Financial **N**arrative **P**rocessing: a Corpus-based NLP approach

Mahmoud El-Haj

CASS, Lancaster University

Research Projects

CorpComm	CFIE	SenseSourcing	VardSourcing
• ESRC	• ESRC	• ESRC/UCREL	• Jisc

Arabic Dialects	Strategic Contents
• UCREL	• ICRA/UCREL/LUMS

Financial Narrative Projects

| Corporate Financial Information Environment (CFIE) | → | Understanding Corporate Communications (CorpComm) | → | Extend for 18 months |

Project Team

CorpComm [1]

CFIE [2]

CASS
Prof Tony McEnery[1]

Dr Mahmoud El-Haj[1,2]

Dr Andrew Hardie[1]

SCC
Dr Paul Rayson[1,2]

LUMS
Prof Steve Young[1,2]

MBS
Prof Martin Walker[1,2]
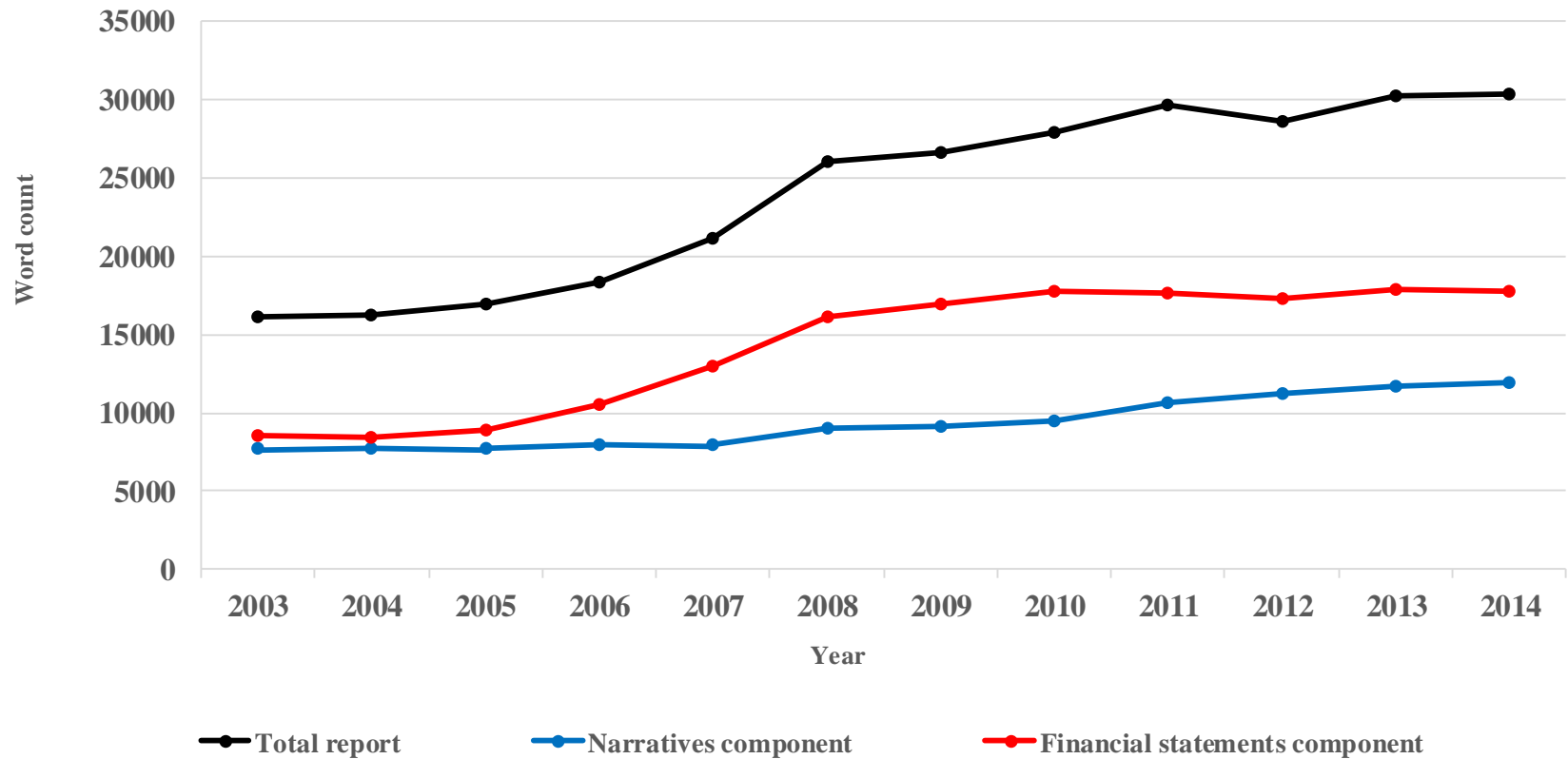
Dr Thomas Schleicher[2]

