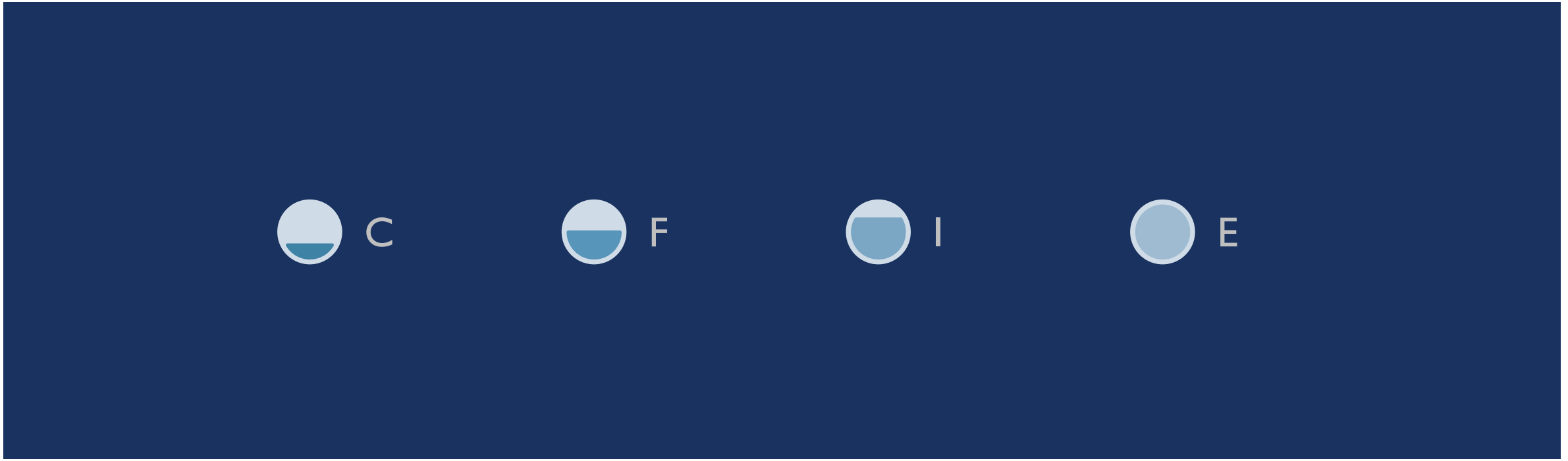
CORPORATE FINANCIAL INFORMATION ENVIRONMENT (CFIE)

MAHMOUD EL-HAJ

SCC, LANCASTER UNIVERSITY



TEAM AND FUNDING

Prof Martin Walker
MBS

Prof Steve Young
LUMS

Dr Paul Rayson
SCC

Dr Mahmoud El-Haj
SCC

Dr Vasiliki Athanasakou
LSE

Dr Thomas Schleicher
MBS

CFIE?

- Primary ways that firms communicate with capital market participants.
- Together with information from:
  - analysts,
  - financial journalists,
  - rating agencies and
  - other market commentators that are external to the firm
- combine to form the Corporate Financial Information Environment (CFIE)

PURPOSE AND MOTIVATION

- study the causes and consequences of corporate disclosure and financial reporting outcomes.
- aim to uncover the determinants of financial reporting quality
- and the factors that influence the quality of information disclosed to investors beyond the financial statements.

