Chapter 1

Multilingual Financial Narrative Processing: Analysing Annual Reports in English, Spanish and Portuguese

Mahmoud El-Haj[*], Paul Rayson[**], Paulo Alves[***], Carlos Herrero-Zorita[****] and Steven Young[*****]

This chapter describes and evaluates the use of Information Extraction and Natural Language Processing methods for extraction and analysis of financial annual reports in three languages: English, Spanish and Portuguese. The work described retains information on document structure which is needed to enable a clear distinction between narrative and financial statement components of annual reports and between individual sections within the narratives component. Extraction accuracy varies between languages with English exceeding 95 %. We apply the extraction methods on a comprehensive sample of annual reports published by UK, Spanish and Portuguese non-financial firms between 2003 and 2014.

1. Introduction

Companies use a number of different methods to communicate with their shareholders and investors and to report to the financial markets. These include annual financial reports, quarterly reports, preliminary earnings announcements, conference calls and press releases. Much previous research has focussed on the quantitative numerical elements of these reports. In addition, researchers in accounting and finance have been able to carry out small scale manual analyses of the narrative textual elements of these reports for many years, but recently a trend has emerged of applying automatic natural language processing techniques to enable replication of these earlier studies on a much larger scale, as well as to improve the accu-

[*]Corresponding author: School of Computing and Communications, Lancaster University, UK
[**]School of Computing and Communications, Lancaster University, UK
[***]Universidade Católica Portuguesa, Portugal
[****]Department of Linguistics, Universidad Autónoma de Madrid, Spain
[*****]Management School, Lancaster University, UK

racy and depth of the metrics that are studied. Much of the previous text mining work has been performed in the US context where annual 10-K filings are required to follow a rigid structure with a standard set of headings, and are written in plain text. This enables more straightforward selection of relevant sections for further analysis. In contrast, in the UK and elsewhere, annual report structure is much less rigid and companies produce glossy brochures with a much looser structure, and this has prevented large-scale long-term narrative research until recently. In this paper, we describe not only the structure detection and extraction process that we have designed and implemented for English annual reports, but also our initial work to extend this research to other national contexts, in this case to Spain and Portugal. We report on our experiments to port the system from UK annual report analyses to those published in Spanish and Portuguese, and describe the adaptations made to the system to enable this. Our resulting software is made freely available for academic research.

2. Related Work

Previous related work on financial narrative analysis has taken place in a number of areas including accounting and finance research, natural language processing and corpus linguistics. Some early approaches in the accounting and finance literature employed painstaking manual approaches and were therefore limited in scale due to time constraints. Further studies have become larger scale but are still using manually constructed word lists for detecting features without considering local context for disambiguation purposes or more advanced machine learning methods. Well known studies include one by Feng Li in 2010[1] which considered forward-looking statements in 10-K and 10-Q filings in the US and found a link between positive statements and better current performance and other indicators. Li also found that general content analysis dictionaries such as Diction, General Inquirer and LIWC) are not helpful in predicting future performance. Loughran and McDonald[2] also found that negative words in the general purpose Harvard Dictionary were not typically considered as negative in financial contexts, and so were less appropriate than domain specific versions. They also considered US 10-K reports for their study. Schleicher and Walker[3] found that companies with impending performance downturns bias the tone in outlook sections of the financial narrative. A good survey of text analysis methods in accounting and finance research was also recently published by Loughran and McDonald.[4]

In the natural language processing research area, previous research has been carried out to extract document structure mainly from scientific articles and books.[5–7] Other than this, there has been much recent work in using text mining and sentiment analysis in particular to Twitter with the goal of predicting stock market performance[8–12] although presumably any really successful methods would not be published.

From the other end of the language analysis spectrum, in linguistics, there has been a large amount of research on the language of business communication. Merkl-Davies and Koller[13] introduced the Critical Discourse Analysis (CDA) approach to the accounting community and showed how it can be used to systematically analyse corporate narrative documents to explore how grammatical devices to obfuscate and guide interpretations. Brennan and Merkl-Davies[14] considered communication choices and devices which contribute to the phenomena of impression management, where individuals or companies use language to present themselves favourably to others.

