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

Mahmoud El-Haj
Lancaster University

Objectives

LANCASTER UNIVERSITY

- 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)

LANCASTER UNIVERSITY

Starbucks Corporation

McDONALD'S CORPORATION

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

Annual Report (sample 2)

Contents

Annual Report (sample 3)

ARRIVA

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 exa
and uncertainties identified. The Board considers that these are the m
do not comprise all those associated with Marks & Spencer and are n
not presently known to management, or currently deemed to be less r

Risk	Impact
Economic downturn Our current priorities place a greater emphasis on managing our busir and continuing to invest for the long term, to be well placed when the	
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

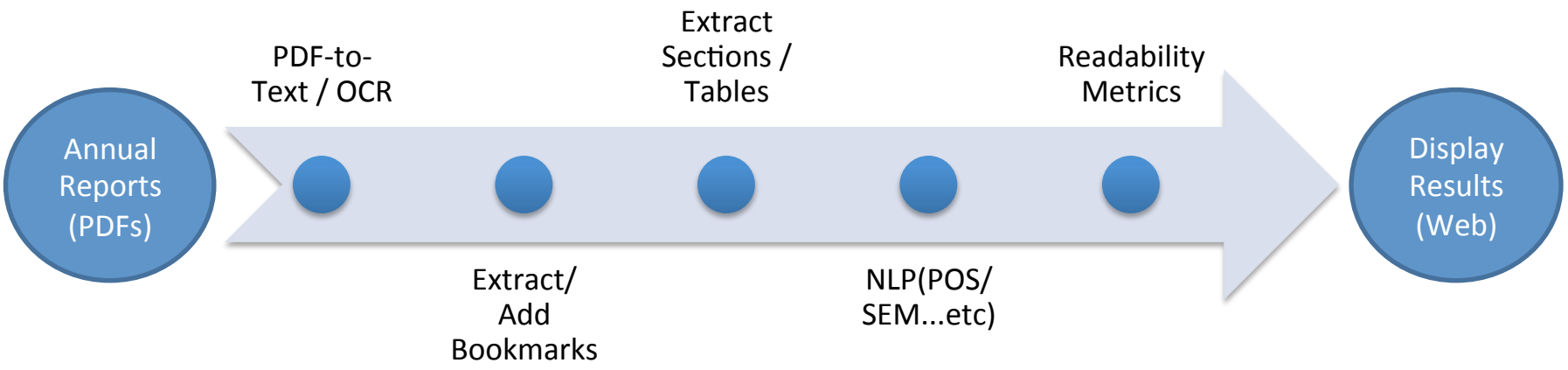
Principal risks and uncertainties
In addition to the opportunities we have to grow and develop our business, the Group faces a range of risks and uncertainties as part of both its day to day operations and its corporate activities. The processes that the Board has established to safeguard both shareholder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

LANCASTER UNIVERSITY

- 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

LANCASTER UNIVERSITY

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

LANCASTER UNIVERSITY

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

LANCASTER UNIVERSITY

- 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

LANCASTER UNIVERSITY

- 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:

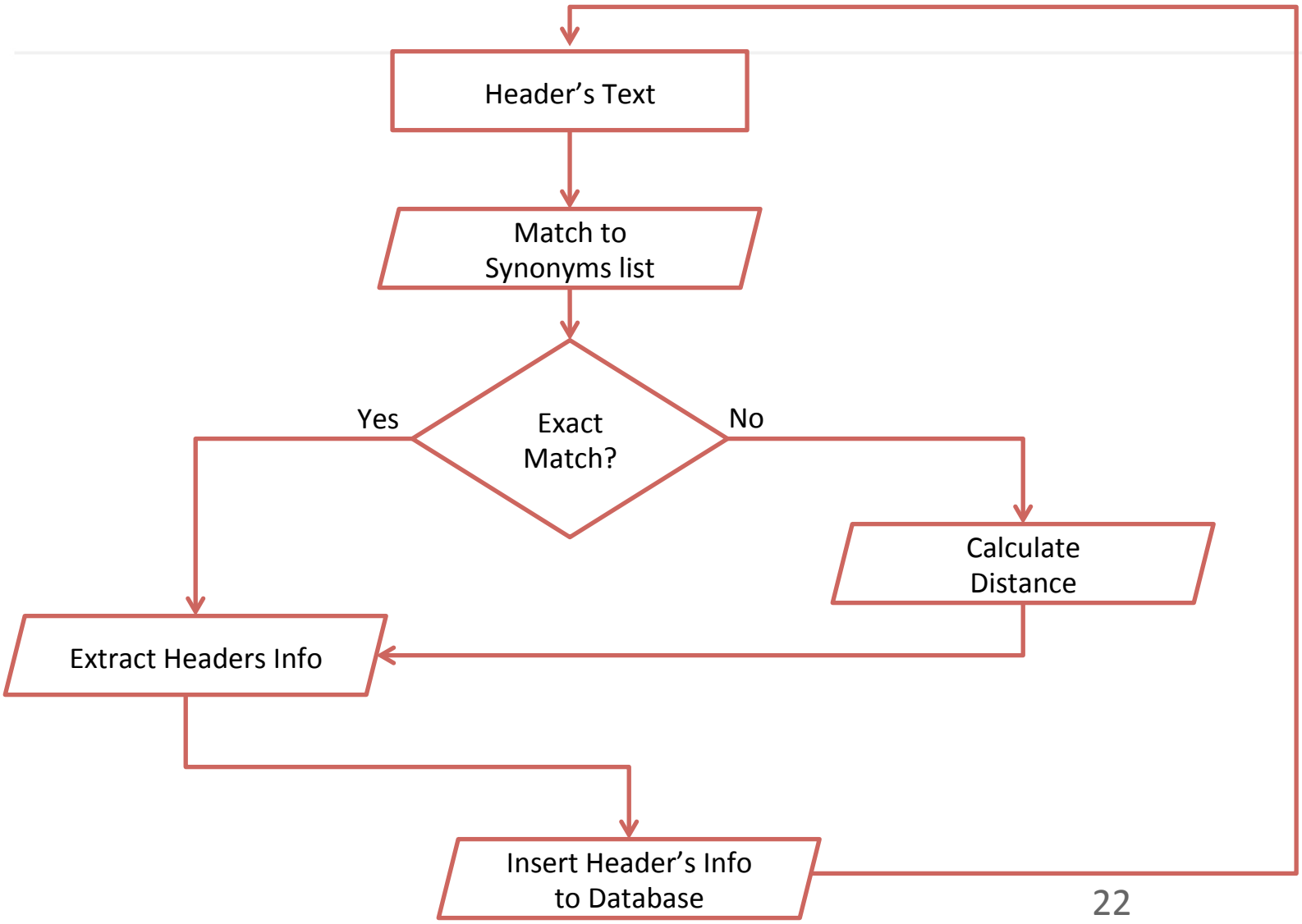
- the minimum number of word edits (insertion, deletion, substitution) required to change one sentence into the other.

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



LANCASTER UNIVERSITY

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