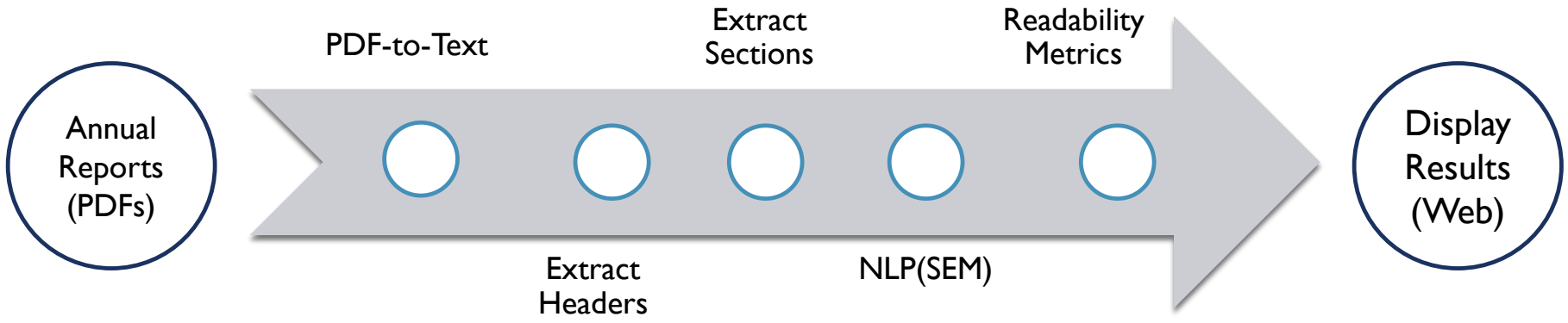
AIM OF WORK PRESENTED

- we aim to scale up the application of current readability metrics and improve their granularity.
- to improve on previous work, we need to apply the metrics to individual sections of firms' annual reports.
- a necessary prerequisite for our work is to automatically determine the structure of these reports.

DATASET

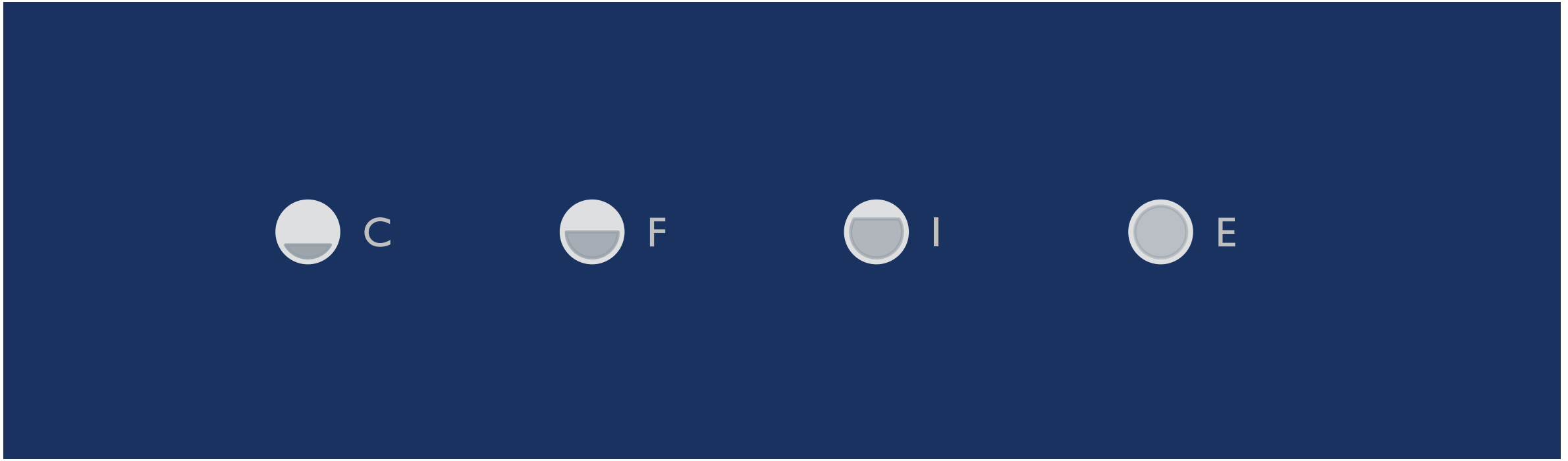
- 1,500 searchable financial annual reports
- of around 200 of the largest UK firms listed on the LSE
- with an average of 7 annual reports for each firm
- between the years 2003 and 2012.

CFIE ANALYSIS PIPELINE



APPLY NLP METHODS USED IN PRIOR US STUDIES TO UK ANNUAL REPORTS?

UK VS. US FILINGS



US FILINGS

- US companies must submit:

1. 10-K: Annual
2. 10-Q: Quarterly
3. 8-K: Special Events (between 10-K and 10-Q)
4. Annual Report

10-K ANNUAL FORM

Each 10-K contains 4 parts and 15 items

- **PART I**
- **ITEM 1.** Description of Business
- **ITEM 2.** Description of Properties
- **ITEM 3.** Legal Proceedings
- **ITEM 4.** Mine Safety Disclosures
- **PART II**
- **ITEM 5.** Market for Registrant's Common Equity….
- **ITEM 6.** Selected Financial Data
- **ITEM 7.** Management's Discussion and Analysis….
- **ITEM 8.** Financial Statements and Supplementary Data
- **ITEM 9.** Changes in and Disagreements ….
- **PART III**
- **ITEM 10.** Directors, Executive Officers and Corporate Governance
- **ITEM 11.** Executive Compensation
- **ITEM 12.** Security Ownership of Certain Beneficial Owners….
- **ITEM 13.** Certain Relationships and Related Transactions….
- **ITEM 14.** Principal Accounting Fees and Services
- **PART IV**
- **ITEM 15.** Exhibits, Financial Statement Schedules….

