The Progression of Financial Narrative Processing (FNP)

Mahmoud El-Haj

Lancaster University

@DocElhaj

http://www.lancaster.ac.uk/staff/elhaj/

Talk Objectives

Part 1

- Introduction to FNP projects at Lancaster University
- Concept of textual analysis
- Why it’s potentially important in accounting and financial (AccFin)
- Review (some of) the main AccFin textual analysis methods
- Methods and techniques used in CFIE-FRSE

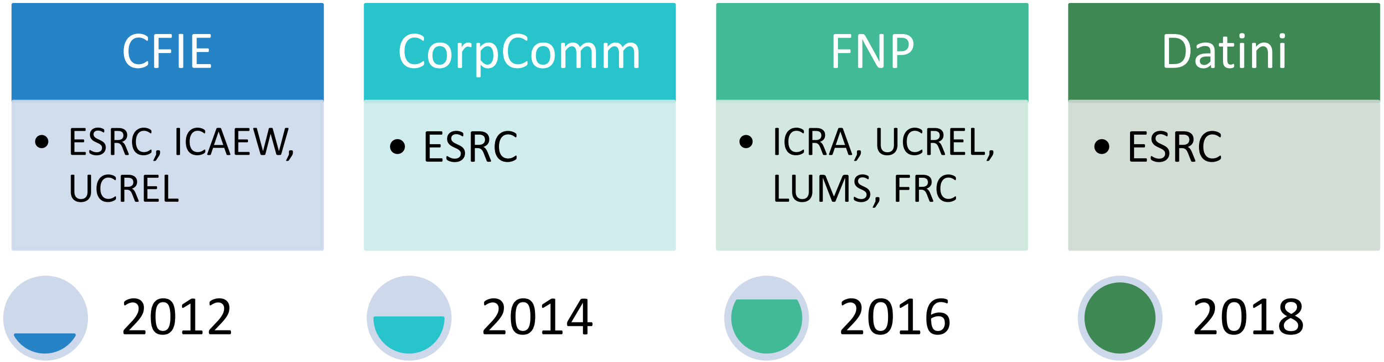
Part 2

- Hands on Demo of CFIE-FRSE software

Part 1

FINANCIAL NARRATIVE PROCESSING

Lancaster University Financial Narrative Projects



CFIE

- ESRC, ICAEW, UCREL

2012

CorpComm

- ESRC

2014

FNP

- ICRA, UCREL, LUMS, FRC

2016

Datini

- ESRC

2018

Datini Project Team

Steven Young
LUMS, Lancaster University

Paul Rayson
SCC, Lancaster University

Martin Walker
Manchester Business School

Mahmoud El-Haj
SCC, Lancaster University

Paulo Alves
Universidade Católica
Portuguesa

Vasiliki Simaki
LAEL, Lancaster University

What is it about?

The projects analyse U.K. financial narratives, their association with financial statement information, and their informativeness for investors.

Contributions

- Developed automated methods for extracting narrative content and structure from UK annual reports provided as PDFs
- First large-sample tests of the incremental predictive ability of UK annual report narratives
- First study to examine the incremental and differential predictive ability of alternative annual report narrative sections
- First study to model disagreement between preparers’ and to examine the impact of disagreement on the predictive properties of narratives

Qualitative Information

Analysis of qualitative information has a long tradition in computer science (Natural Language Processing) and Linguistics (Corpus Linguistics)

Methods only recently started to gain traction in AccFin

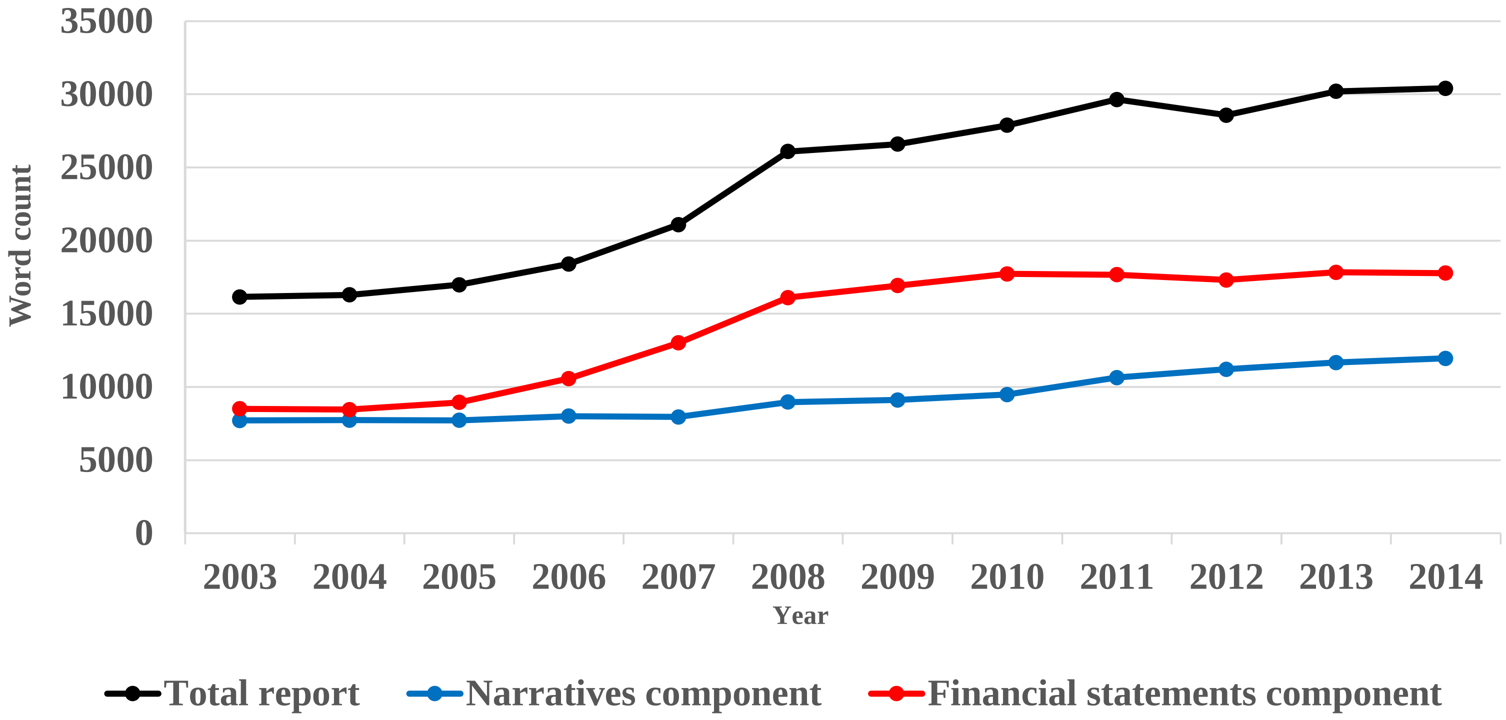
- Earlier work on disclosures involved manual analysis of small samples

AccFin data?

- Estimates: 90% of data created in the last 10 years is qualitative/unstructured data (80% of which in a business context)
- Rapid growth in non-traditional information sources (Social media, blogs, etc.)