In terms of the context for financial narrative analysis in other countries, as some authors explain,[15–19] apart from the economic data, information explaining the intellectual capital, organisation and human activities and resources is key for a company's visibility on the market and transparency of information with shareholders. This has driven Spanish entities in recent years to voluntarily include them in annual reports and especially sustainability reports.[19] Oliveras et al.[20] show that between 2000 and 2002 among the 12 most important Spanish companies this type of information has increased significantly, particularly intellectual, human and structural capital a claim supported also by Villacorta.[21] Tejero Romero[18] analyses annual reports between 2004 and 2008 located in companies' websites and also claims there has been an increase, more especifically concerning management control and network systems followed by research. This has been more significant in entities related to technology and communications and construction.

3. Dataset

We collected annual reports from UK, Spanish and Portuguese large firms. For UK annual reports, we gathered more than 10,000 annual reports for non-financial firms listed on the London Stock Exchange for the years in the range 2003 and 2014. The extraction methods have been tested and evaluated on English annual reports and were later adapted to work

with other languages. For Spanish, we gathered one hundred annual reports from the biggest companies of the country from the year 2015. The selection criteria was made using the Orbis international Database, focusing only on the very large businesses and excluding those with no recent financial data as well as banks and public authorities/governments. The query returned a list of 9,126 companies. The next step was to manually retrieve the annual reports of the top 100 companies from their public web pages. Our Portuguese annual report sample consists of 576 reports, issued by 77 firms, for the period 2006-2015. All firms are listed on the Portuguese Stock Exchange. The annual reports were collected automatically from Perfect Information[a].

Problems of collecting and distributing this data

This initial step proved to be problematic, since not every company had made available their annual report on their websites. Spanish legislation obliges companies to display their financial statements and corporate governance reports, but not the narrative annual reports. Therefore, the decision to provide a report relies entirely on the company, which will either present it freely at their website along with the financial statements, display it in a restricted section only available to shareholders, or not show it. This issue becomes more frequent among subsidiaries or branch companies that belong to a bigger group, which, if international, may only present its annual report in English. Overall, reports were found in nearly 1 in 7 companies from the Orbis Database. In order to obtain 100 documents, 638 web pages were accessed.

Description of Dataset

Before describing the dataset it is worth explaining what is an annual report. An annual report is a comprehensive report on a company's activities throughout the preceding year. Annual reports are intended to give shareholders and other interested people information about the company's activities and financial performance. They may be considered as grey literature.

It was not until legislation was enacted after the stock market crash in 1929 that the annual report became a regular component of corporate financial reporting. Typically, an annual report will contain the following

[a]www.perfectinfo.com

sections:

- Financial Highlights
- Letter to the Shareholders
- Narrative Text, Graphics and Photos
- Management's Discussion and Analysis
- Financial Statements
- Notes to Financial Statements
- Auditor's Report
- Summary Financial Data
- Corporate Information

The annual reports dataset files are all in PDF format, and variation in formatting makes it a challenge for automatically extracting and detecting structure. The annual reports vary in respect to their style and number of pages. In contrast to the US, stock exchange-listed firms in UK, Spain and Portugal do not present their financial information and accompanying narratives in a standardised format when creating annual reports. Firms in the aforementioned countries have much more discretion regarding the structure and content of the annual report. Added to this is the problem of nomenclature: no standardised naming convention exists for different sections in UK annual reports so that even firms adopting the same underlying structure and content may use different terminology to describe the same section(s).

Table 1 shows the dataset size in words in addition to the number of reports for each language.

Table 1. Dataset Size

Language	Reports	Words
English (UK)	11,009	300M
Portuguese	396	7.50M
Spanish	100	2.40M

4. Extraction Methods

We used Information Extraction (IE) and Natural Language Processing (NLP) methods to detect the structure of the annual reports and extract sections and their narratives. The methods automatically detect the annual

report's table of contents, synchronise page numbers in the native report with page numbers in the corresponding PDF file, and then use the synchronised page numbers to retrieve the textual content (narratives) for each header (hereinafter section) listed in the table of contents. Section headings presented in the table of content are used to partition retrieved content into the audited financial statements component of the report and the "front-end" narratives component, with the latter sub-classified further into a set of generic report elements including the letter to shareholders, management commentary, the governance statement, the remuneration report, and residual content.

4.1. Structure Extraction Process

In this section, we discuss in detail the process of detecting the structure for UK, Spanish and Portuguese annual reports. As mentioned in Section 3 we processed more than 10,000 UK annual reports in PDF file format and used the same methods at a later stage to analyse a smaller sample of Spanish and Portuguese annual reports.