10-K ANNUAL (STARBUCKS VS. MCDONALD'S)

Starbucks Corporation

McDONALD'S CORPORATION

UK ANNUAL REPORTS

- Free style (no standard structure)
- Use of images, text, hyperlinks, …etc.
- PDF format

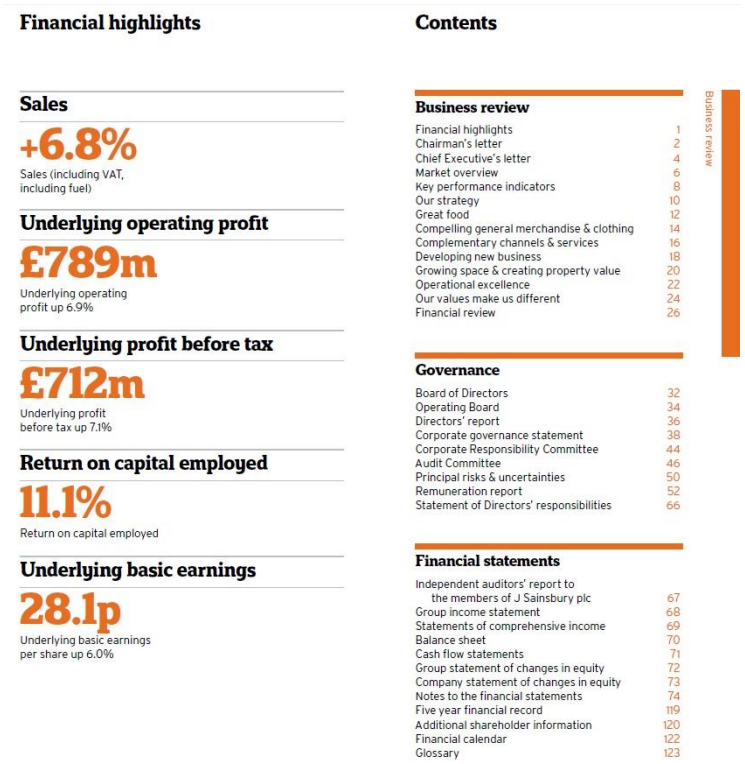
UK ANNUAL REPORT SAMPLES

- Content and structure varies across firms.
- Management have more discretion over what, where, and how much information on topics such as risk, strategy, performance, etc. is reported.

This makes the extraction and analysis task more challenging;

but it provides research opportunities.

UK ANNUAL REPORTS SAMPLE



Financial highlights

Sales

+6.8%

Sales (including VAT, including fuel)