LSE
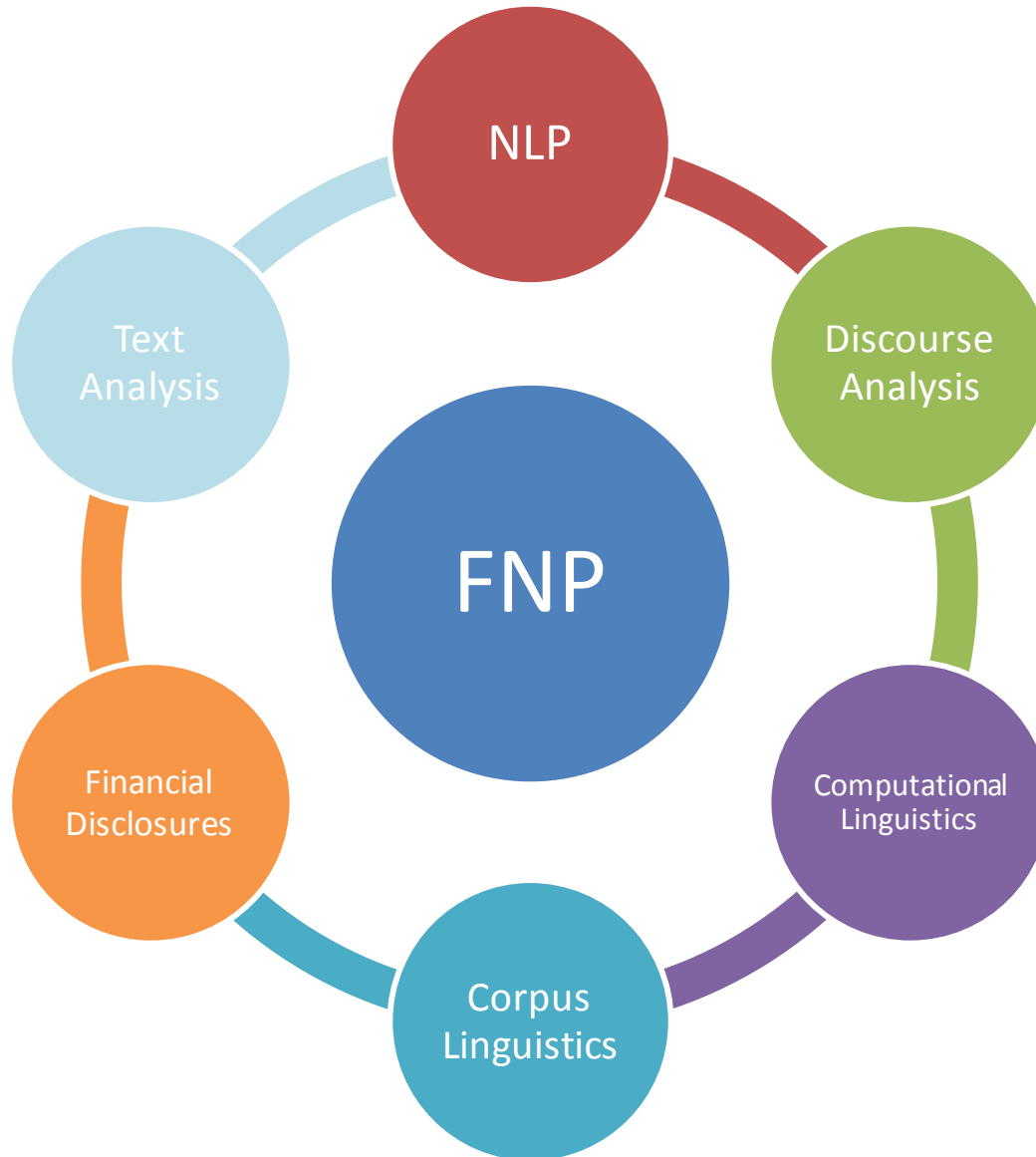Dr Vasiliki Athanasakou[2]

What is it about?

Project analyses U.K. financial narratives, their association with financial statement information, and their informativeness for investors.