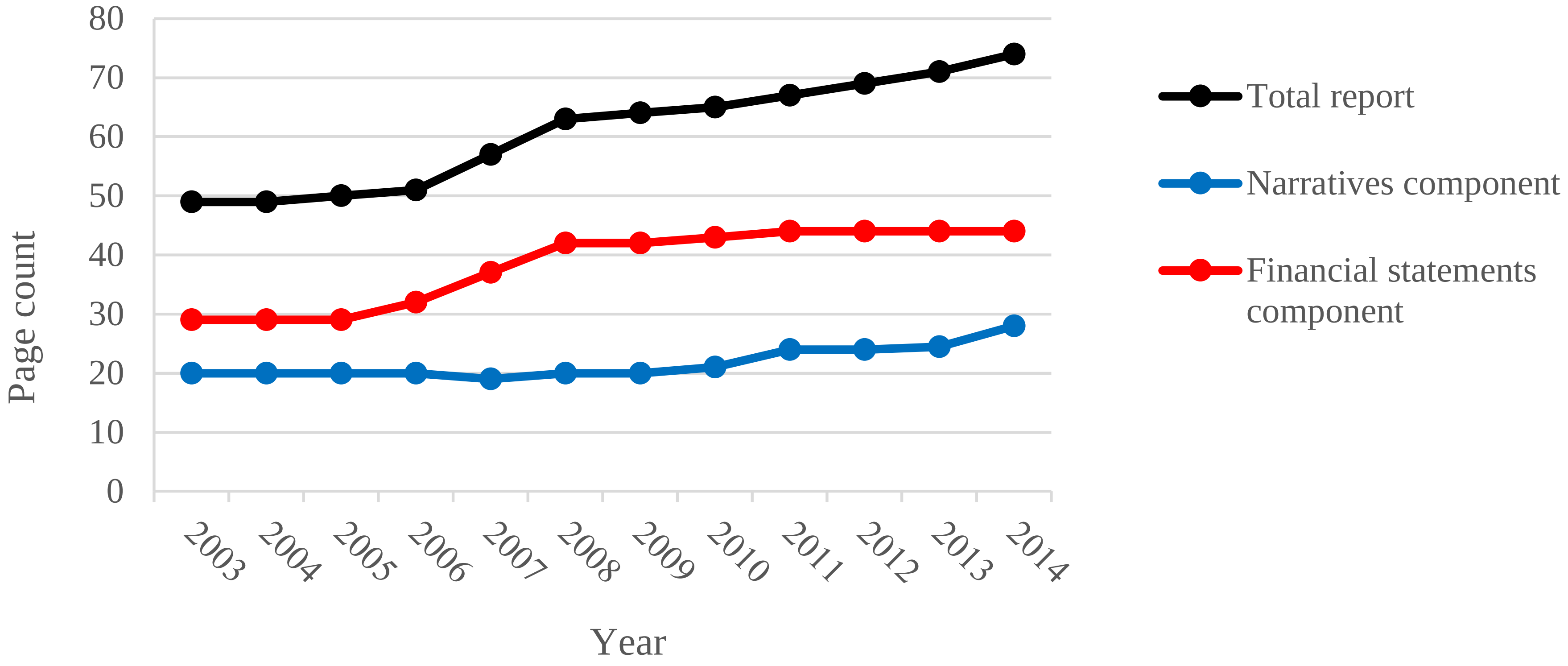
Why Financial Narratives

A doubling of the median word count over the sample period

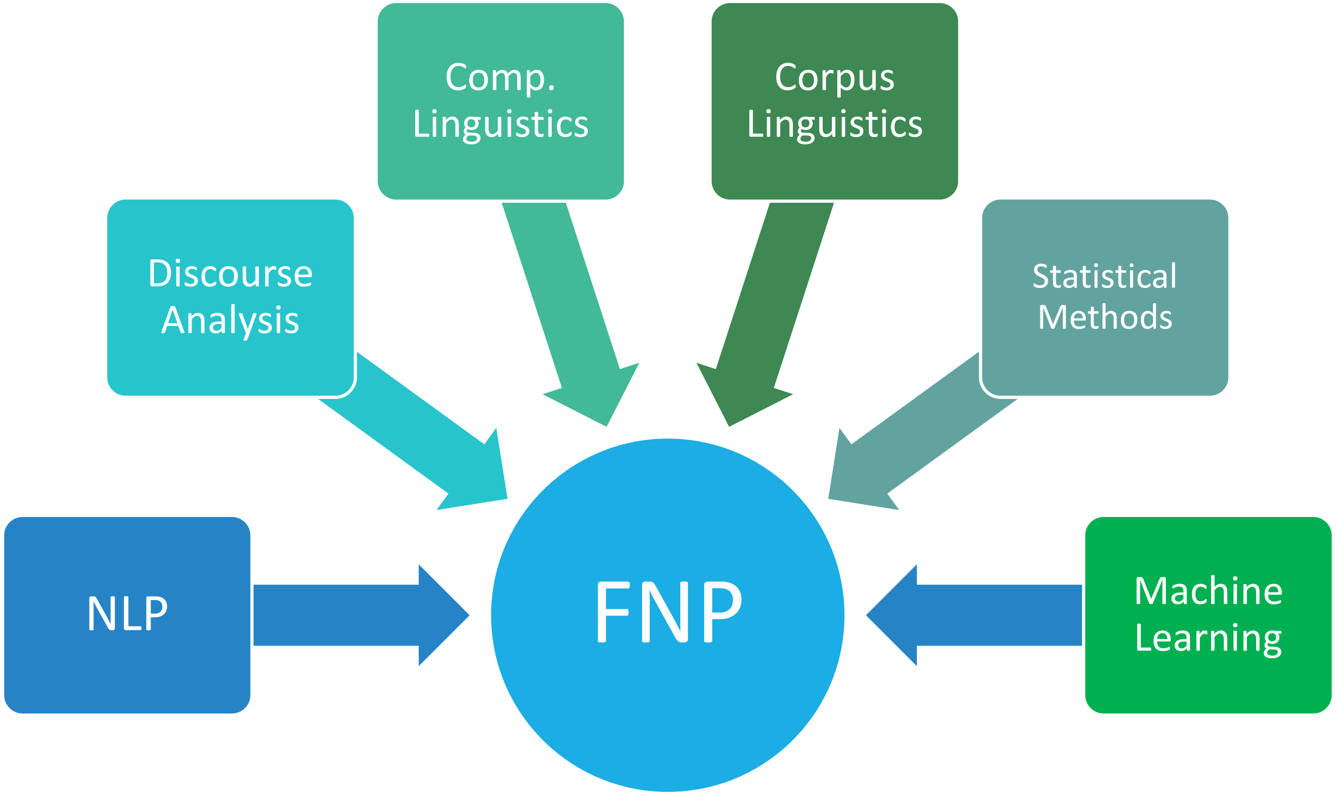


Why Financial Narratives

A doubling of the median page count over the sample period



FNP: Fields of Study



NLP

Discourse Analysis

Comp. Linguistics

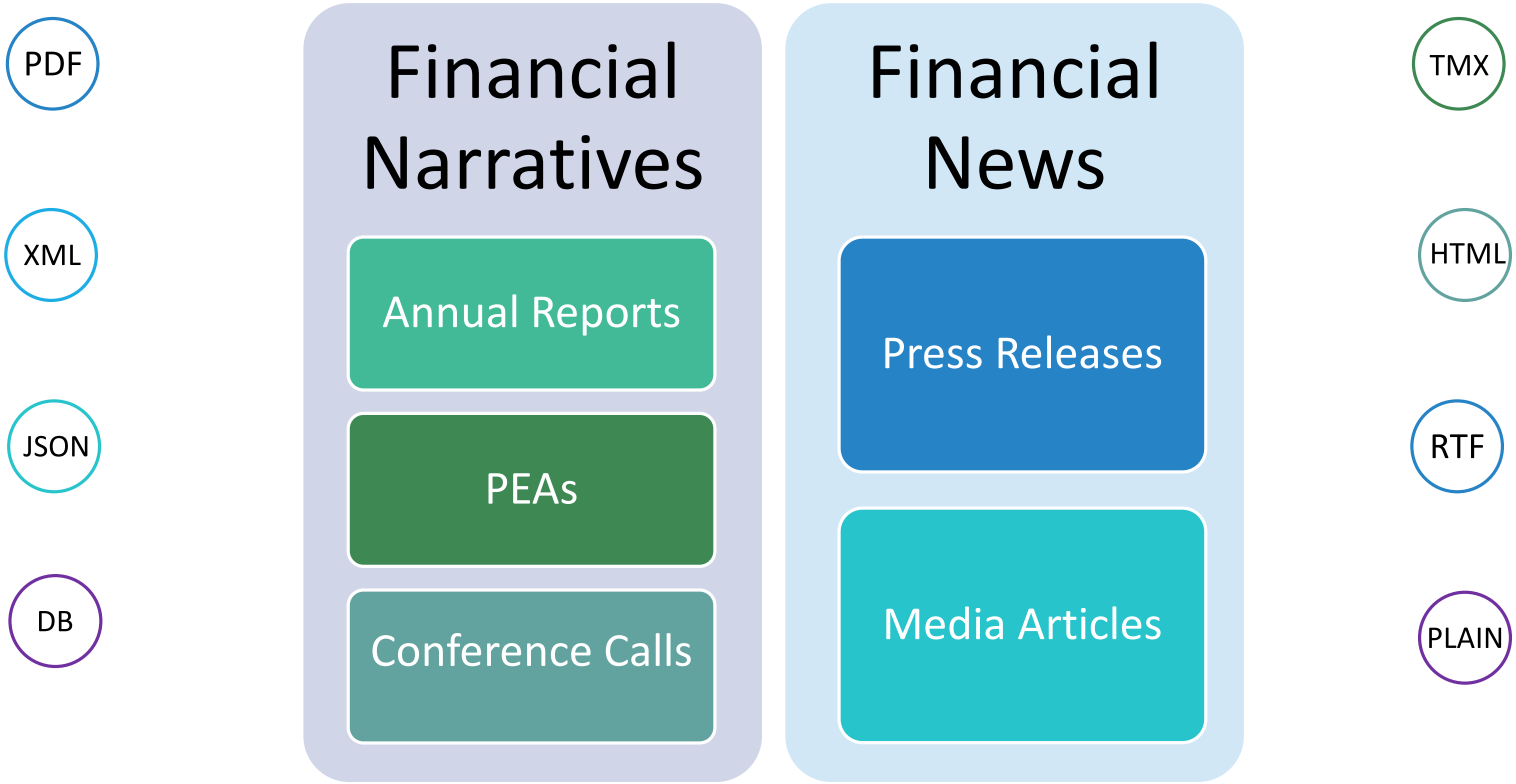
Corpus Linguistics

Statistical Methods

Machine Learning

FNP

Financial Big Data



PDF

XML

JSON

DB

Financial Narratives

Annual Reports

PEAs

Conference Calls

Financial News

Press Releases

Media Articles

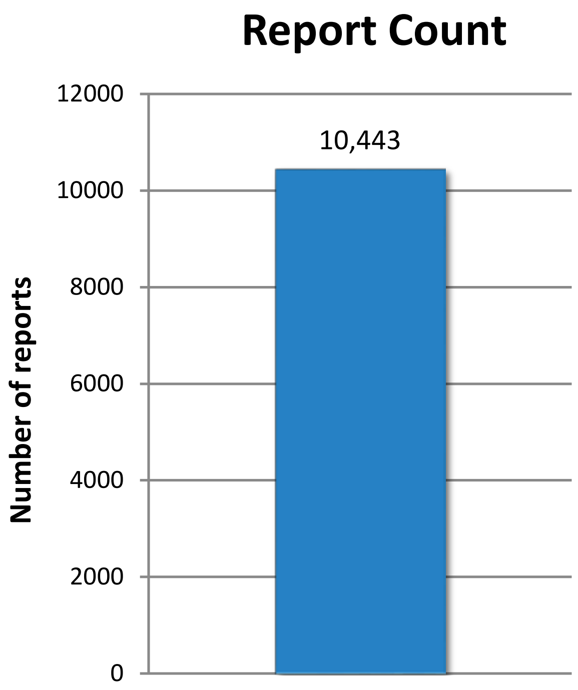
TMX

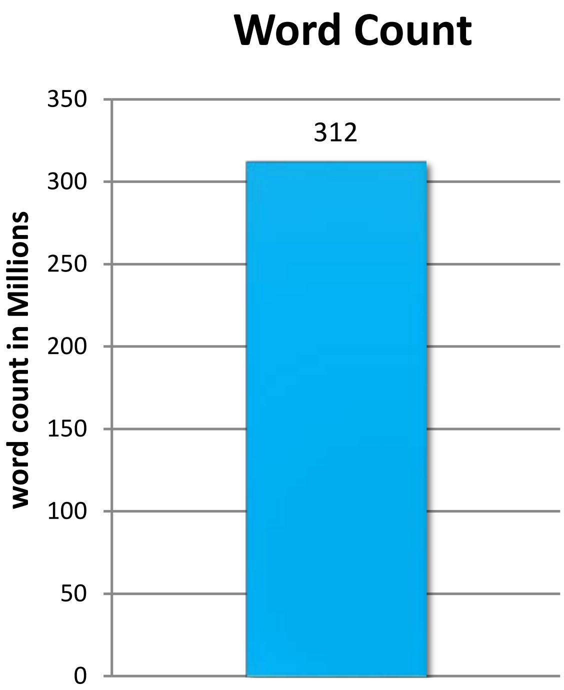
HTML

RTF

PLAIN

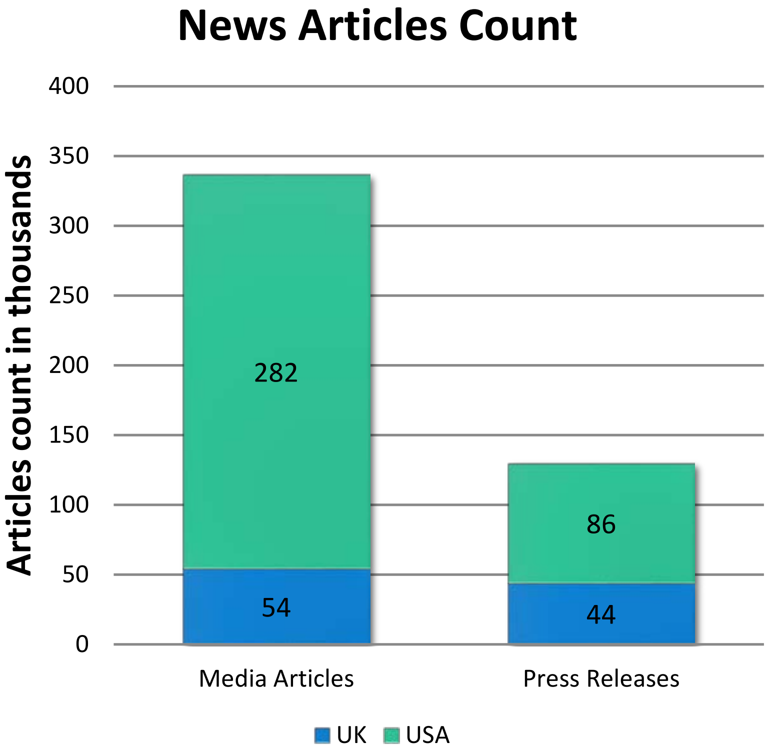
Annual Reports: how big?

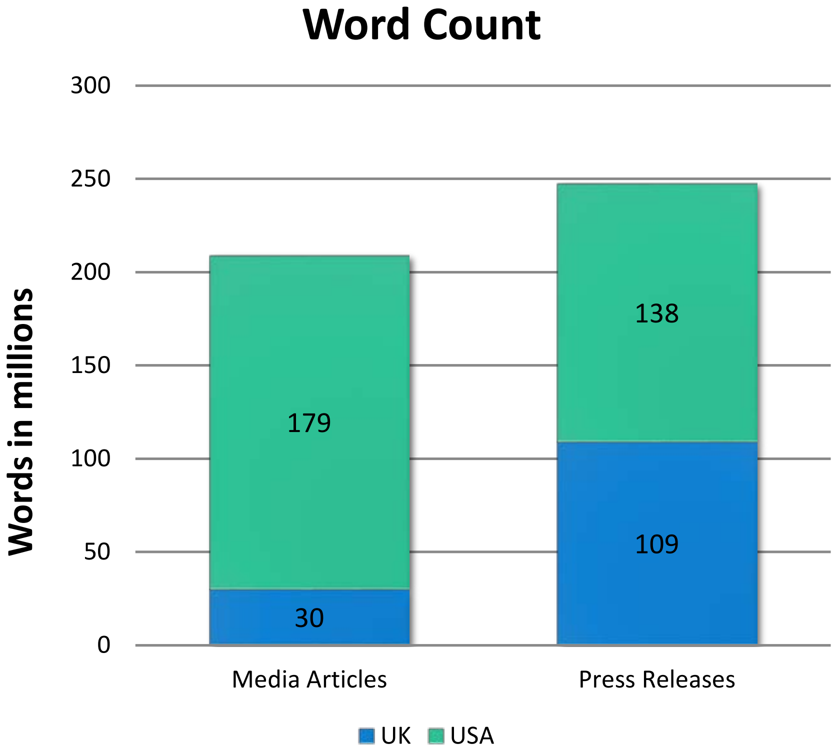




Regularly updated and growing…

Financial Media: how big?





Harvesting Data

- Application of NLP and corpus methods involves large volumes of text
  - SEC, Thomson, Factiva, StreetEvents, Twitter …etc

- Accounting researchers often use Perl/Python script to harvest and process documents (Python SECEdgar).