Underlying operating profit

£789m

Underlying operating profit up 6.9%

Underlying profit before tax

£712m

Underlying profit before tax up 7.1%

Return on capital employed

11.1%

Return on capital employed

Underlying basic earnings

28.1p

Underlying basic earnings per share up 6.0%

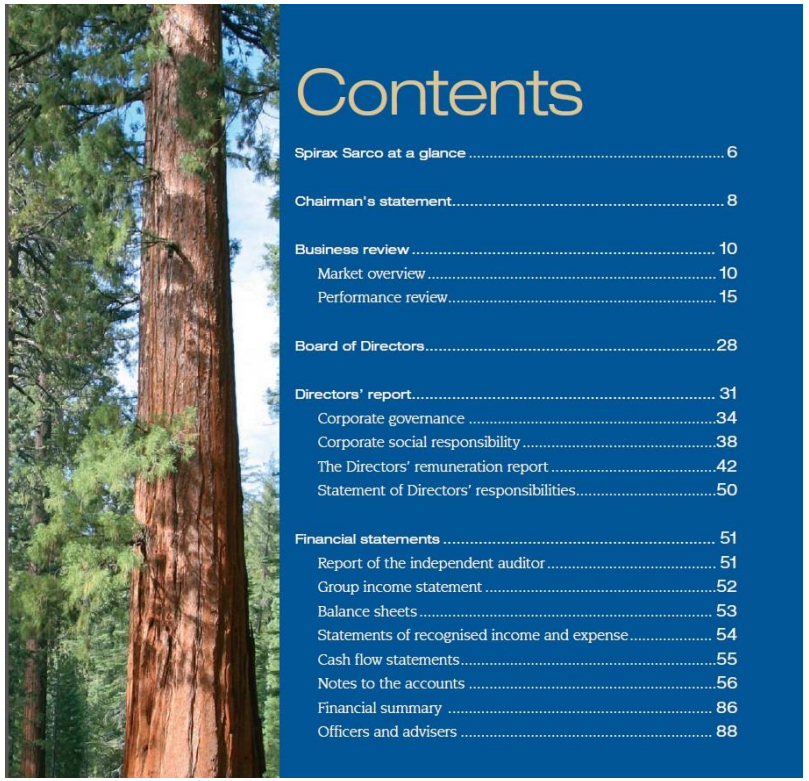
Contents

Business review

Business review

Governance

Financial statements

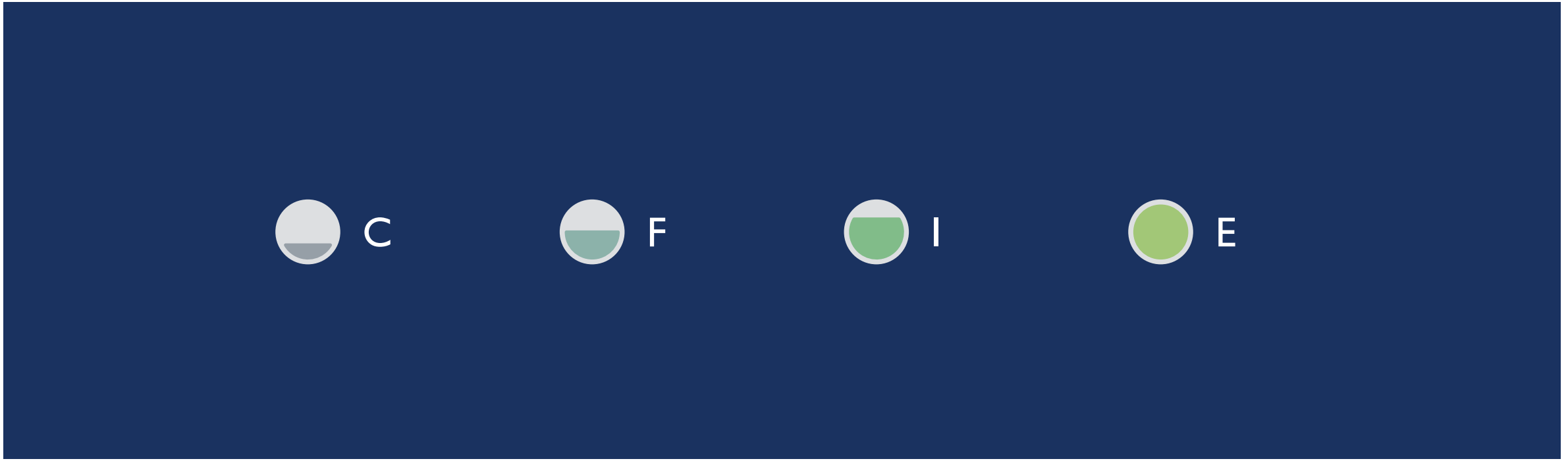


Contents

ARRIVA

EXTRACTION PROCESS

WHAT ARE WE LOOKING TO EXTRACT?

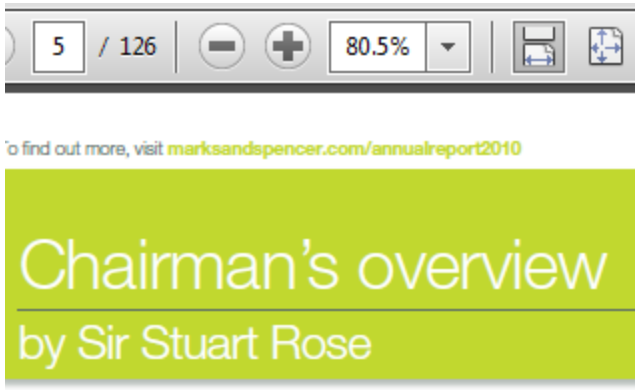


HEADERS AND THEIR SECTIONS

- We are looking to extract the following headers and their narratives for further processing:
    1. Chairman's statement
    2. CEO Review
    3. Corporate Government Report
    4. Directors Remuneration Report
    5. Directors Report and Business Review
    6. Directors Responsibilities Statement
    7. Directors Report
    8. Financial Review
    9. Key Performance Indicator
    10. Operational Review
    11. Highlights

HOW?