Unlike the US Stock exchange, firms in the UK do not follow a standard reporting template when writing annual reports. Firms and management in the UK have more discretion regarding the format, structure and the contents of the annual reports. On the other hand the US Securities and Exchange Commission forces firms to follow a standard format and a pre–labeled annual reports template which they publish in HTML file format. This has helped in creating a reporting standard making it easy for investors, firms and analysts to access and acquire information automatically from a bulk of annual reports. This is different in the UK where firms tend to publish their annual reports in PDF file format. Despite being cross-platform and a portable file format it is deemed a difficult task to automatically extract information from PDF annual reports since companies' reports vary significantly especially when it comes to the contents and the section headers. In order to automatically analyse a large dataset of UK annual reports we first needed to automatically detect the structure of the PDF annual reports so we can extract the information needed.

To detect and extract the structure of the annual reports each PDF file goes through the following five steps 1) detecting the contents-page, 2) parsing the detected contents-page and extracting the sections, 3) detecting page numbering, 4) adding the extracted sections to the annual report PDFs as bookmarks, and 5) using the added bookmarks to extract the narrative

sections under each heading.

4.1.1. Detecting the Contents Page

An annual report contents page includes information about the main sections of the report and its associated page numbers. Information in the contents page helped us detect the structure of the annual reports. However, detecting the contents page was not a straightforward task. We created a list of gold–standard section names extracted manually from the contents page of a random sample of 50 annual reports. We filtered the gold–standard keywords by removing duplicates and preserving the structure of how they appeared in the annual reports. We matched each page in the annual report against the list of section names in gold–standard, then we 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. a section name followed by a page number, such as "Chairman's Statement 13").

4.2. Parsing the Contents Page

In order to get the structure of the annual report we automatically parse the selected contents page by extracting each section and its associated page number. To do this we matched each line of text in the selected contents page against a regular expression commands that will extract any line starting or ending with a number between 1 and the number of pages of the annual report.

We built a simple filtering tool that filters out any block of text that matches our regular expression commands. This is done by removing text containing addresses, dates, and postal codes. The filtering tool can also detect email addresses, websites, references to branches and locations using regular expression commands and a gazetteer.

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 Bothwell Road) might still be confusing for the tool. We tackled this problem by matching the list of extracted sections against a list of gold–standard section synonyms which we explain in more details in Section 4.5.

The structure of the PDF files makes it difficult to extract text in its

actual format. Extracting plain text from PDFs results in many line breaks being added in between the text. This makes extracting a section that is split into two lines a difficult task. To tackle the problem of broken sections (i.e., sections appearing on two lines or more), we implemented an algorithm to detect broken section headers and fix them by concatenating lines 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. This method was also adapted to Spanish and Portuguese prepositions and other stop words needed to concatenate lines of text by forming a list of most common stop words for each language.

4.3. Detecting Page Numbering

The page numbers appearing on the contents page do not usually match with the actual page numbers in the PDF files. For example, page 4 in the annual report could refer to page 6 in the PDF file, which may lead to incorrect extraction[b]. We address this problem by creating a page detection tool that crawls through annual report pages taking three consecutive pages in each iteration. The tool aims to extract a pattern of sequential numbers with an increment of 1 (e.g. 31, 32, 33) but with the complex structure of the PDF files this has been proven to be a difficult task. The tool starts by reading the contents three pages at a time starting from the report's number of pages minus one. For example, assume we are trying to detect the page numbering pattern for a report of 51 pages. The tool starts by extracting text from pages 48, 49 and 50. A regular expression command is then used to extract all the numbers in each page contents that is made up of maximum three digits creating a vector of numbers for each page. Figure 1 shows a sample of 3 vectors for the pages 48, 49 and 50. As shown in the Figure the algorithm will only keep numbers that are within a range of 10 pages those linked with a green arrow. The algorithm will then try to form a pattern of sequential numbers with an increment of 1. The Figure shows that the pattern 49, 50 and 51 has been found which is equivalent to a one page difference (*page–increment*) between the reports page numbering and those found in the PDF file. The tool will repeat the same process for all the pages in the annual report until it reaches pages 1, 2 and 3 where it stops.

[b]The algorithm responsible for extraction sections text uses start and end page numbers to locate the text and therefore accurate page numbers are required.

