

Problems and Prospects for Natural Language Processing of UK Corporate Narrative Disclosures.

Mahmoud El-Haj
Lancaster University

Objectives

- Apply NLP methods used in prior US studies to UK annual reports.
- Tools options, and what seems most promising for our purposes.
- Get control over the headings and detailed content of digital PDF files.
- Early highlights from our sample of 250 reports.

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

	PART I
Item 1	Business
Item 1A	Risk Factors
Item 1B	Unresolved Staff Comments
Item 2	Properties
Item 3	Legal Proceedings
Item 4	(Removed and Reserved)
	PART II
Item 5	Market for the Registrant's C
Item 6	Selected Financial Data
Item 7	Management's Discussion an
Item 7A	Quantitative and Qualitative I
Item 8	Financial Statements and Sup
	Report of Independent Regist
Item 9	Changes in and Disagreemen
Item 9A	Controls and Procedures
Item 9B	Other Information
	PART III
Item 10	Directors, Executive Officers
Item 11	Executive Compensation
Item 12	Security Ownership of Certai
Item 13	Certain Relationships and Re
Item 14	Principal Accountant Fees an
	PART IV
Item 15	Exhibits and Financial Stater

	Part I.
Item 1	Business
Item 1A	Risk Factors and Cautionary
Item 1B	Unresolved Staff Comments
Item 2	Properties
Item 3	Legal Proceedings
Item 4	Mine Safety Disclosures
	Part II.
Item 5	Market for Registrant's Comn
Item 6	Selected Financial Data
Item 7	Management's Discussion ar
Item 7A	Quantitative and Qualitative I
Item 8	Financial Statements and Su
Item 9	Changes in and Disagreemer
Item 9A	Controls and Procedures
Item 9B	Other Information
	Part III.
Item 10	Directors, Executive Officers.
Item 11	Executive Compensation
Item 12	Security Ownership of Certain
Item 13	Certain Relationships and Re
Item 14	Principal Accountant Fees an
	Part IV.
Item 15	Exhibits and Financial Stater

UK Annual Reports

-
- Free Style (no standard structure)
 - Use of Images, Text
 - 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.

Annual Report (sample 1)

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

Financial highlights	1
Chairman's letter	2
Chief Executive's letter	4
Market overview	6
Key performance indicators	8
Our strategy	10
Great food	12
Compelling general merchandise & clothing	14
Complementary channels & services	16
Developing new business	18
Growing space & creating property value	20
Operational excellence	22
Our values make us different	24
Financial review	26

Business review

Governance

Board of Directors	32
Operating Board	34
Directors' report	36
Corporate governance statement	38
Corporate Responsibility Committee	44
Audit Committee	46
Principal risks & uncertainties	50
Remuneration report	52
Statement of Directors' responsibilities	66

Financial statements

Independent auditors' report to the members of J Sainsbury plc	67
Group income statement	68
Statements of comprehensive income	69
Balance sheet	70
Cash flow statements	71
Group statement of changes in equity	72
Company statement of changes in equity	73
Notes to the financial statements	74
Five year financial record	119
Additional shareholder information	120
Financial calendar	122
Glossary	123

Annual Report (sample 2)



Contents

Spirax Sarco at a glance	6
Chairman's statement.....	8
Business review	10
Market overview	10
Performance review.....	15
Board of Directors	28
Directors' report	31
Corporate governance	34
Corporate social responsibility	38
The Directors' remuneration report	42
Statement of Directors' responsibilities.....	50
Financial statements	51
Report of the independent auditor	51
Group income statement	52
Balance sheets	53
Statements of recognised income and expense.....	54
Cash flow statements.....	55
Notes to the accounts	56
Financial summary	86
Officers and advisers	88

Annual Report (sample 3)