Not consistent across ARs



Doesn't always refer to the correct page

EXTRACTION STEPS

- 1) detecting the contents-page
- 2) parsing the detected contents-page and extracting the headers
- 3) detecting page numbering
- 4) adding the extracted headers to the annual report PDFs as bookmarks
- 5) using the added bookmarks to extract the narrative sections under each heading

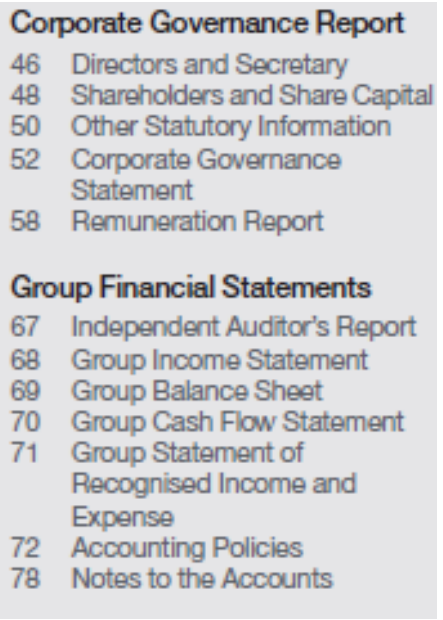
- The processes run on searchable (text-based) PDFs; we will consider using OCR techniques to process non-searchable (scanned) PDFs in a later stage.

1) DETECTING THE CONTENTS PAGE

- created a list of gold–standard section names extracted manually from a random sample of 50 annual reports
- matched each page in the annual report against the gold-standard list
- selected the page with the highest matching score as the potential contents page
- the score was calculated by an increment of 1 for each match.
- To improve the matching process and avoid false positives, we match the gold–standard keywords against lines of text that follow a contents-page-like style (e.g. section name followed by page number, such as *Chairman's Statement 13*).

2) PARSING THE CONTENTS PAGE

- We automatically parsed the detected contents page to extract section names and their associated pages
- matched each line of text in the potential contents page against a regular expression command that will extract any line starting or ending with a number between 1 and the number of pages of the annual report.
- We differentiate between dates and actual page numbers to avoid extracting incorrect section headers.
- However, lines containing text such as an address (e.g., 77 London Road) might still be confused.
- We tackled this problem by matching the list of extracted headers against a list of gold–standard header synonyms.
- To tackle the problem of broken headers we concatenating sentences that end or begin with prepositions such as 'of', 'in' ...etc.
- The algorithm also concatenates sentences ending with singular or plural possessives, symbolic and textual connectors (e.g. 'and', 'or', '&'...etc), and sentences ending with hyphenations.



Corporate Governance Report

46 Directors and Secretary
48 Shareholders and Share Capital
50 Other Statutory Information
52 Corporate Governance Statement
58 Remuneration Report

Group Financial Statements

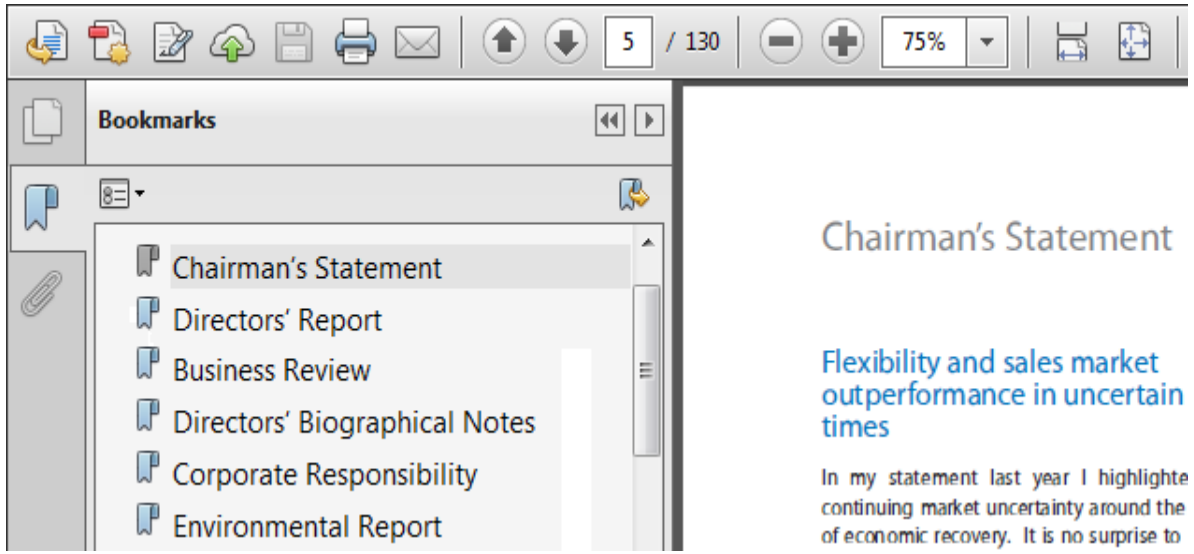
67 Independent Auditor's Report
68 Group Income Statement
69 Group Balance Sheet
70 Group Cash Flow Statement
71 Group Statement of Recognised Income and Expense
72 Accounting Policies
78 Notes to the Accounts