- Some types of qualitative data are easier to access than others → HTML vs PDF

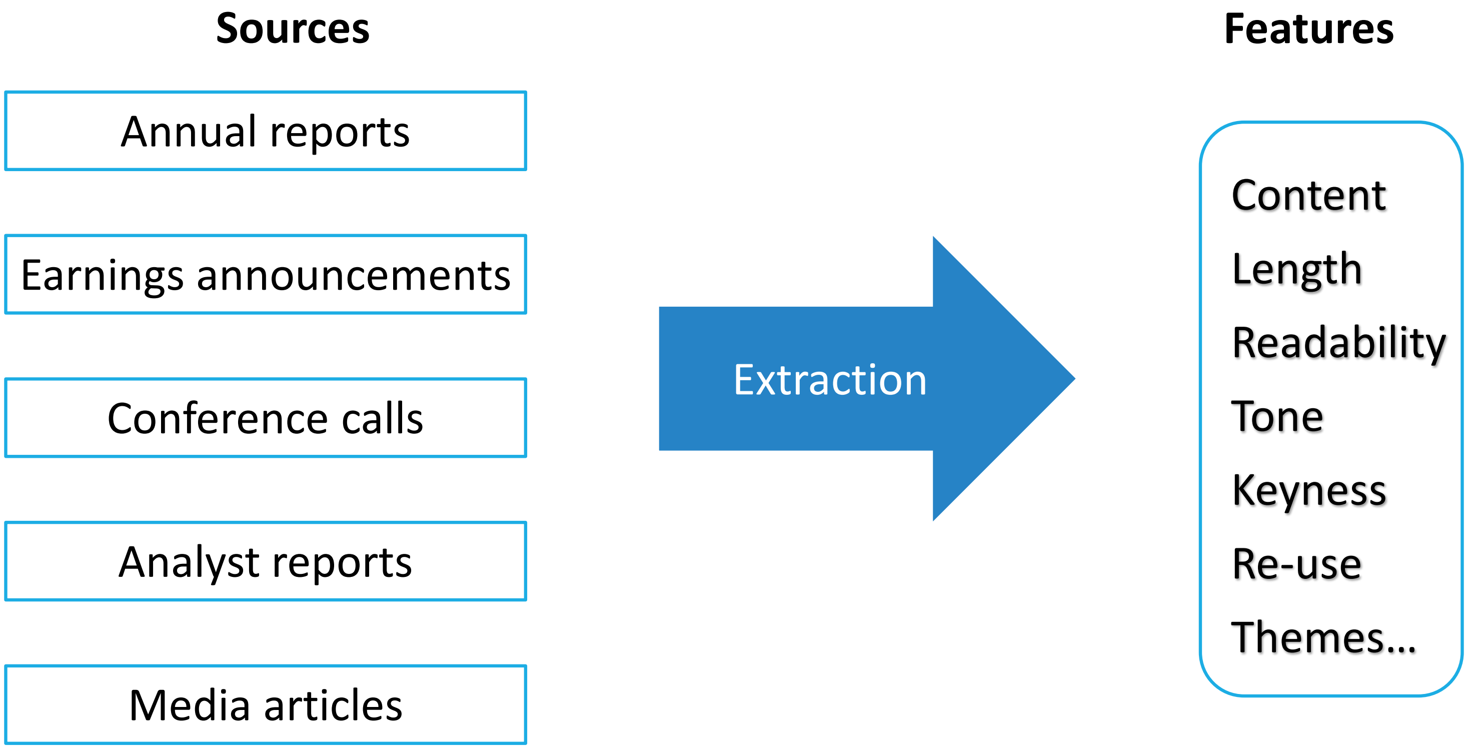
Analysing AccFin Datasets

Word-Level analysis: AccFin research

- Dictionary methods
- Readability
- Text similarity

bag-of-words: words considered in isolation from their context, meaning, grammatical usage, etc.

Common AccFin Features



Sources

Annual reports

Earnings announcements

Conference calls

Analyst reports

Media articles

Extraction

Features

Content

Length

Readability

Tone

Keyness

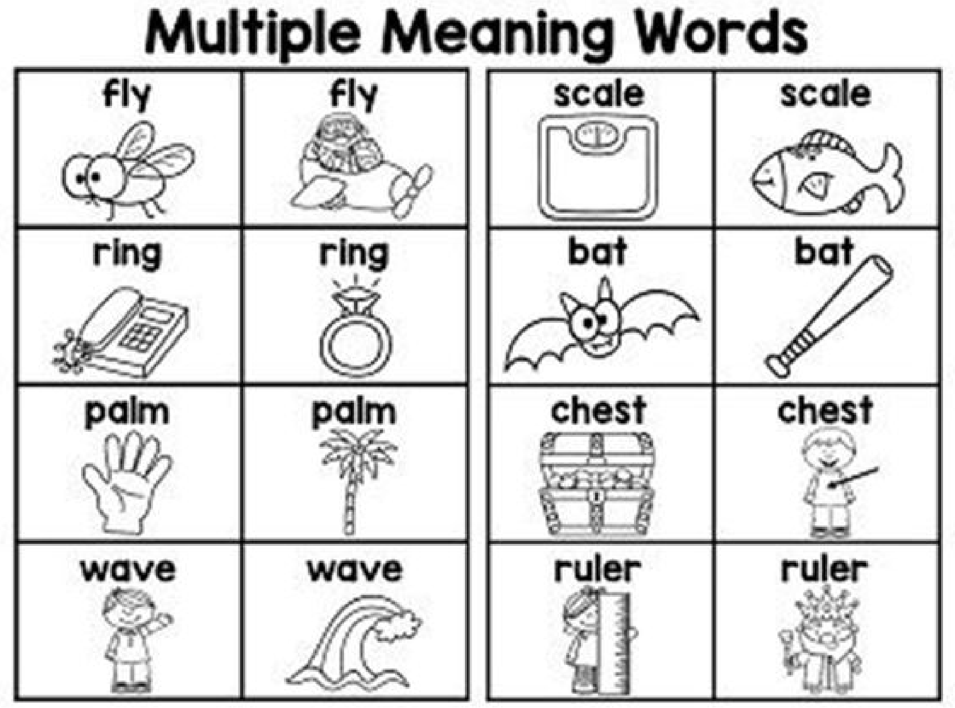
Re-use

Themes…

Approaches to Text Analysis

- **Bag-of-words methods:**
  - Readability: Ease of understanding for English writing (Fog, Flesch)
  - Text similarity: Similarity of language between two or more sections of text (Cosine Similarity).
- **Statistical approaches:**
  - Text classifiers
  - Clustering
  - Information extraction
- **Natural Language Processing**
  - Helps the previous two methods by giving meaning to words:
  - Part-of-speech (POS) tagging
  - Semantic analysis