02	Who we are and what we do	42	Corporate governance
06	24 hours in the life of Arriva	46	Statement of directors' responsibilities
08	Our growth story	47	Independent auditors' report on the group financial statements
10	Our markets	48	Financial statements
12	Chairman's statement	52	Accounting policies
14	Chief executive's review	56	Notes to the accounts
22	Financial review	82	Five-year financial summary
26	Corporate responsibility	83	Parent company financial statements
32	Board of directors	90	Statement of directors' responsibilities on the parent company financial statements
34	Directors' report	91	Independent auditors' report on the parent company financial statements
37	Directors' remuneration report	92	Financial calendar, registered office and advisors

Annual Report (sample 4)

1

Principal risks and uncertainties

There are risks and uncertainties which could impact the Group's long identify, manage and mitigate business risk. The table below gives examples of risks and uncertainties identified. The Board considers that these are the most significant risks and uncertainties that the Group faces. These risks do not comprise all those associated with Marks & Spencer and are not intended to be an exhaustive list. There are risks and uncertainties which are not presently known to management, or currently deemed to be less significant.

Risk	Impact
Economic downturn	
Our current priorities place a greater emphasis on managing our business for the short term and continuing to invest for the long term, to be well placed when the economy recovers.	
Strategy	
We fail to set the strategic direction to balance short-term and long-term profitability	Adverse effect on financial results
Finance	
We fail to protect brand and profitable revenues whilst driving cost savings	Adverse effect on financial performance and brand reputation

2

Principal risks and uncertainties

In addition to the opportunities we have to grow and develop our business, the Group faces a range of risks and uncertainties as part of both its day to day operations and its corporate activities. The processes that the Board has established to safeguard both shareholder value and the assets of the Group are described in the Corporate Governance Report.

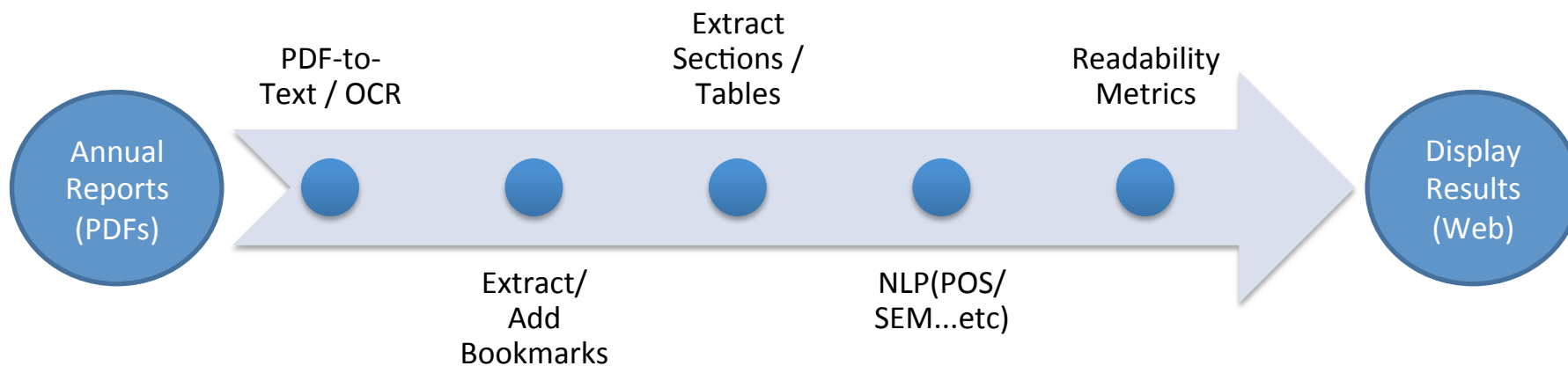
The narrative below describes those specific risks and uncertainties that the Directors believe could have the most significant impact on the Group's long term value generation. The risks and uncertainties described below are not intended to be an exhaustive list.