Why Financial Narratives

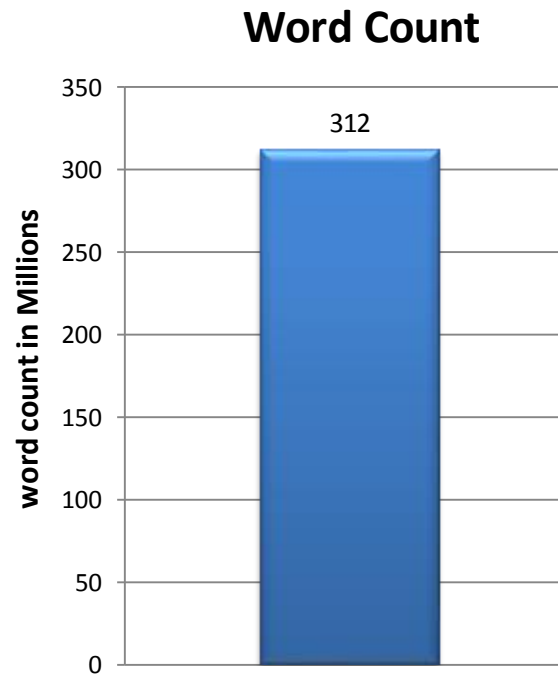a doubling of the median word count over the sample period
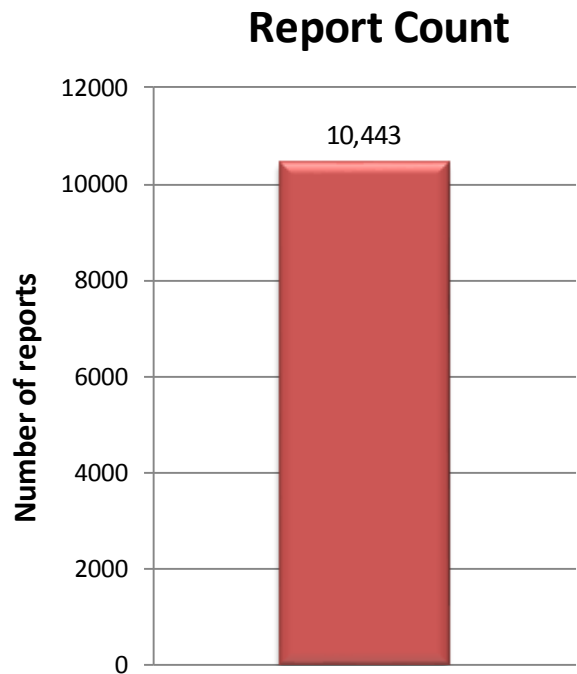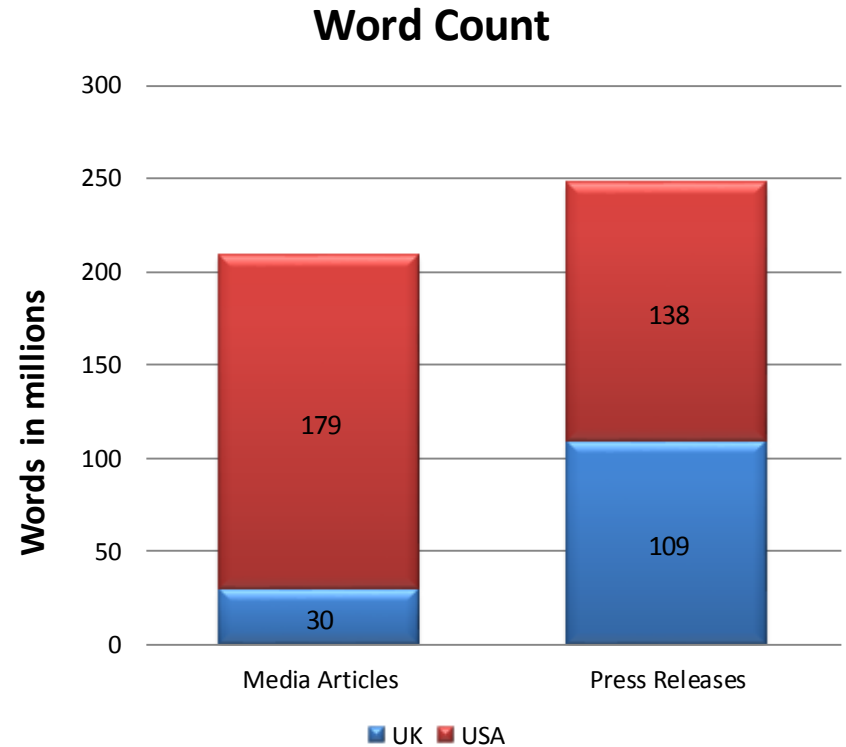
FNP: Fields of Study