LANCASTER UNIVERSITY

- 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

LANCASTER UNIVERSITY

- 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.

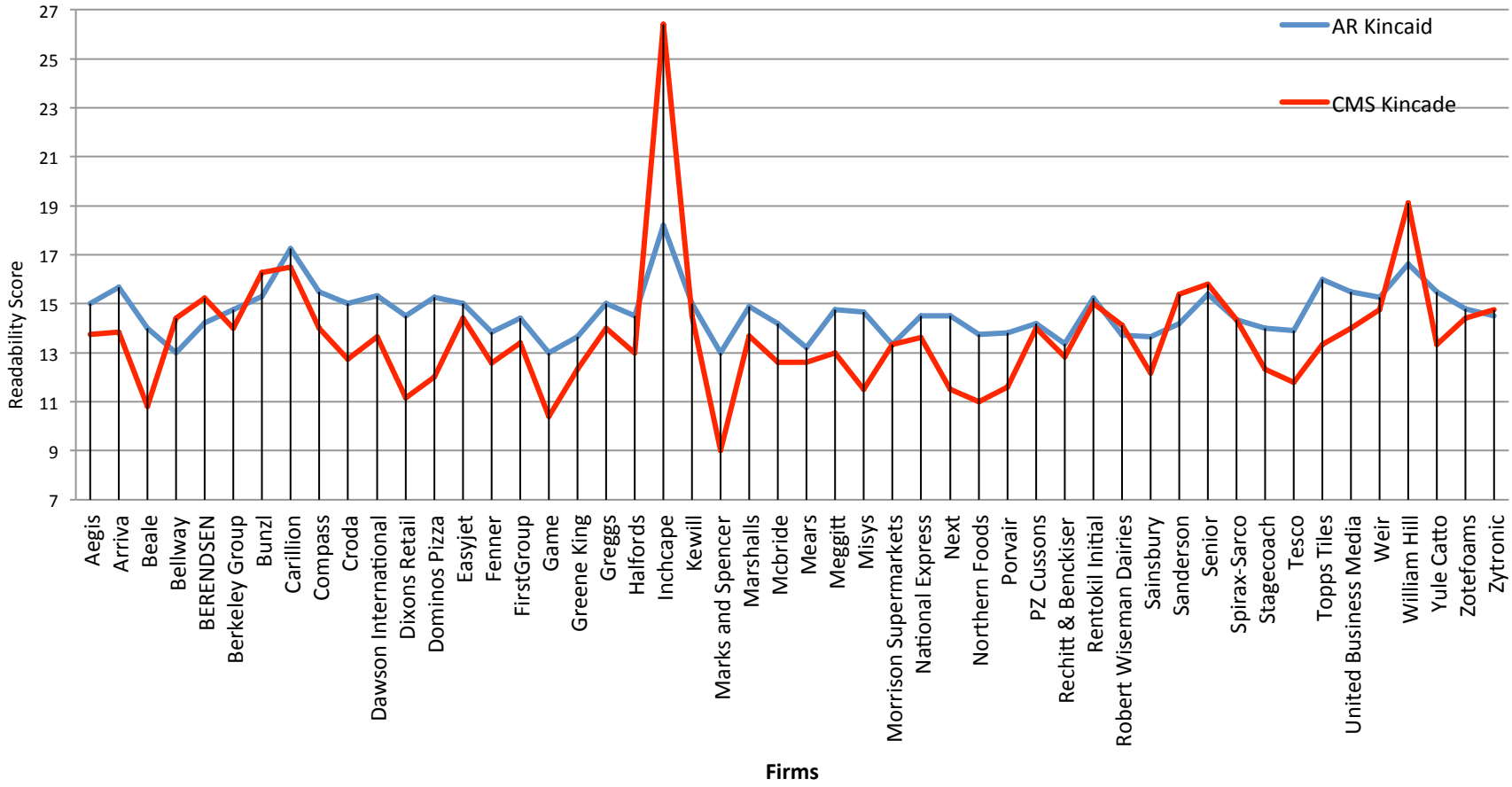
Early highlights from our sample of 250 reports.