Bag-of-words limitation



Duck , saw, bat, tear, bank, can, tie, park, wind, second, row, refuse, bow, minute…etc

NLP topics

- Text Categorisation
- Morphology
- Multimodality
- Ontologies
- Part of Speech Tagging
- Tweet Corpora and Analysis
- Twitter-Related Analysis
- Social Media
- Word Sense Disambiguation
- Prosody and Phonology
- Crowdsourcing
- Corpus Querying and Crawling
- Grammar and Syntax
- Parallel and Comparable Corpora

- Machine Translation and Evaluation
- Sentiment Analysis and Emotion Recognition
- Corpora for Language Analysis
- Information Extraction and Retrieval
- Multimodality
- Multiword Expressions
- Named Entity Recognition
- Parsing
- Summarisation
- Word Sense Disambiguation
- Multilingual Corpora
- Lexicons
- Semantics
- Sentiment Analysis and Opinion Mining

FNP 2018 Workshop:
http://wp.lancs.ac.uk/cfie/fnp2018

Text Analysis: AccFin vs CS/CL

AccFin has only scratched the surface:

- Reliance on basic NLP techniques primarily involving bag-of-words methods
- Little use of corpus methods

> 10-20 years behind developments in computational linguistics and machine learning

Text Analysis: AccFin

Majority of work on simple structured documents:

- 10-Ks via EDGAR
- Conference call transcripts
- Media articles

Other more sophisticated documents:

- PDF annual reports → no standardised structure, poor accessibility, infographics
- Comment letters → different styles, various formats, irrelevant content
- Regulatory documents → no standardised structure, PDF files
- Preliminary Earning Announcements (commentary level)

Working with Annual Reports

Can we apply methods used in prior US studies to UK annual reports?

Before we can answer this, let’s have a look at the differences between UK and US annual reports.

US Filings

SEC Edgar: US companies must submit:

1. 10-K: Annual
2. 10-Q: Quarterly
3. 8-K: Special Events
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….

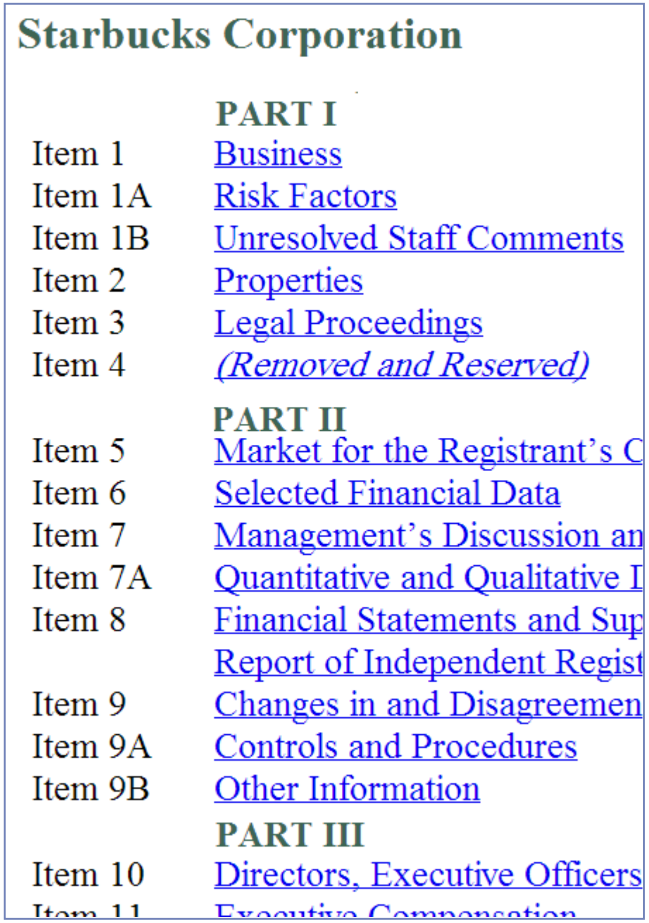
ITEM 13. Certain Relationships and Related Transactions….

ITEM 14. Principal Accounting Fees and Services

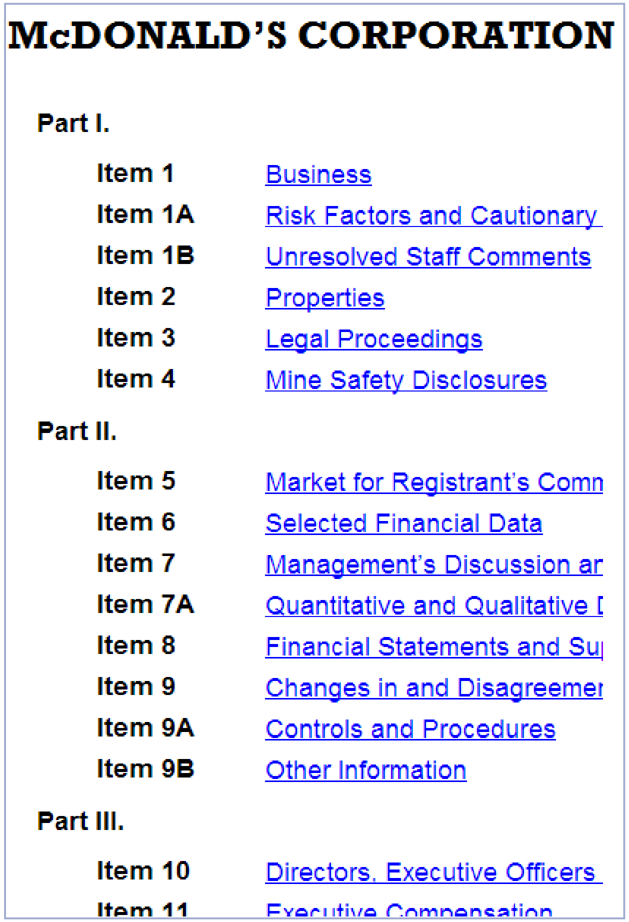
PART IV

ITEM 15. Exhibits, Financial Statement Schedules….

10-K Annuals (Starbucks vs McDonald's)



Starbucks Corporation

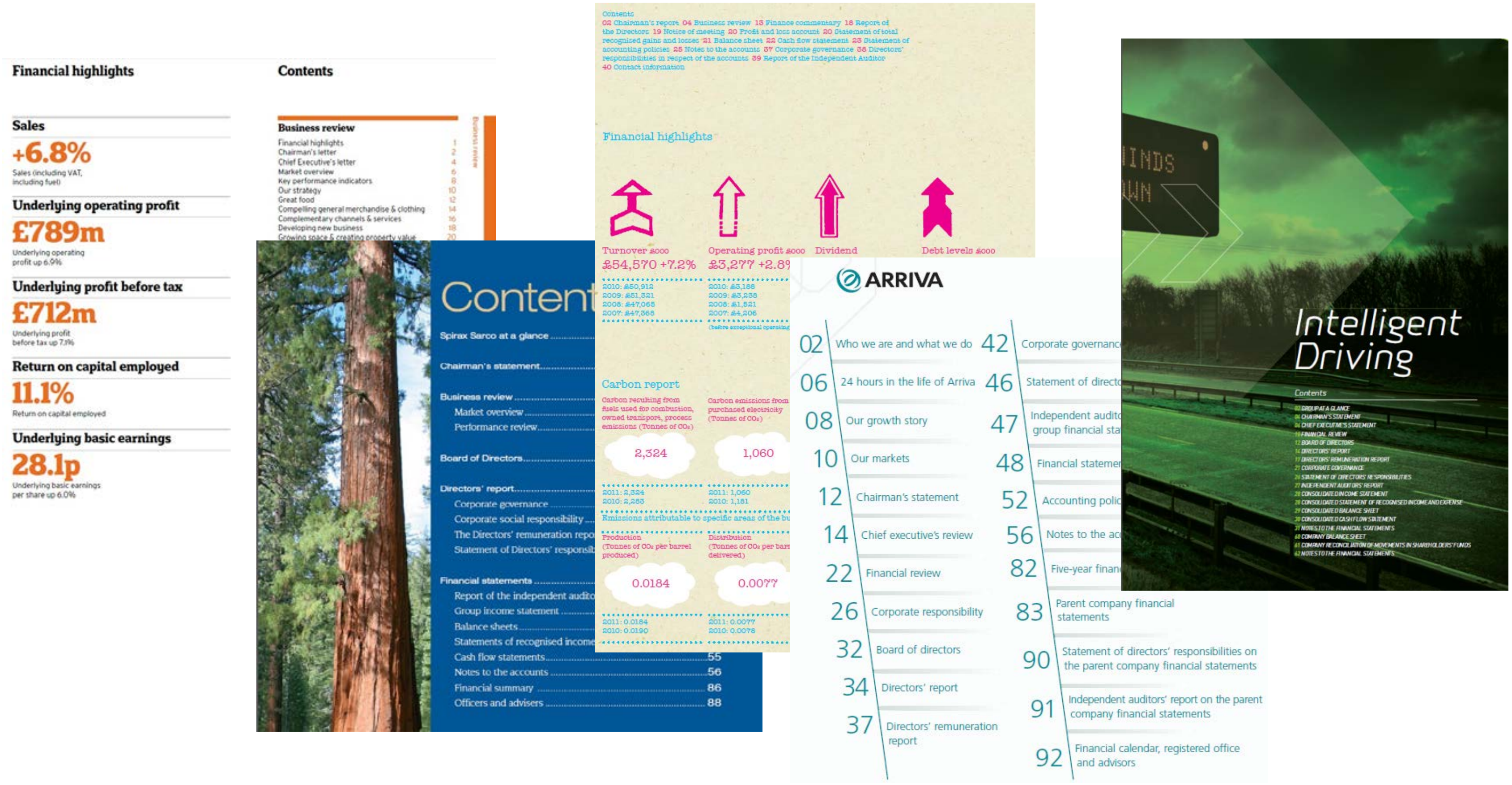


McDONALD'S CORPORATION

UK Annual Reports

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

Content

Contents

Financial highlights

Turnover £000
£54,570 +7.2%
2010: £50,912
2009: £51,321
2008: £47,065
2007: £47,368

Operating profit £000
£3,277 +2.8%
2010: £3,188
2009: £3,238
2008: £1,521
2007: £4,206
(before exceptional operating

Dividend

Debt levels £000

Carbon report

Carbon resulting from fuels used for combustion, owned transport, process emissions (Tonnes of CO_2)
2,324
2011: 2,324
2010: 2,283