3) DETECTING PAGE NUMBERING

- The page numbers appearing on the contents page do not usually match with the actual page numbers in the pdf files.
- Created a simple page detection tool that crawls through a dynamic number of three consecutive pages with the aim of extracting a pattern of sequential numbers with an increment of 1 (e.g. 31, 32, 33).
- Running this process we got an accuracy rate of 94%.
- Manual examination of the remaining 6% revealed the following reasons for non-detection: 1) encoding, 2) formatting and 3) design.

4) ADDING HEADERS AS BOOKMARKS

- Using the headers and their correct page numbers we implemented a tool to insert the extracted contents page headers as bookmarks (hyperlinks) to sample PDFs.
- This process helped in extracting narratives associated with each header for further processing



5) EXTRACTING HEADERS' NARRATIVES (PART 1)

- Automatically crawl through the data collection and extract all inserted bookmarks and their associated pages.
- Since UK firms do not follow a standard format when creating annual reports, a long list of synonyms are possible for a single header.
- For example the header "Chairman's Statement" may also appear as "Chairman's Introduction", "Chairman's Report" or "Letter to Shareholders".
- To solve this problem we, semi automatically and by the help of an expert in accounting and finance, created a list of synonyms for each of the 11 generic annual report headers.
- This was done by extracting all headers containing "Chairman", "Introduction", "Statement", "Letter to"...etc from a sample of 250 annual reports of 50 UK firms (the quoted unigrams were selected by the same expert).
- We refined the list by removing redundancies. The accounting expert then manually examined the list and deleted irrelevant or inappropriate headers.
- We used the refined list as gold–standard synonyms to extract all the headers related to each of our generic headers

5) EXTRACTING HEADERS' NARRATIVES (PART 2)

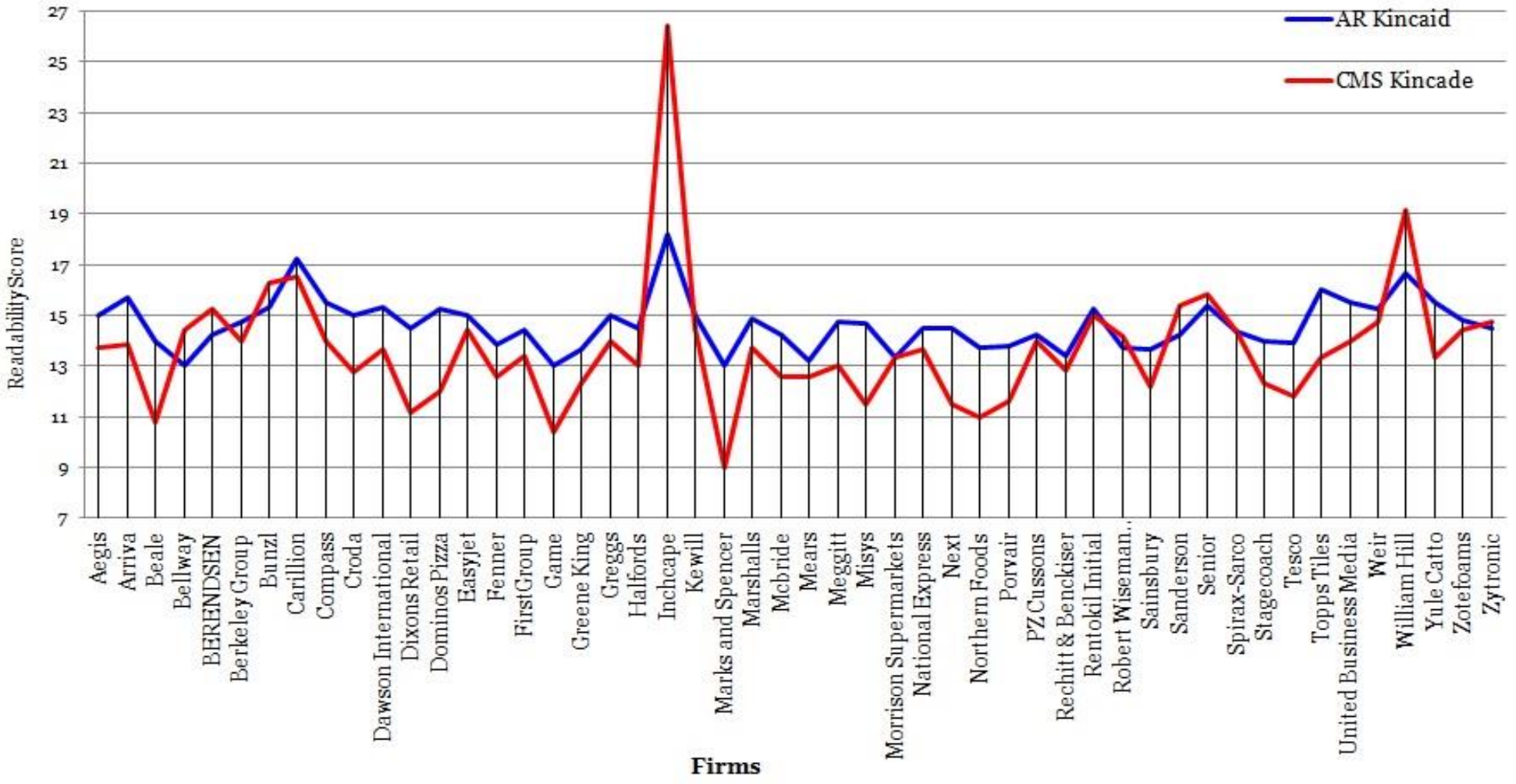
- To tackle different word–order or additional words included in the headline (e.g. "The Statement of the Chairman") we used Levenshtein Distance string metric algorithm to measure the difference between two headers.
- The Levenshtein distance between two words is the minimum number of single-character edits (insertion, deletion, substitution) required to change one word into the other.
- To work on a sentence level we modified the algorithm to deal with words instead of characters.
- All the headers with a Levenshtein distance of up to five were presented to the accounting expert.