Readability Measure Charts

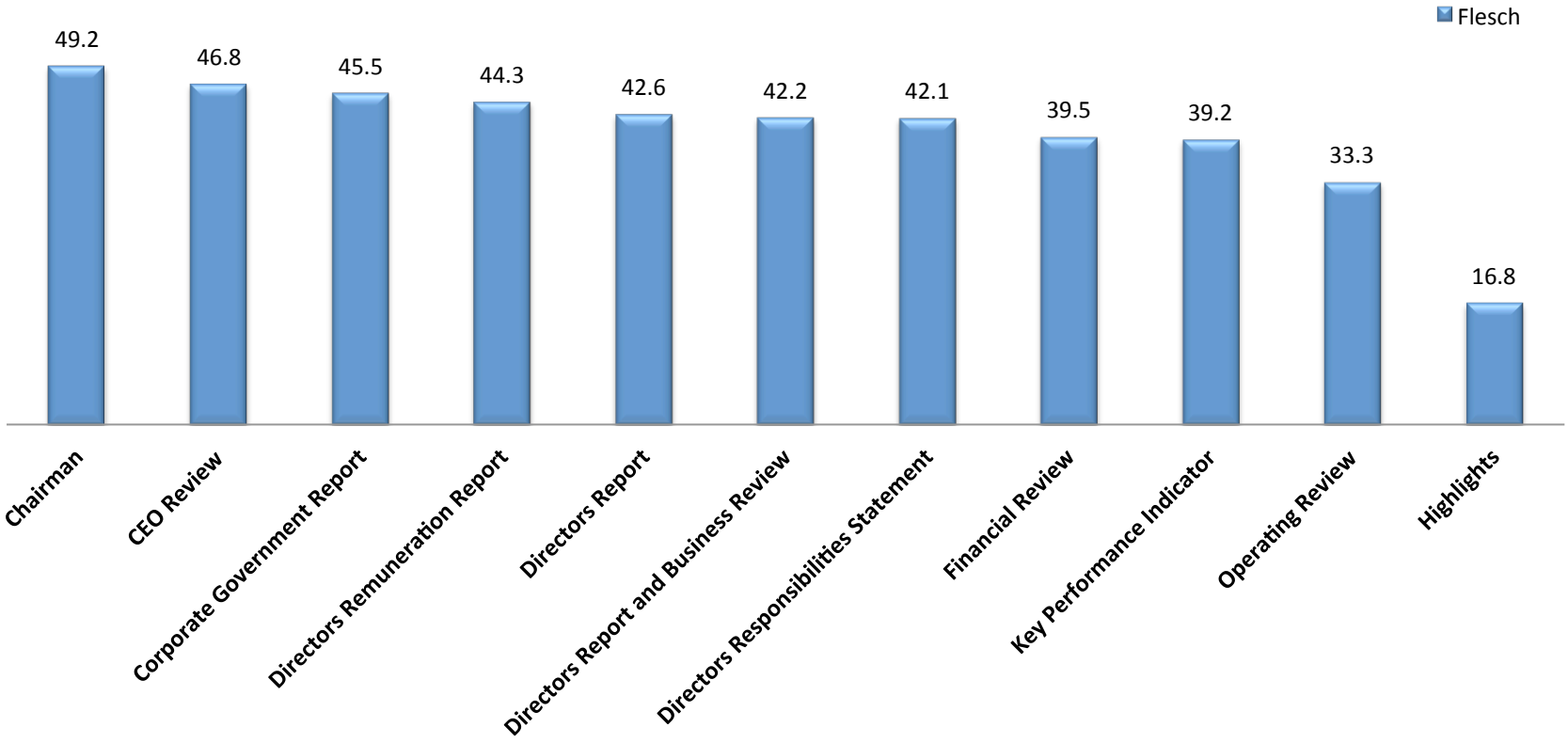
LANCASTER UNIVERSITY

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

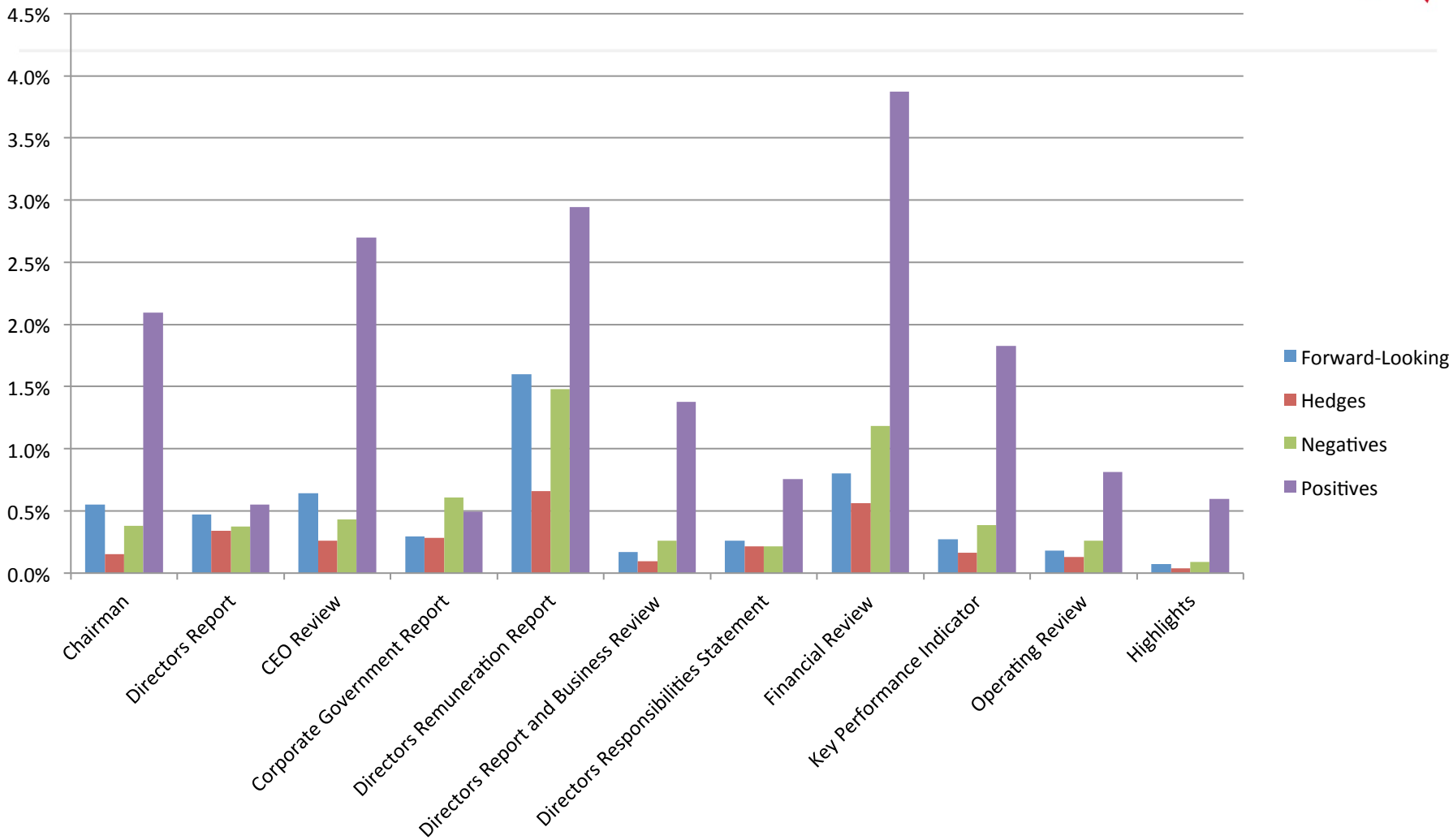
Annual Report vs. Chairman's Statement



Sections Readability Measure



Wordlist Frequencies Across Sections



Challenges

Challenges

LANCASTER UNIVERSITY

- 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

LANCASTER UNIVERSITY

Thank You…

Questions?

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

email:
m.el-haj@lancaster.ac.uk