Carbon emissions from purchased electricity (Tonnes of CO_2)
1,060
2011: 1,060
2010: 1,181

Emissions attributable to specific areas of the bu

Production (Tonnes of CO_2 per barrel produced)
0.0184
2011: 0.0184
2010: 0.0190

Distribution (Tonnes of CO_2 per barr delivered)
0.0077
2011: 0.0077
2010: 0.0078

ARRIVA

Intelligent Driving

Contents

Non-10-K Annual Reports

Lang & Stice-Lawrence (2015) first large sample analysis of non-10-K AR (> 87,000 PDFs)

++

- Converting files to plain text format
- Isolating running text with a Perl script

−−

- Unable to distinguish disclosures in the footnotes to the financial statements from commentary in the narrative component of the report
- Unable to distinguish between disclosures from distinct sections of the narrative component
- No information on document structure → important dimension of disclosure

Non-10-K Annual Reports

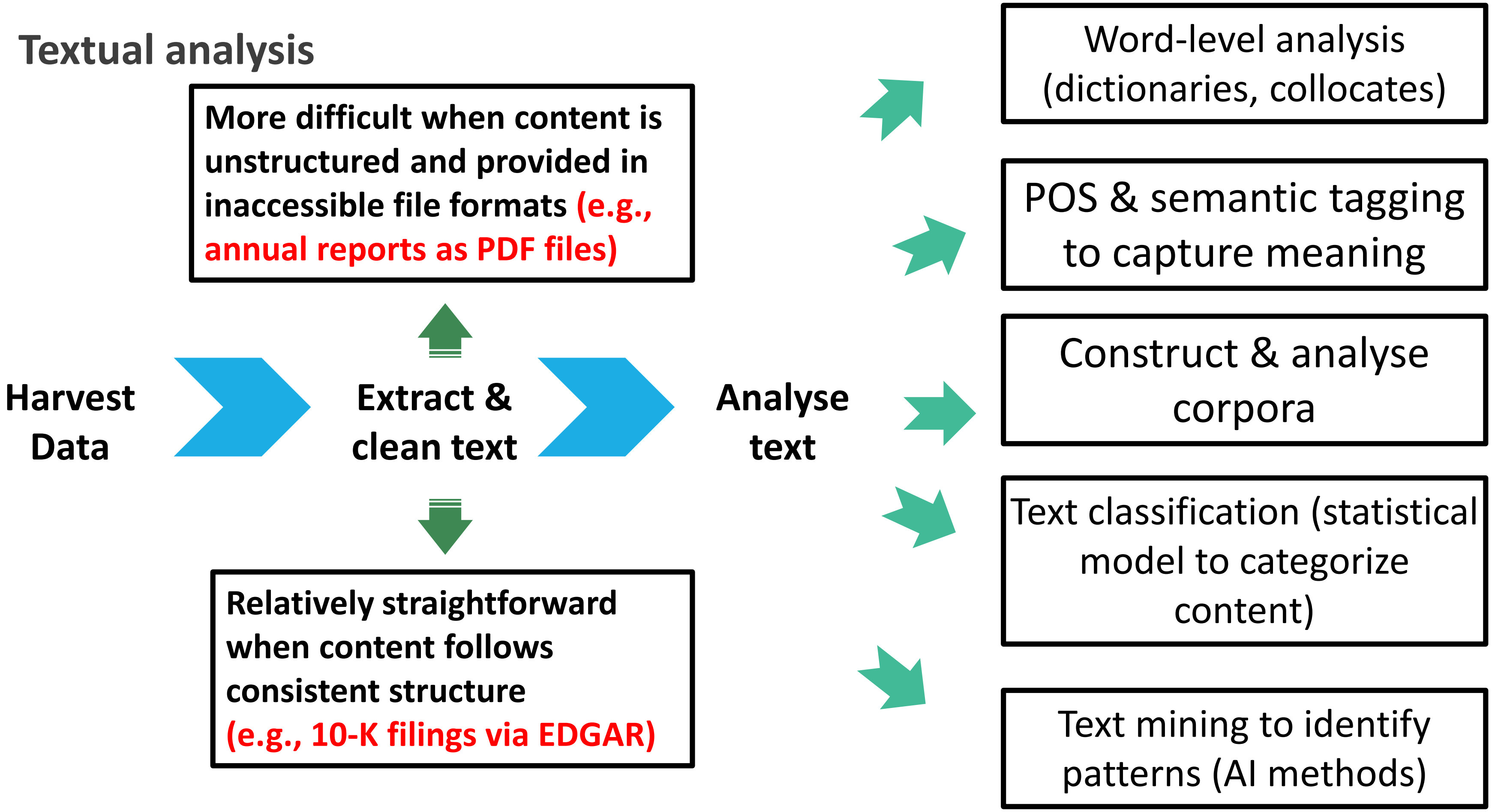
Lancaster FNP: we developed a software tool for extracting and classifying narrative content from digital PDF annual reports

- Detect table of contents
- Parse extracted table of contents
- Synchronize page numbers in the digital PDF file
- Determine start and end of each section
- Partition contents into financial statements and the “front-end” narratives component
- Sub-classify narratives (shareholders’ letter, CEO review, CFO review, governance statements, remuneration reports)

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

General Flow Diagram

Textual analysis



More difficult when content is unstructured and provided in inaccessible file formats (e.g., annual reports as PDF files)

Harvest Data

Extract & clean text

Analyse text

Relatively straightforward when content follows consistent structure (e.g., 10-K filings via EDGAR)

Word-level analysis (dictionaries, collocates)

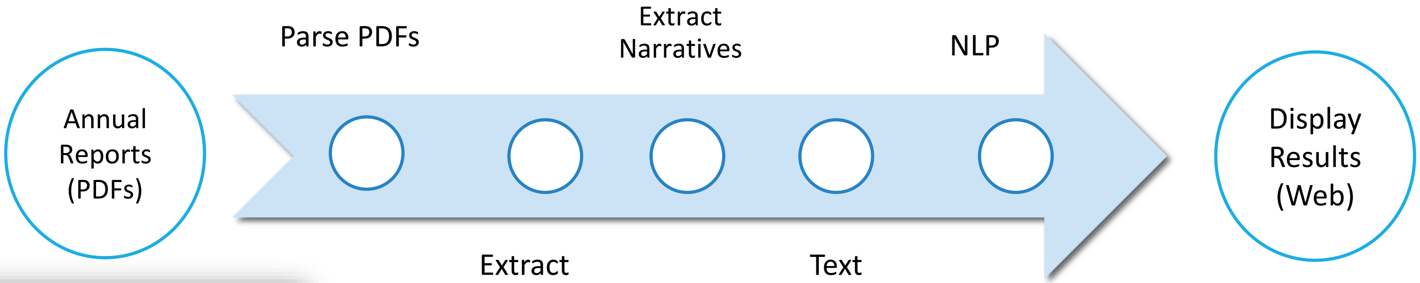
POS & semantic tagging to capture meaning

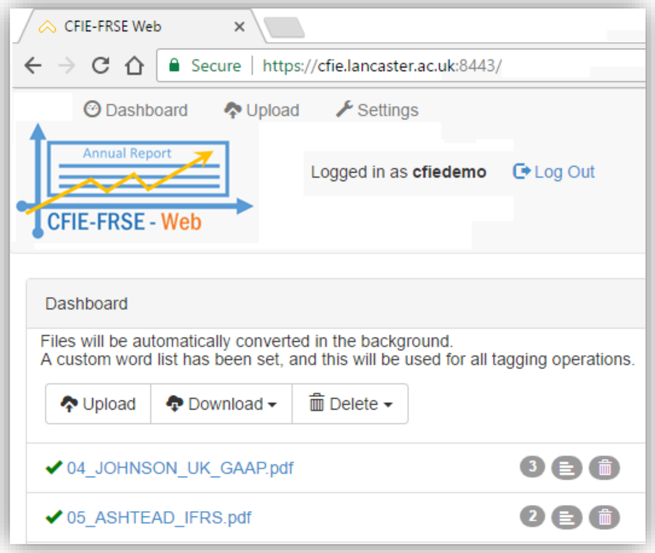
Construct & analyse corpora

Text classification (statistical model to categorize content)

Text mining to identify patterns (AI methods)

CFIE-FRSE Tool





Logged in as **cfiedemo**

Dashboard

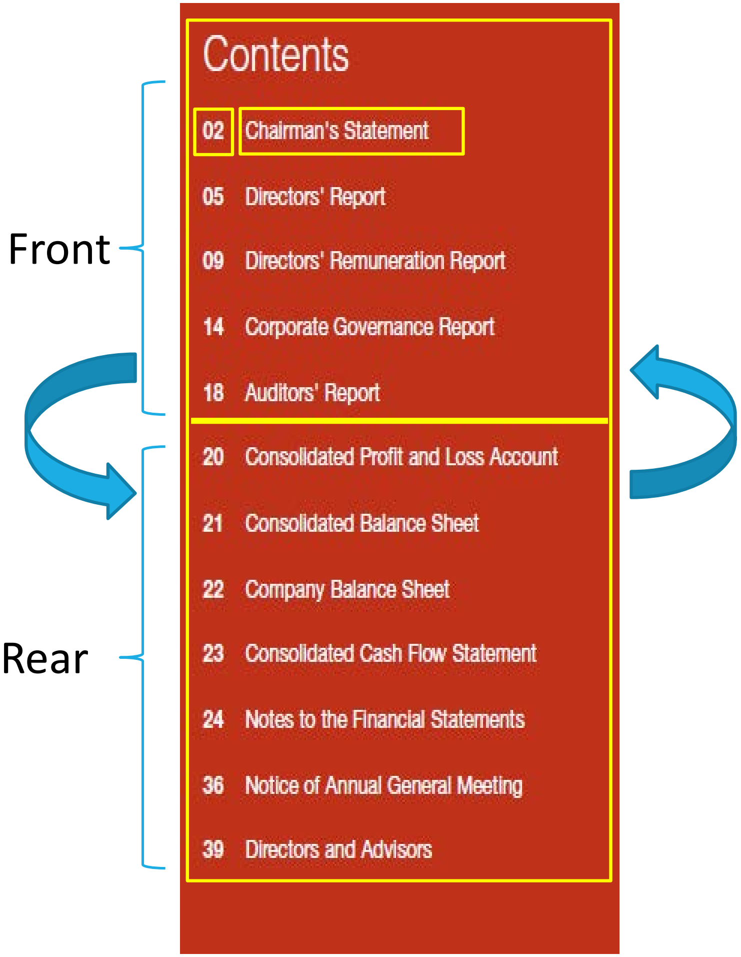
Files will be automatically converted in the background.
A custom word list has been set, and this will be used for all tagging operations.