1. Chairman's statement
2. CEO Review
3. Corporate Government Report
4. Directors Remuneration Report
5. Directors Report and Business Review
6. Directors Responsibilities Statement
7. Directors Report
8. Financial Review
9. Key Performance Indicator
10. Operational Review
11. Highlights

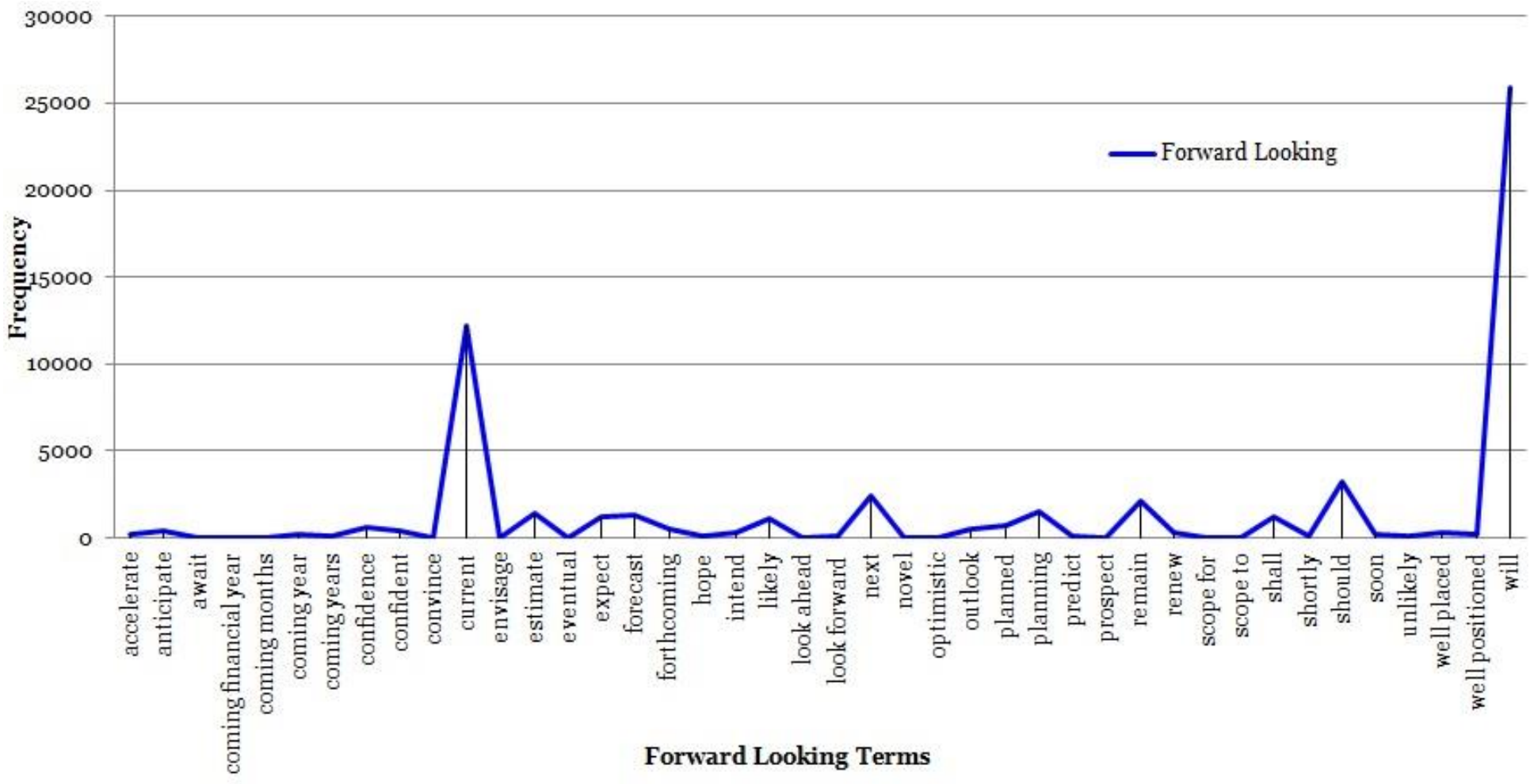
ANALYSIS AND READABILITY MEASURES

- For a sample of 250 annual reports we analysed each report and its extracted sections by calculating text readability scores using Flesh and Fog readability measures.
- We also counted word frequencies using forward looking, hedging, positive and negative words–lists.

READABILITY: ANNUAL REPORT VS CHAIRMAN'S SECTION



FORWARD LOOKING FREQUENCIES



EVALUATION

- To ensure quality, we used domain experts to judge the quality of the document structure extraction process.
- We took a random sample of 100 previously unseen annual reports that had bookmarks automatically added to them through the extraction process.
- The expert human evaluators were presented with an evaluation form and asked to compare the automatically assigned bookmarks to the contents page of the same annual report.
- An expert in the accounting and finance domain went through the extracted headers and their narrative sections to judge the quality of the extraction process, the expert also updated the gold–standard list with any new unseen synonyms.

Document Name	Year	Number of Headers in PDF	Number of Extracted Headers	Number of Exact Matches	Number of Partial Matches	Number of Wrong Headers	Page Numbers Correct?	If NO, what is the difference between PDF and Report page numbers?	Notes:
2 ERGO 31AUG04	2004	31	24	24	0	0	Yes		
3I GROUP PLC_07	2007	40	39	38	0	1	Yes		Picked a footer
ACAL PLC-09	2009	24	24	22	2	0	No	2	

EVALUATION RESULTS

- The evaluators' input was used to calculate Recall/Precision and F measure.
- The manual evaluation was performed in two separate stages following the same evaluation process.
- Stage 1 helped identify the most common errors that led to incorrect extraction and detection of either the contents page and its headers or the annual report's page numbering.
- Stage 2 was performed after fixing errors discovered by the human evaluators.

EVALUATION: STAGE 1 AND 2

	Stage 1		Stage 2	
	Count	Percent	Count	Percent
# of PDFs	105	-	105	-
Headers in PDFs	2473	-	2473	-
Extracted Headers	2479	-	2502	-
Exact Matches	2101	84.8%	2202	88.01%
Partial Matches	189	7.6%	105	4.20%
Wrong Headers	189	7.6%	195	7.8%
Missing Headers	183	7.4%	166	6.6%
Correct Headers	2290	92.6%	2307	93.3%
Detected Page number	80	76.2%	94	89.5%
Detected Contents Pages	97	92.4%	97	92.4%

RECALL/PRECISION AND F MEASURE

- An extracted header is considered 'strictly relevant' only if it is an exact match of a PDF's header.
- The header is considered 'broadly relevant' if it is either an exact match or a partial match of a PDFs header.
- Results reveal the fixes applied helped increase recall and precision rates by extracting more relevant headers.

	Stage 1	Stage 2
Strict Recall	0.8496	0.8904
Strict Precision	0.8475	0.8801
Strict F-1 Score	0.8485	**0.8852**
Broad Recall	0.9260	0.9329
Broad Precision	0.9238	0.9221
Broad F-1 Score	0.9249	**0.9274**

THANKS
ANY QUESTIONS

CFIE UREL: http://ucrel.lancs.ac.uk/cfie