Risks inherent in bidding for contracts

One of the principal methods of increasing shareholder value is winning new contracts. Inherent in bidding for new contracts is a risk that assumptions are made in the bid model that turn out to be undeliverable for any number of

Tools options, and what seems most promising for our purposes

NLP and Readability Tools



Extracting Text from PDFs

- Open Source
 - XPDF
 - PDF-to-HTML
 - LA-PDF-Text
 - Apache PDFBox
 - Improved Apache PDFBox
 - Multivalent
 - iText ✓
 - PDFTextStream ✓
- Commercial
 - Adobe X Pro ✓
 - OmniPage

Readability Measures and other Dimensions

- Used Phantom, an open source Java library.
 - Flesch-Kincaid
 - Gunning Fog Index
- Other Dimensions (word-lists):
 - Forward Looking
 - Hedges
 - Tone (Positive, Negative)
 - Uncertainty
 - Challenges

Get control over the headings and detailed content of digital PDF files.

Sections

- We are focusing on extracting the following sections:
 - **Annual Report**
 1. Chairman’s statement
 2. Directors Report
 3. CEO Review
 4. Corporate Government Report
 5. Directors Remuneration Report
 6. Directors Report and Business Review
 7. Directors Responsibilities Statement
 8. Financial Review
 9. Key Performance Indicator
 10. Operating Review
 11. Highlights
 - **Financial Statement**
 1. Independent auditors' report.
 2. Mandatory financial statements (including statement of financial position, cash flow statement)
 3. Other information (including information for shareholders...etc.)

Pre-processing

- To ease the matching process, the headers had to undergo the following steps (using Regular Expressions):
 1. Convert all headers to lowercase.
 2. Remove punctuations (e.g. single-quote).
 3. Get the singular or the plural form of headers text (e.g. “director”, “chairman”, “responsibility” – instead of “directors”, “chairman’s”, “responsibilities”).

Levenshtein Distance

- is a string metric for measuring the difference between two sequences.
- we modified Levenshtein distance algorithm to work on a word level instead of character level.

This transforms the algorithm's meaning to:

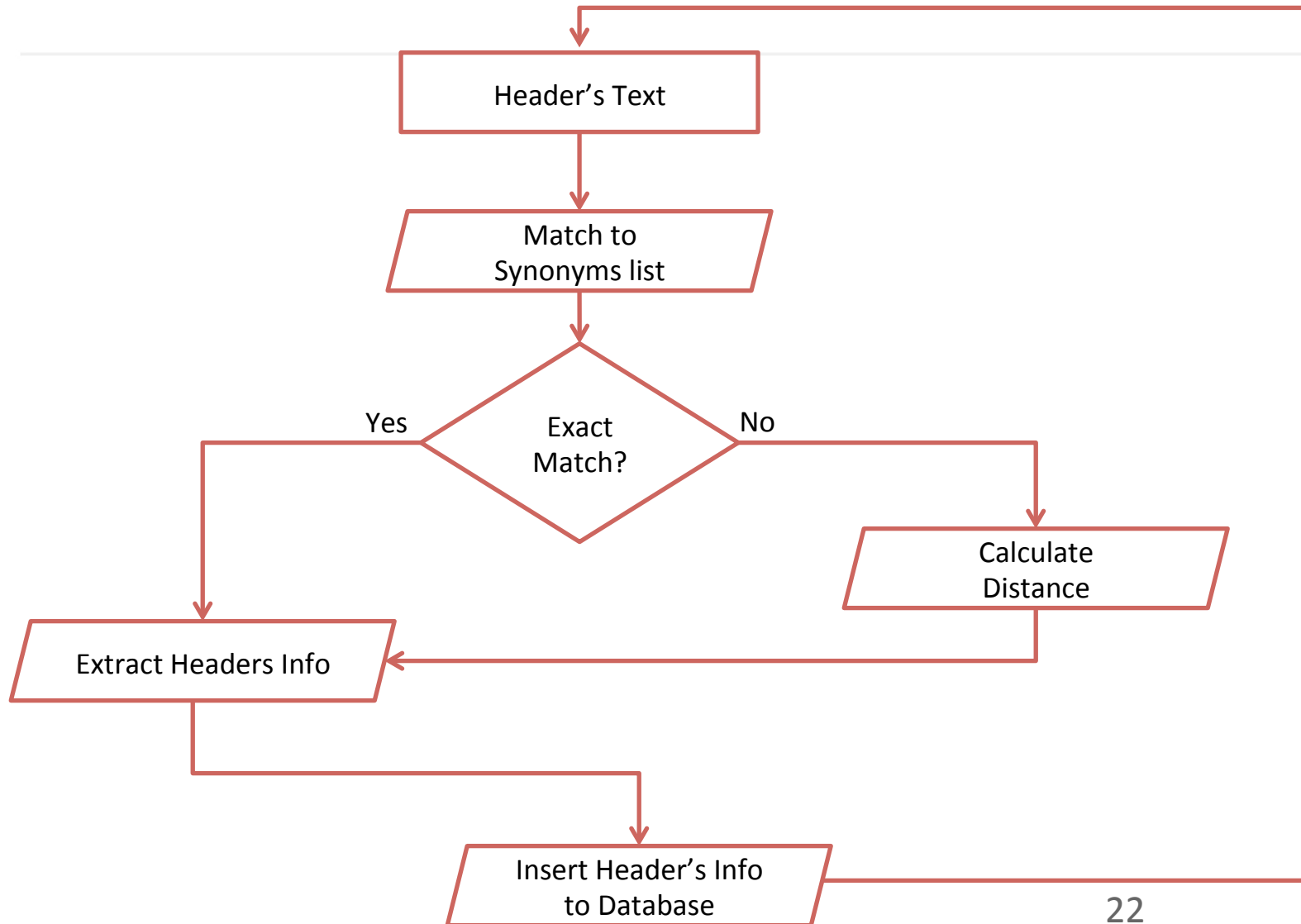
- the minimum number of word edits (insertion, deletion, substitution) required to change one sentence into the other.