✓ 04_JOHNSON_UK_GAAP.pdf 3

✓ 05_ASHTEAD_IFRS.pdf 2

UK Annual Reports

- Steps in extraction process:
  - Detect contents page
  - Parse contents page
  - Extract section
  - Detect section type
  - Reorder section



Contents

Report Classification

Heuristic Approach https://cfie.lancaster.ac.uk:8443/

Machine Learning https://github.com/drelhaj/MachineLearning

Classify narratives (front) component into:

- Chairman’s statement
- 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.)

SSRNPaper

El-Haj (2018):

Retrieving, Classifying and Analysing Narrative Commentary in Unstructured (Glossy) Annual Reports Published as PDF Files

https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2803275

Output Evaluation

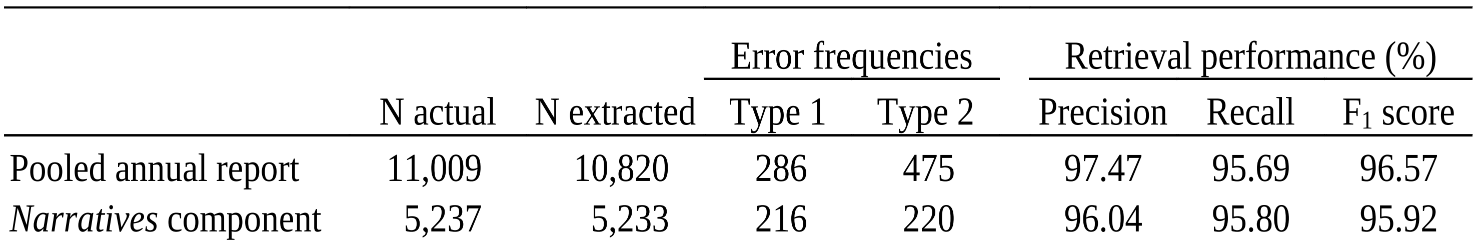
Manual evaluation

- Precision (Type I errors) → false positives
- Recall (Type II errors) → false negatives
- Overall accuracy (F_1) → harmonic mean of precision and recall

- Compare extracted section headers with table of contents
- Assigned page numbers with actual page numbers from table of contents
- Examine accuracy of section classification
- Evaluations based on 586 reports selected at random (approx. 5% of initial population)
  - 11,720 annual report sections

Output Evaluation

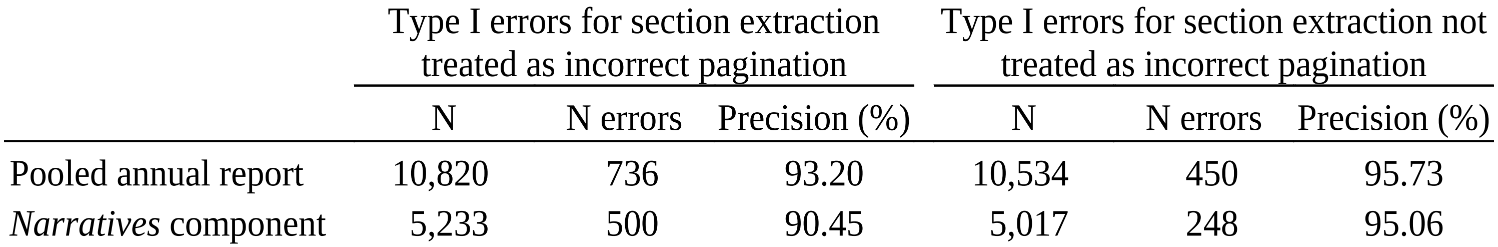
Panel A: Section extraction



			Error frequencies		Retrieval performance (%)		
	N actual	N extracted	Type 1	Type 2	Precision	Recall	F_1 score
Pooled annual report	11,009	10,820	286	475	97.47	95.69	96.57
Narratives component	5,237	5,233	216	220	96.04	95.80	95.92

Output Evaluation

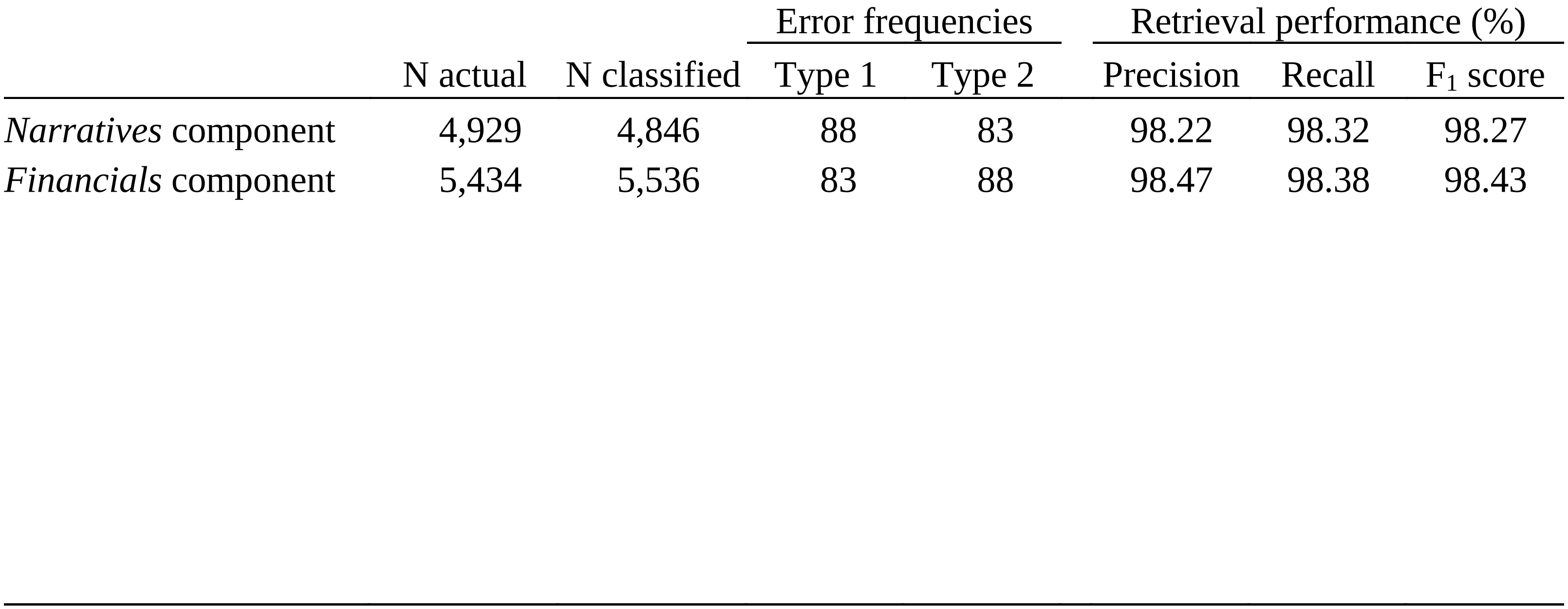
Panel B: Page number synchronization



	Type I errors for section extraction treated as incorrect pagination			Type I errors for section extraction not treated as incorrect pagination		
	N	N errors	Precision (%)	N	N errors	Precision (%)
Pooled annual report	10,820	736	93.20	10,534	450	95.73
Narratives component	5,233	500	90.45	5,017	248	95.06

Output Evaluation

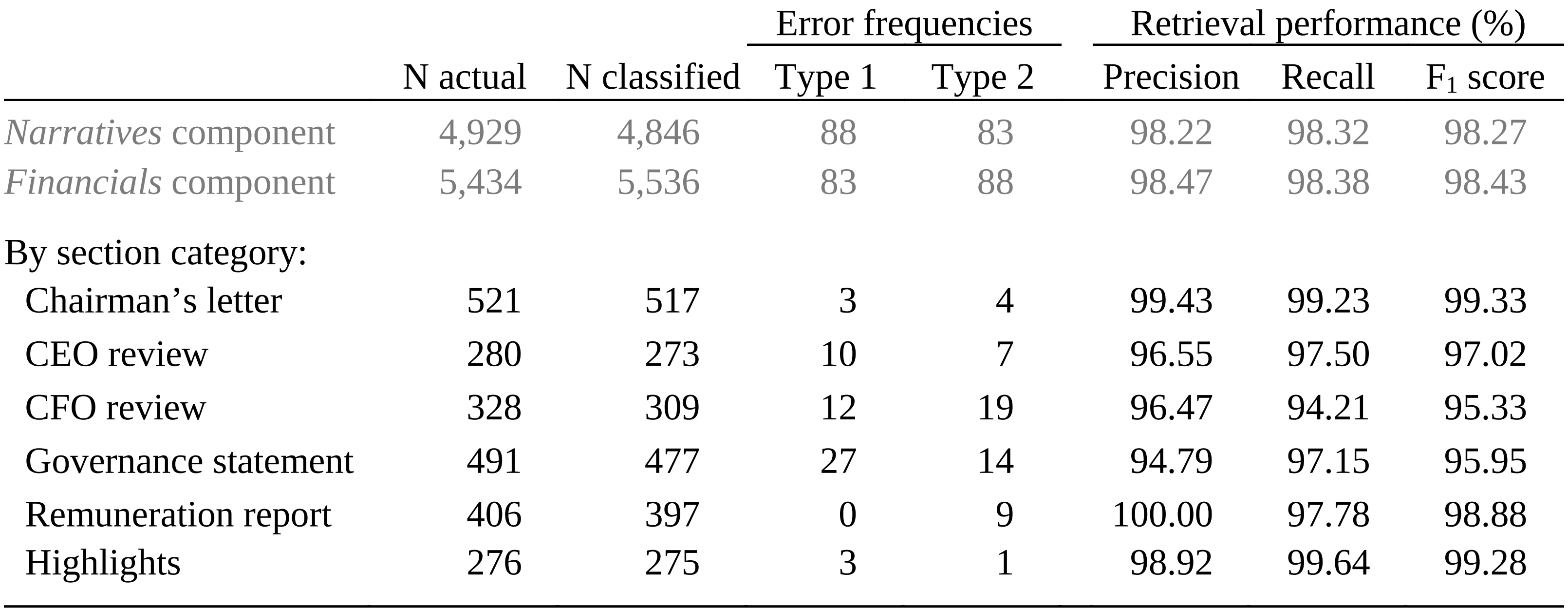
Panel C: Document classification



	N actual	N classified	Error frequencies		Retrieval performance (%)		
			Type 1	Type 2	Precision	Recall	F_1 score
Narratives component	4,929	4,846	88	83	98.22	98.32	98.27
Financials component	5,434	5,536	83	88	98.47	98.38	98.43