Big Data

PDF

XML

JSON

DB

Financial Narratives

Annual Reports

PEAs

Conference Calls

Financial News

Press Releases
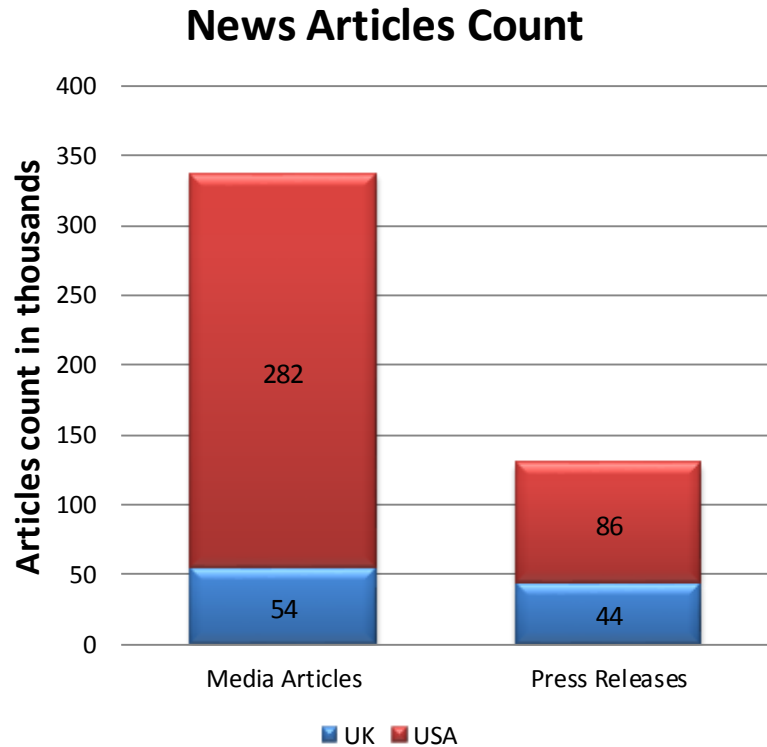
Media Articles

TMX

HTML

RTF

PLAIN
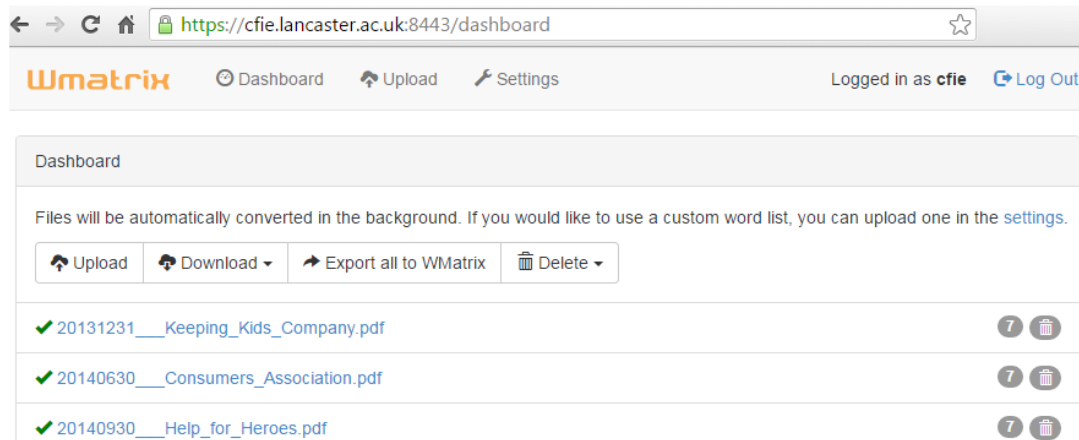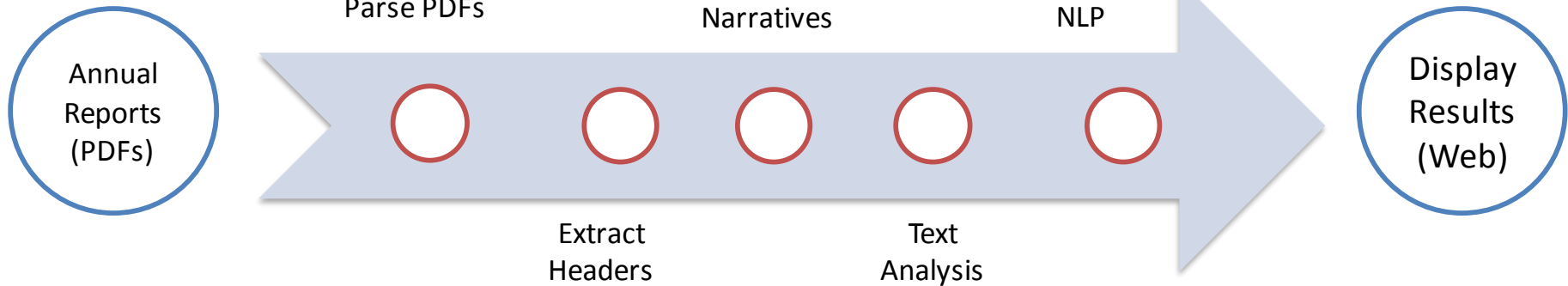
Annual Reports: how big?