Extraction Process

Example: “Chairman’s statement”

1. Select all headers containing any of the chairman header’s words (e.g. “Chair”, “statement”, “introduction”, “letter to”..etc).
2. Provide the list to Accounting and Finance expert to create gold-standards (synonyms list).
3. Insert the headers & their info into a database-table.
4. Run the selected headers through a decision-making (keep one chairman’s header per firm).

Extracting Headers Diagram



Matching Statistics

Show matching statistics of a selected number of headers.

Chairman's Statement

- List of Synonyms:
 - chairman and chief executive introduction
 - chairman and chief executive statement
 - chairman foreword
 - chairman introduction
 - chairman letter
 - chairman review
 - chairman statement
 - executive chairman statement
 - letter to shareholder
 - statement by the chairman and the chief executive
- Total Matches: 237
- Exact Matches: 229
- Partial Matches: 8
 - chairman statement by lord smith of kelvin
 - chairman statement on corporate governance

Corporate Governance Report

- List of Synonyms:
 - corporate governance report
 - corporate governance statement
 - corporate governance
 - director statement on corporate governance
 - corporate governance review
 - report on corporate governance
 - director report on corporate governance
 - statement on corporate governance
- Total Matches: 232
- Exact Matches: 232
- Partial Matches: 0

Directors Remuneration Report



- List of Synonyms:
 - director remuneration report
 - director remuneration statement
 - board remuneration report
 - board report on director remuneration
 - director remuneration
 - director report on remuneration
 - director report remuneration
 - remuneration committee report
 - remuneration report
 - report of the board on director remuneration
 - report of the remuneration committee
 - report on director remuneration
 - the director remuneration report
- Total Matches: 238
- Exact Matches: 237
- Partial Matches: 1 (policy on executive director remuneration)

Directors Report

- List of Synonyms:
 - Report of the director
 - Director Report
- Total Matches: 232
- Exact Matches: 211
- Partial Matches: 21
- Errors: 23/232
- Exact Match Errors: 3/211
- Partial Match Errors: 20/21

Error Analysis

Running the above on the 250 document set we found the following cases where a document:

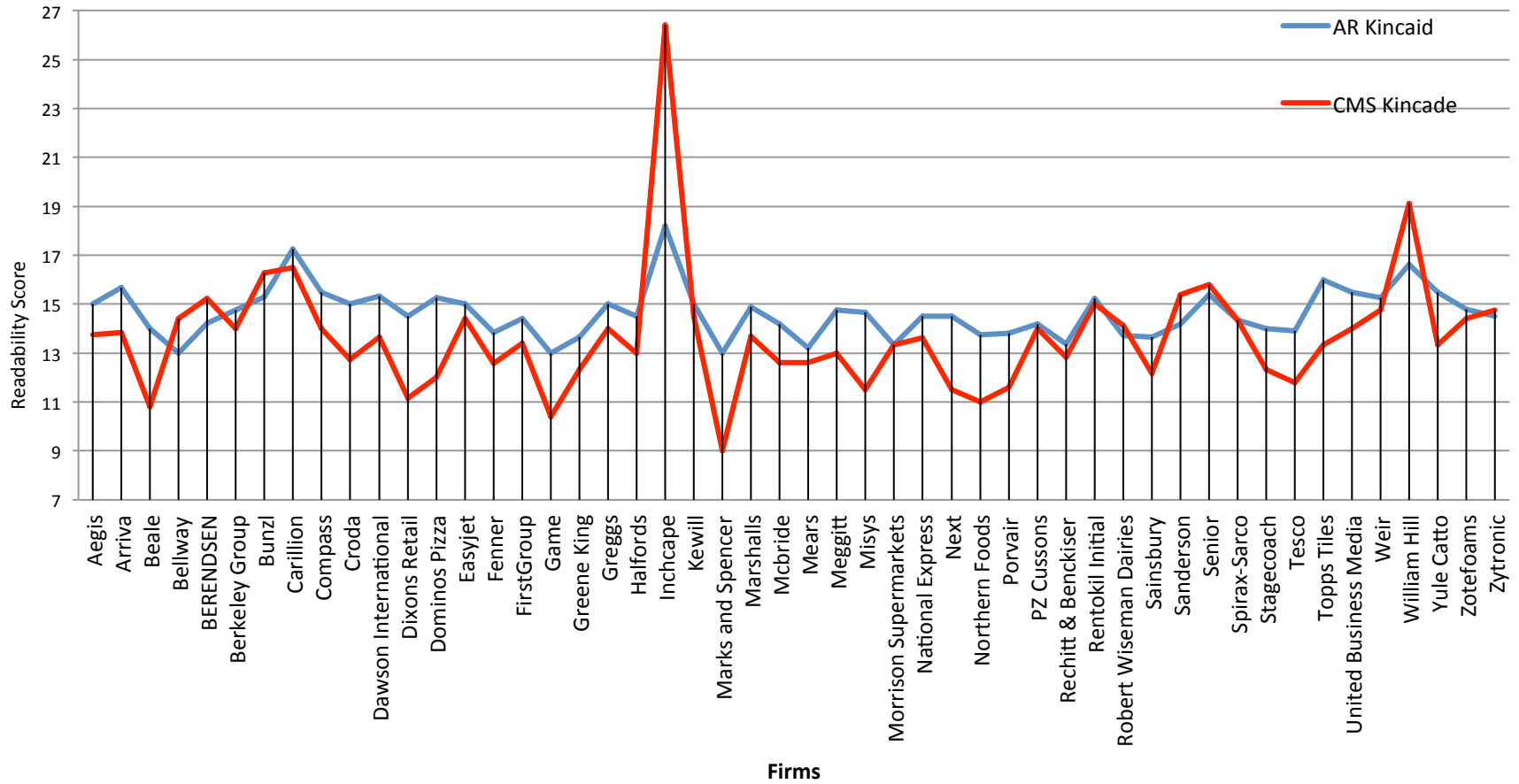
- does not contain bookmarks.
- bookmarks available are incorrect/non-textual.
- does not contain any the header and their synonyms.

Early highlights from our sample of 250 reports.