Output Evaluation

Panel C: Document classification



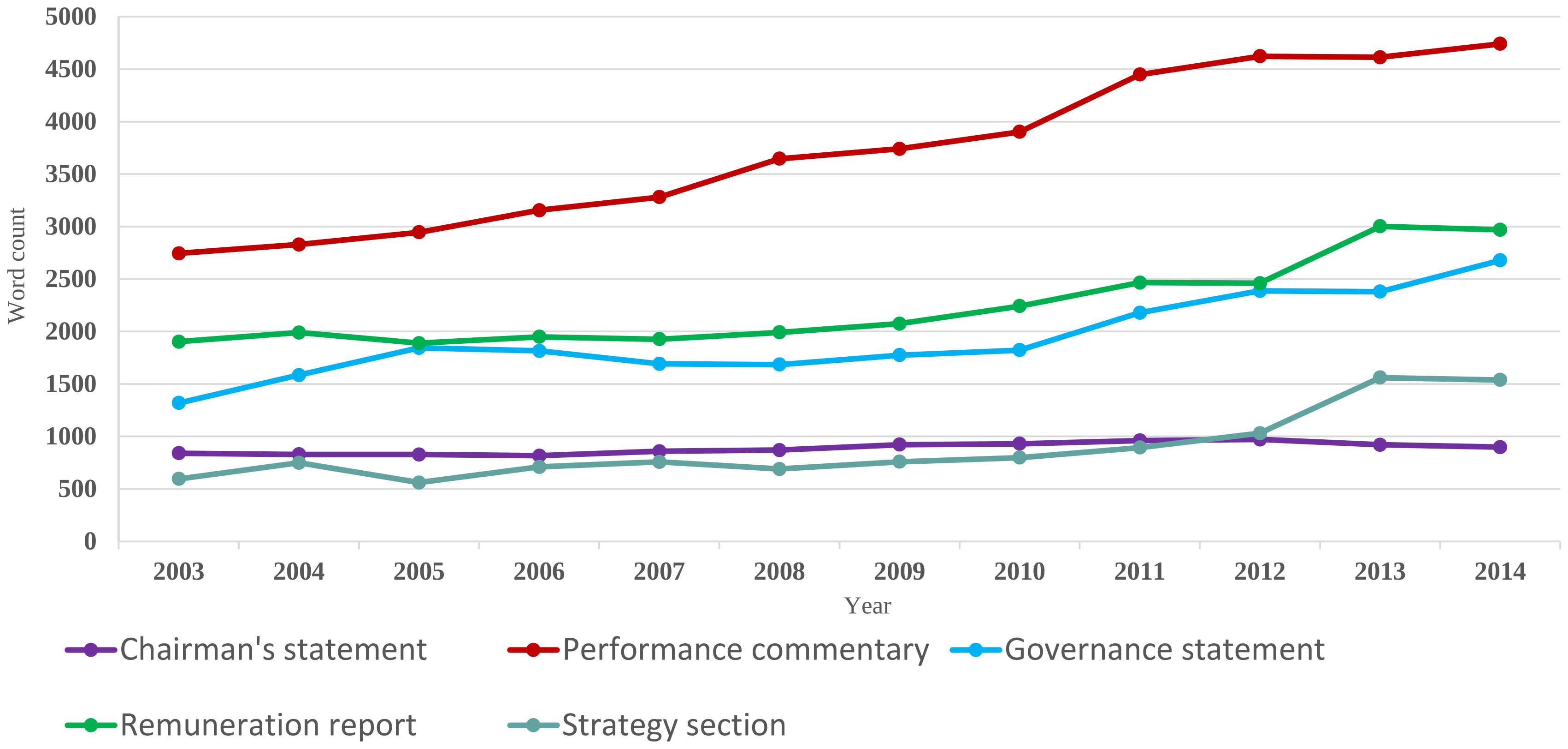
			Error frequencies		Retrieval performance (%)		
	N actual	N classified	Type 1	Type 2	Precision	Recall	F_1 score
Narratives component	4,929	4,846	88	83	98.22	98.32	98.27
Financials component	5,434	5,536	83	88	98.47	98.38	98.43
By section category:							
Chairman’s letter	521	517	3	4	99.43	99.23	99.33
CEO review	280	273	10	7	96.55	97.50	97.02
CFO review	328	309	12	19	96.47	94.21	95.33
Governance statement	491	477	27	14	94.79	97.15	95.95
Remuneration report	406	397	0	9	100.00	97.78	98.88
Highlights	276	275	3	1	98.92	99.64	99.28

Output

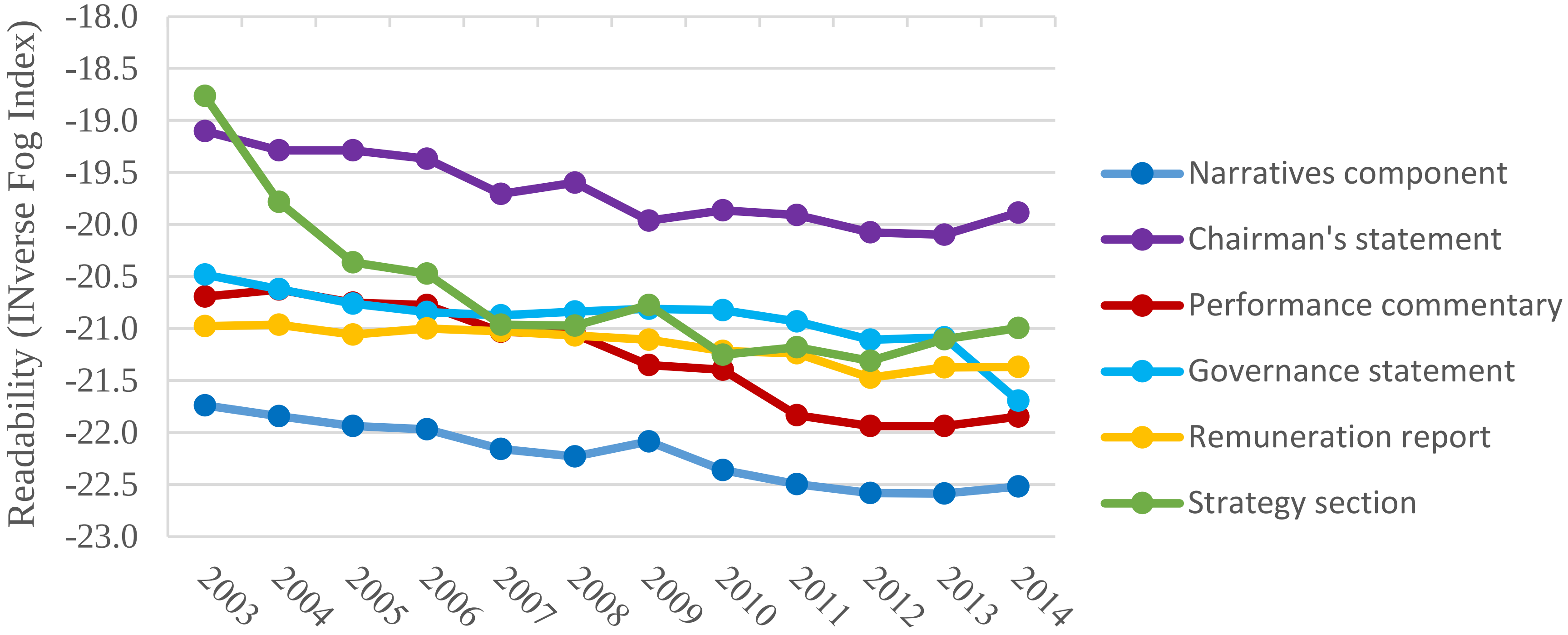
- NLP publicly available tool (CFIE-FRSE)
- First large scale study of UK annual reports structure.
- First ever published disclosure scores for UK annual reports.
- Approaches used help speed up the analysis process and close the gap between firms and investors.
- Leads to better understanding of corporate financial decisions and corporate financial performance.

Reporting growing over Time

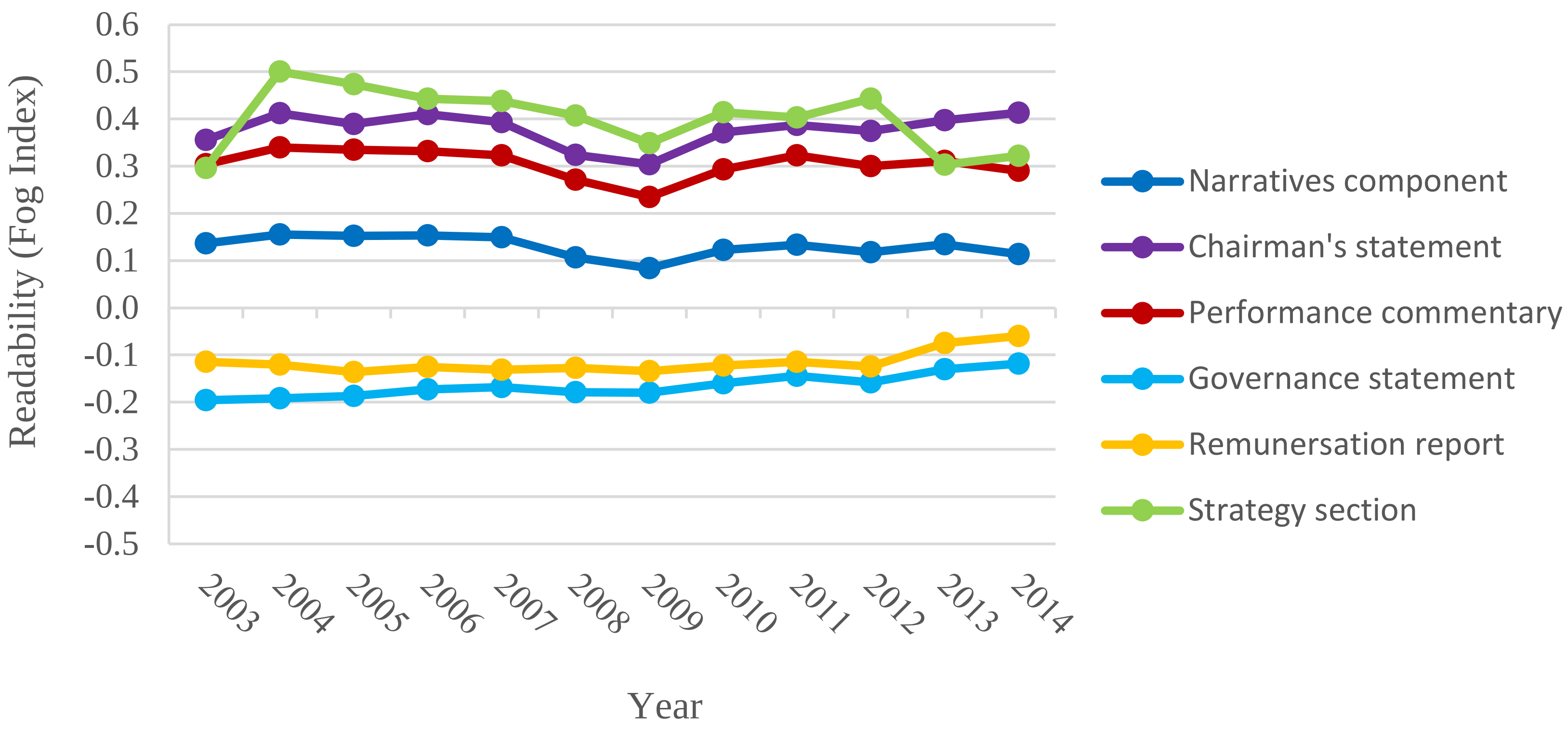
How some of the main types of content have grown over the sample period?