Financial Media: how big?

Output

- NLP publicly available tool (Wmatrix-Import)
- 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 understand corporate financial decisions and corporate financial performance.

Wmatrix

Wmatrix-Import

Parse PDFs

Extract
Narratives

NLP

Annual
Reports
(PDFs)

Display
Results
(Web)

Extract
Headers

Text
Analysis

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

Financial highlights

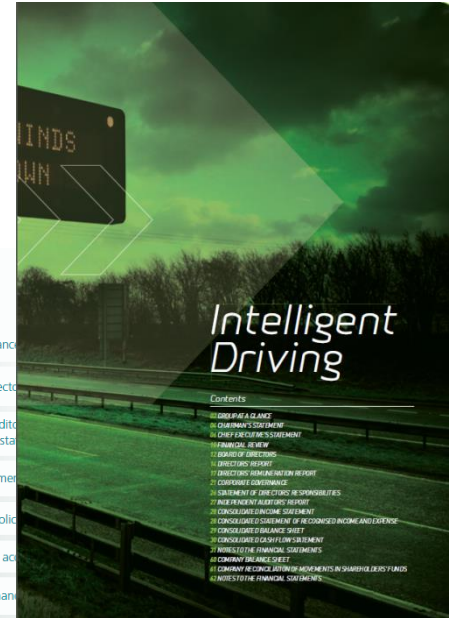| Turnover £000	Operating profit £000	Dividend	Debt levels £000
£54,570 +7.2%	£3,277 +2.8%		
2010: £50,912	2010: £3,188		
2009: £51,521	2009: £3,235		
2008: £47,065	2008: £1,521		
2007: £47,068	2007: £4,206		
	(before exceptional operating)		

Carbon report

Carbon resulting from fuels used for combustion, owned transport, process emissions (Tonnes of CO₂)	Carbon emissions from purchased electricity (Tonnes of CO₂)
2,324	1,060
2011: 2,324	2011: 1,060
2010: 2,285	2010: 1,181

Emissions attributable to specific areas of the business

Production (Tonnes of CO₂ per barrel produced)	Distribution (Tonnes of CO₂ per barrel delivered)
0.0184	0.0077
2011: 0.0184	2011: 0.0077
2010: 0.0190	2010: 0.0078

Content

ARRIVA

Intelligent Driving

UK Annual report tool: Extraction

- Steps in extraction process:

  - Detect contents page
  - Parse contents page
  - Extract section
  - Detect section type
  - Reorder section

Contents

Front

02	Chairman's Statement
05	Directors' Report
09	Directors' Remuneration Report
14	Corporate Governance Report
18	Auditors' Report