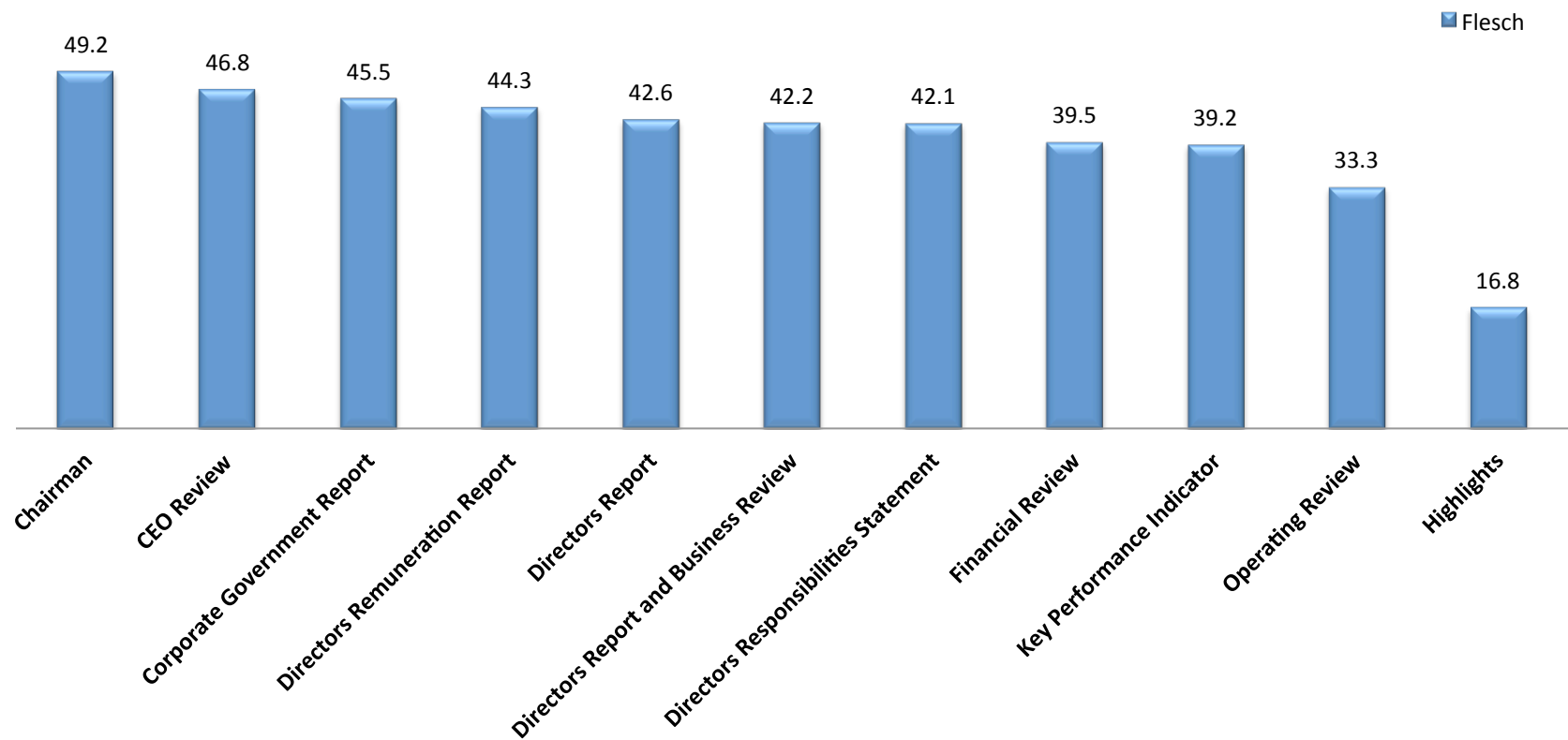
Readability Measure Charts

- Annual Reports (AR) Sample:
 - 250 Searchable PDFs.
 - 50 Firms.
 - 5 ARs over 5 years.
 - 2005-2012.