Readability (inverse Fog)



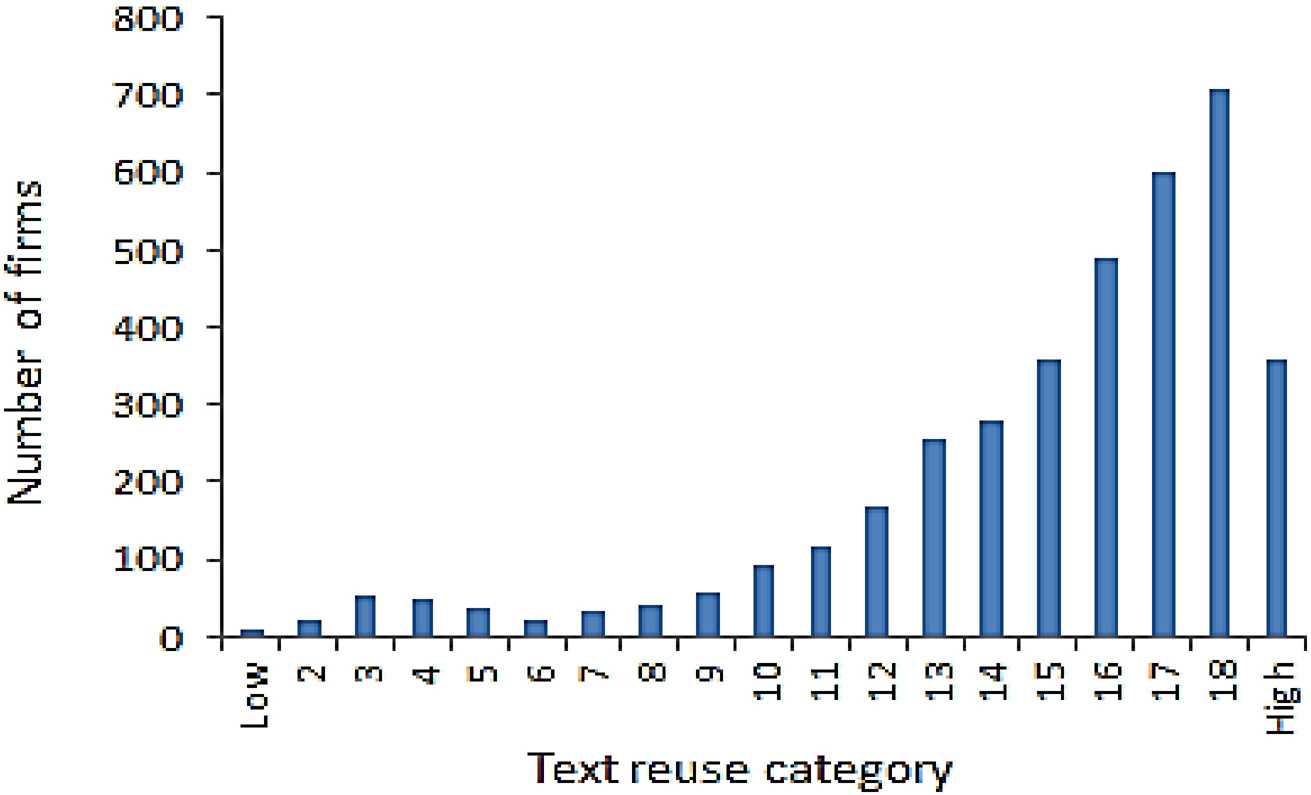
Net Tone



Readability (Fog Index)
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-0.1
-0.2
-0.3
-0.4
-0.5
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
Year
Narratives component
Chairman's statement
Performance commentary
Governance statement
Remunersation report
Strategy section

Text Reuse (Boilerplating)

Governance statements: Distribution of similarity



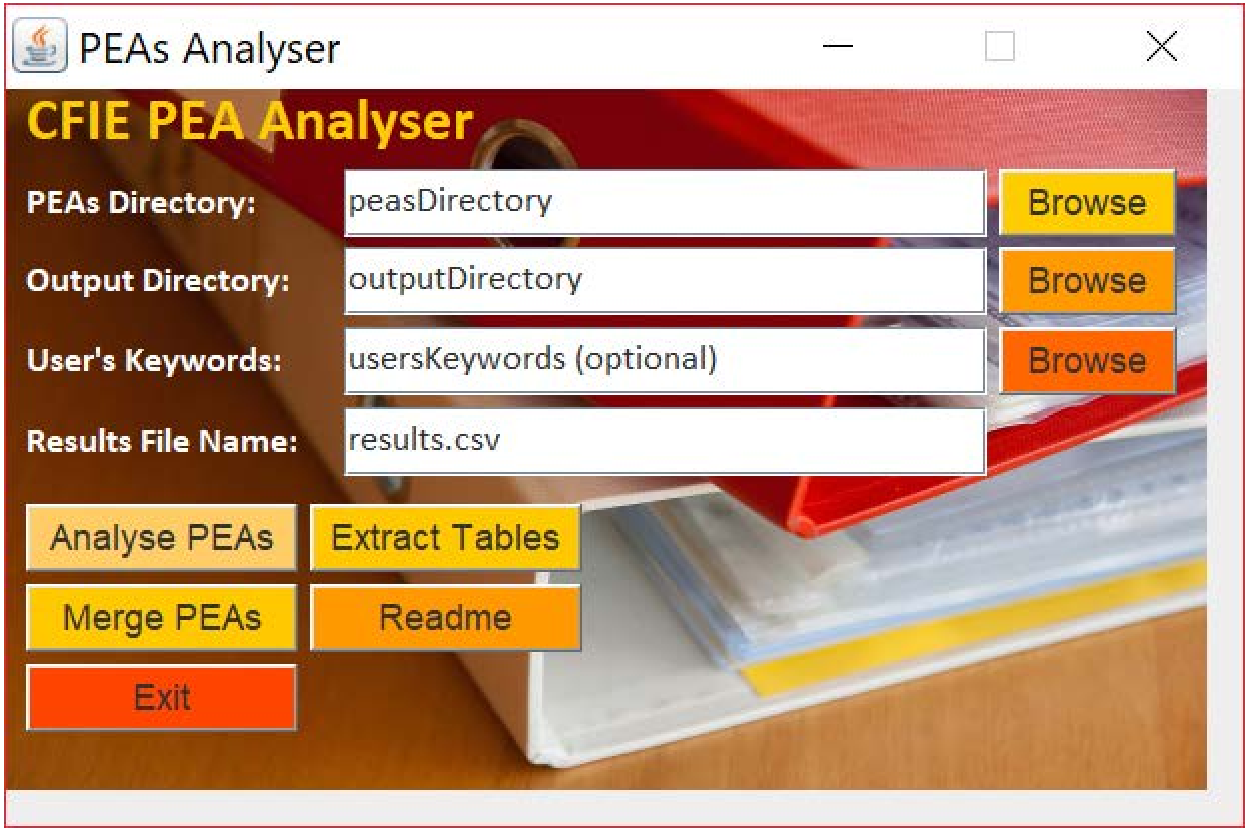
Limitations

Designed for large sample analysis

- Accuracy of extraction process estimated to be around 95%
- Extraction process can result in errors at the individual firm level (which are removed or diluted in large sample work)
- Cannot capture the richness of *how* information is presented (e.g., graphics, charts, tables, etc.)
- No attempt (yet) to parse text in the financial statements

Coming Up

Preliminary Earning Announcement Analyser



PEAs Analyser

CFIE PEA Analyser

PEAs Directory: peasDirectory

Output Directory: outputDirectory

User's Keywords: usersKeywords (optional)

Results File Name: results.csv

Summary & Conclusions

- Develop and validate a method for extracting the content and structure of UK annual reports published as PDF files
  - Retaining the structure of the report creates opportunities for research examining new features of disclosure

- Provide the first large-sample evidence regarding the predictive ability of UK annual report narratives
  - Narratives in their entirety are incrementally predictive for earnings beyond financial statement data
  - Difference annual report sections are associated with different predictive qualities
  - Abnormal managerial optimism is associated with lower predictive ability and independent chairman commentaries help to negate managerial bias

Thanks

More about the projects: http://ucrel.lancs.ac.uk/cfie/

CFIE-FRSE - WEB: https://cfie.lancaster.ac.uk:8443/

CFIE-FRSE – Desktop: https://drelhaj.github.io/CFIE-FRSE

Machine Learning: https://github.com/drelhaj/MachineLearning

FNP 2018 Workshop: http://wp.lancs.ac.uk/cfie/fnp2018

Part 2

CFIE-FRSE HANDS-ON DEMO

Hands-on Demo

CFIE-FRSE - WEB:

https://cfie.lancaster.ac.uk:8443/

Sample Annual Reports:

http://bit.ly/2n2sqcY

Or

http://bit.ly/2nfza8u

Wmatrix Tutorial:

http://ucrel.lancs.ac.uk/wmatrix/tutorial/