	13
Corporate Governance	15
Independent Auditors' Report	20
Consolidated Income Statement	22
Consolidated Balance Sheet	23
Company Balance Sheet	24
Consolidated Statement of Recognised Income and Expense	25
Consolidated Reconciliation of Movements in Equity	25
Company Statement of Recognised Income and Expense	25
Company Reconciliation of Movements in Equity	25
Consolidated Cash Flow Statement	26
Company Cash Flow Statement	26
Notes to the Financial Statements	27
Board Report on Directors' Remuneration	54
Notice of Annual General Meeting	58
Shareholder Information	59
Group Five Year Record	60
Corporate Information	Inside Back Cover

UK Annual report tool: NLP

- In addition to performing text extraction, the tool provides a range of text analysis options:
  - Readability metrics
  - Word counts using pre-determined lists (e.g., forward looking, uncertainty, tone, etc.)
  - Word counts based on user-defined wordlists
  - Comparison with reference corpus (word, parts of speech and semantic level)
  - Concordance and collocates
  - Upload and analyse user-defined text file

UK Annual report tool: NLP

- In addition to performing text extraction, the tool provides a range of text analysis options:
  - Readability metrics
  - Word counts using pre-determined lists (e.g., forward looking, uncertainty, tone, etc.)
  - Word counts based on user-defined wordlists
  - Comparison with reference corpus (word, parts of speech and semantic level)
  - Concordance and collocates
  - Upload and analyse user-defined text file

- Demo to illustrate functionality →

UK Annual report tool: Demo overview

- Overview (wordlists, readability metrics) and interface with WMatrix

- Uploading one or more annual reports and generating output

- Uploading and analysing with a user-defined key word list

- Uploading and analysing a user-defined text file

- Examples of further analysis in MWatrix:
  - Cloud for chairman's statement vs. standard reference corpus: word level
  - Cloud for chairman's statement vs. to standard reference corpus: semantic level
  - Cloud for chairman's statement vs. chairman's statement corpus: word level
  - Cloud for chairman's statement vs. chairman's statement corpus: semantic level
  - Concordance/colocation

UK Annual report tool: Developments

- Within the next month we expect the tool to offer the following:
  - Choice of different readability metrics
  - Text re-use metric to detect boilerplating and incremental changes
  - Reference corpora for:
    - Full Annual Report and Accounts
    - Chairman's Statement
    - CEO Review/Business Review/Operating Review/Financial Review
    - Highlights
    - Corporate Governance
    - Directors' Remuneration
  - Readability scores for all UK annual reports for the period 2003-2013
  - Method for representing key themes (e.g., strategy, risk, CSR and governance, etc.) by section

UK Annual report tool

Questions?

Application: Strategy-related disclosures

- Use tool to study strategy-related content in UK annual reports
  - Final sample comprises approx. 9,500 reports from 2003 to 2012

- Classify report content into following generic sections using synonym lists:
  - Highlights
  - Chairman's Statement
  - CEO's Review
  - Finance Director's Review
  - Other Reviews
  - Strategy and Business Model
  - Directors Report

  - Risk Report
  - CSR Report
  - Governance and Remuneration
  - Auditors Report
  - Accounting Policies
  - Financial Statements
  - Notes to the Accounts
  - Other Statutory Information

Application: Strategy-related disclosures

- Use tool to study strategy-related content in UK annual reports
  - Final sample comprises approx. 9,500 reports from 2003 to 2012

- Classify report content into following generic sections using synonym lists:
  - Highlights
  - Chairman's Statement
  - CEO's Review
  - Finance Director's Review
  - Other Reviews
  - Strategy and Business Model
  - Directors Report
  - Risk Report
  - CSR Report
  - Governance and Remuneration
  - Auditors Report
  - Accounting Policies
  - Financial Statements
  - Notes to the Accounts
  - Other Statutory Information

- Procedure generates initial sample of approx. 200,000 sections
  - Removing non-core narratives and statutory financial statements yields a final sample of approx. 38,000 sections for 7,628 firm-years

Measuring strategic-related content

- Identify 1,156 strategic keywords and phrases from index pages of leading business strategy textbooks plus survey of strategy concept from *SMJ*
  - Porter (1985)
  - Barney & Clark (2007)
  - Rumelt (2011)
  - Magretta (2012)
  - Grant (2013)
  - Ronda-Pupol & Guerras-Martin (2012 *SMJ*)

- Disambiguation and semantic level analysis:
  - Use Bayes Rule to compute the probability that a keyword is used in the context of a discussion about strategy
  - Weight keyword occurrences by its Baysian probability and then sum weighted keywords to produce strategic content score for each firm-year

18 most important words

growth	assets
customers	focus
development	risk
markets	acquisition
strategy/strategic	activities
product	employees
opportunities	position
profit	demand
technology	resources

Results

- Strategy-related content is sticky over time
  - Lagged strategy score explains ≈ 60% of cross-sectional variation

- Strategy-related content:
  - Varies across industry and increases over time
  - Increases in firm size, MTB, financing needs, profit variability, and analysts
  - Decreases in asset growth, relative performance, past performance, and restructuring costs

- One-year change in strategy-related content is negatively related to poor past performance and industry disclosure norms

- Question: What are the consequences of providing more information on strategy?

Summary

- Automated large sample analysis of non-US annual reports is hampered by document format and structure

- Develop a free-to-access software tool for extracting and analysing UK annual report narratives (plus text from other sources)
  - Extracts from pdfs by section
  - Provides simple NLP metrics such as readability, word counts, etc.
  - Interfaces with MWatrix for more sophisticated linguistic analysis at parts of speech and semantic levels

- Functionality of software expected to develop significantly over coming months

- Using methods to examine a series of disclosure-related questions