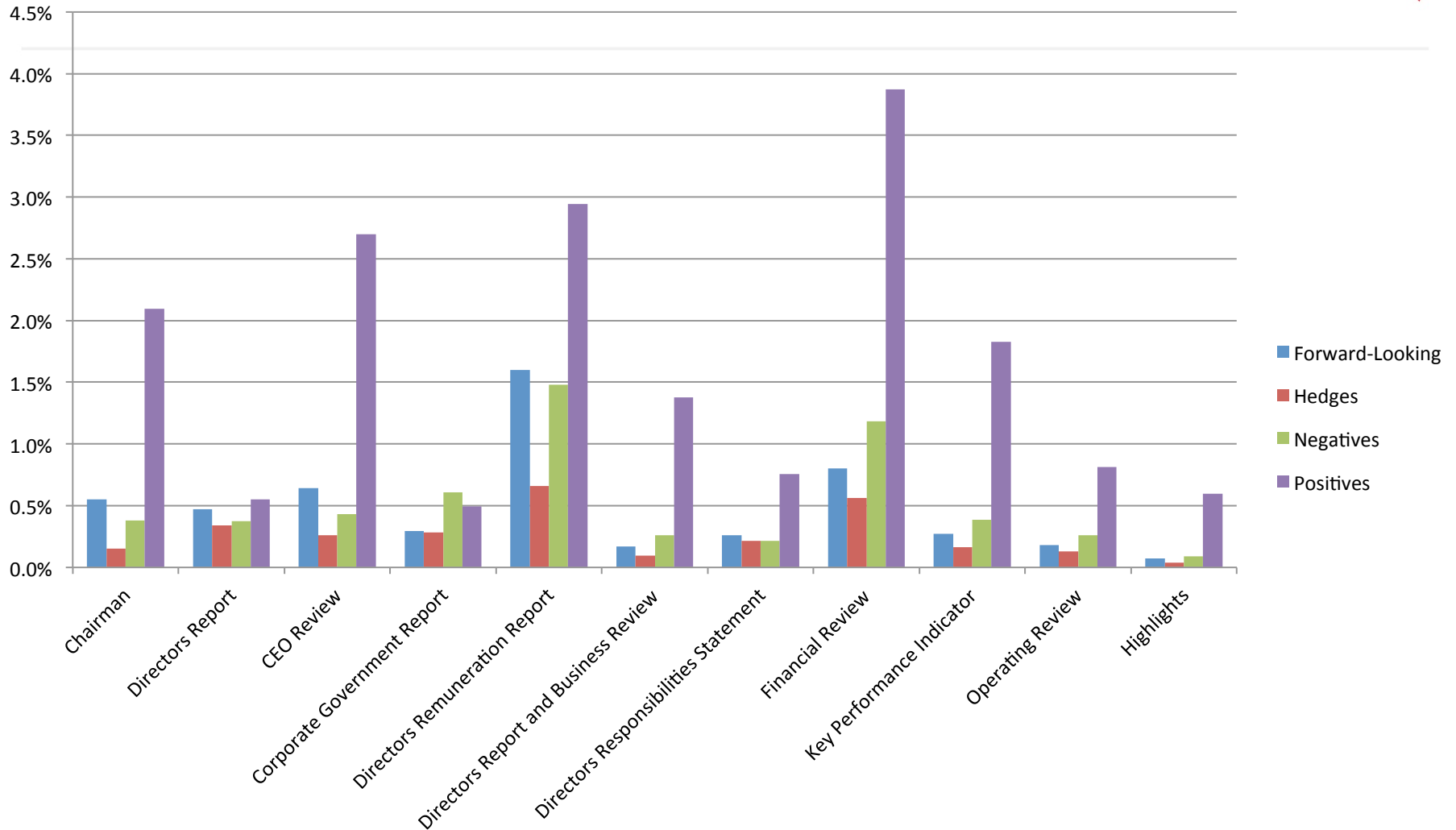
Annual Report vs. Chairman's Statement



Sections Readability Measure



Wordlist Frequencies Across Sections



Challenges



Challenges

- Rely on bookmarks provided by PI
- Idiosyncrasy
 - Sub-headers (directors report, Key performance Indicator) .
- Alternatives:
 - Parse and detect content page.
 - Difficulties
 - Not standard
 - Hard to detect section's end page.
 - PDF vs. AR page numbering
 - Progress

Thank You...

Questions?

CFIE Project:

<http://ucrel.lancs.ac.uk/cfie/>

email:

m.el-haj@lancaster.ac.uk