Rear

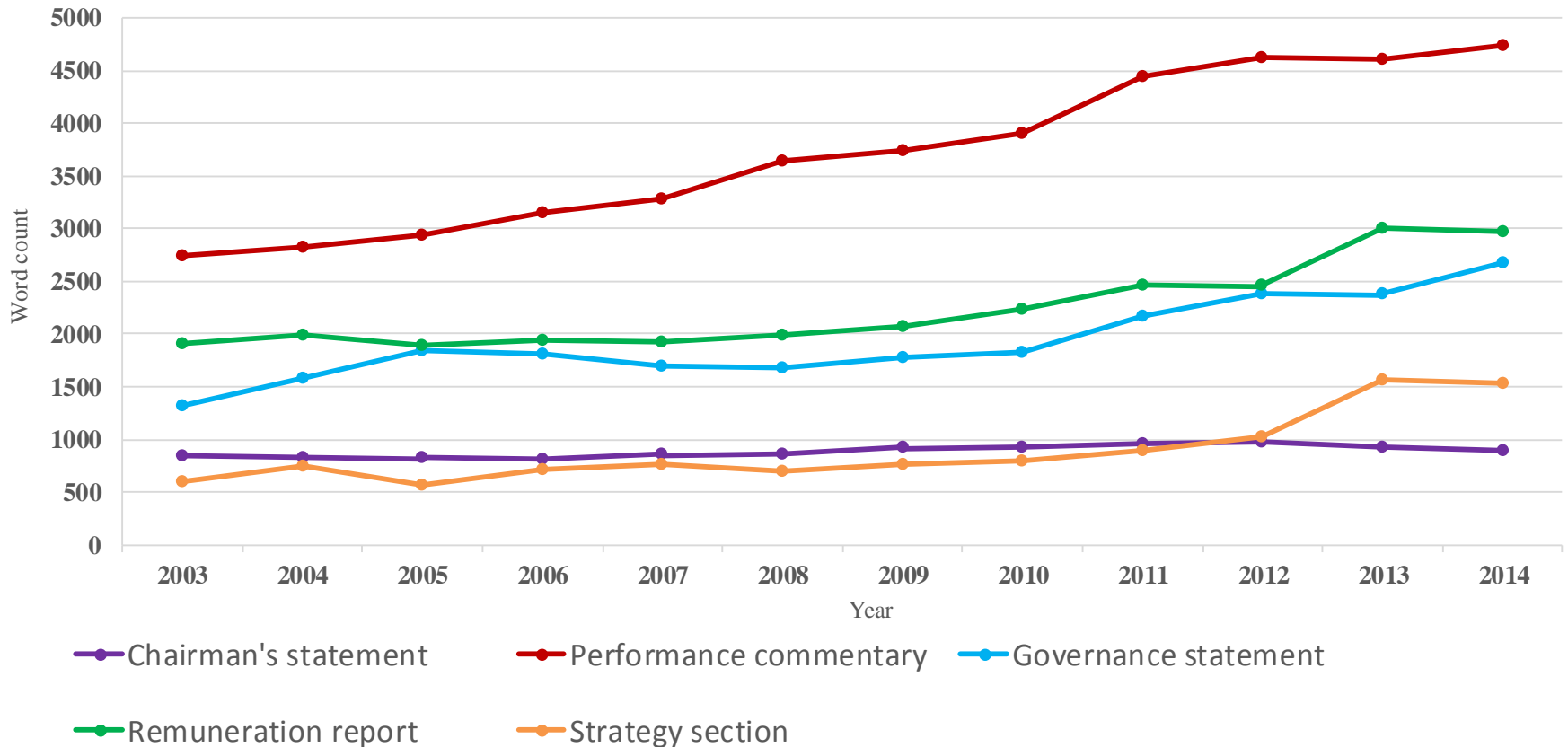|	
20	Consolidated Profit and Loss Account
21	Consolidated Balance Sheet
22	Company Balance Sheet
23	Consolidated Cash Flow Statement
24	Notes to the Financial Statements
36	Notice of Annual General Meeting
39	Directors and Advisors

Techniques

- Automatically detect 8 key headers
- Reorder headers to front and rear
- String matching, regex
- Levenshtein Distance / text similarity
- Machine learning to classify headers
- Crowdsourcing
- Text Reuse
- Hand crafted keyword lists by experts
- Bottom up and top down word lists:
  - Computer based approaches
  - By Linguists

Growing over time

how some of the main types of content have grown over the sample period

Collaboration

FRC	• Funding • Help standardise UK annual reports
Financial Firm, London	• Fraud Detection • Management Change • Management vs Analysts
Financial Services Provided, London	• Interest in the tool • Automatic Structuring • Indexing Annual Reports
Big Four in Accounting	• Interest in the tool • Provided free-to-publish annual reports
Big Data company, London	• Interest in working together
KSU, Saudi Arabia	• Arabic Financial Narratives (Book Chapter)
Sunway University, Malaysia	• Malay Financial Narratives (consultancy)
National Natural Science Foundation of China (NSFC)	• Chinese Financial Narratives (consultancy)

Thanks

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

Wmatrix-Import: https://cfie.lancaster.ac.uk:8443/