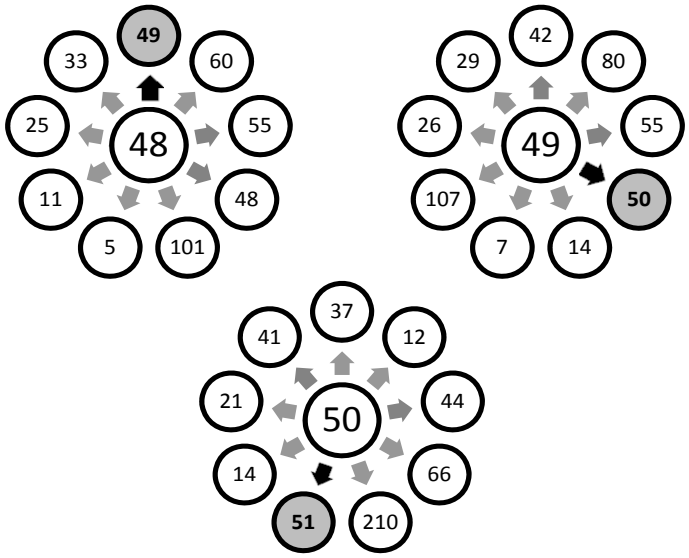

Figura 1. Detecting Page Numbering

As shown in Figure 2 for each 3 vectors the tool will store the page–increment in an array of numbers and at the end of the process the most popular (most frequent) page–increment will be selected as the difference between the annual report and the PDF numbering.

This process on the sample yielded an accuracy rate of more than 95 %. Manual examination of the remaining less than 5 % revealed the following reasons for non-detection: 1) encoding, 2) formatting and 3) design.

4.4. Adding Section Headers as Bookmarks

Using the sections and their correct page numbers from Sections 4.1.1 and 4.3 we implemented a tool to insert the extracted contents page sections as bookmarks (hyperlinks) to sample PDFs. This process helped in extracting narratives associated with each section for further processing (see Section 4.5 below).

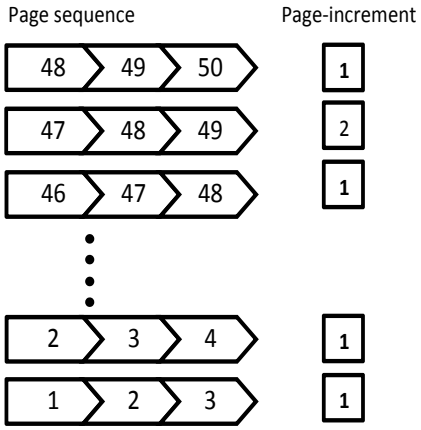

Figura 2. Popular Page Increment

4.5. Extracting Sections' Narratives

We implemented an automatic extraction algorithm to crawl through the data collection and, for each PDF file, 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 section header. For example the section header "Chairman's Statement" may also appear as "Chairman's Introduction", "Chairman's Report" or "Letter to Shareholders". The same case applies to Spanish and Portuguese as well. For example Spanish section headers "Carta del presidente", "Informe del presidente", "Carta al accionista", and "Mensaje del presidente" all refer to the "Chairman Statement" section. To solve this problem, we semi-automatically and by the help of experts in accounting and finance, created a list of synonyms for each of the generic annual report sections (see the list below). This was done by extracting all sections 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 experts). We refined the list by removing redundancies. The accounting experts then manually examined the list and deleted irrelevant or incorrect sections. We used the refined list as gold–standard synonyms to extract all the sections related to each of our generic sections (e.g. all sections about the "Chairman's Statement"). To overcome the problem of different word–order or additional words included in the headline (e.g. "The

Statement of the Chairman") we used *Levenshtein Distance* string metric algorithm[22] to measure the difference between two sections. 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 sections with a Levenshtein distance of up to five were presented to the accounting expert.

We used the above process to create gold–standard synonym lists for the following 11 generic section headers that we wished to extract for further analysis:

1. Chairman Statement
2. CEO Review
3. Corporate Government Report
4. Directors Remuneration Report
5. Business Review
6. Financial Review
7. Operating Review
8. Highlights

Having detected and extracted section headers (or their gold–standard synonyms) and their sections, we then extract the sections' narratives using iText[c], an open source library to manipulate and create PDF documents,[23] to apply our text analysis metrics, which include readability measurement and counting word frequencies using financial domain hand–crafted word lists.

5. Extraction Tools

We used the extraction methods described in Section 4 to create publicly available web and desktop tools for users to automatically and freely analyse annual reports in different languages. The tools deal with multilingual annual reports of firms within the UK, Spain or Portugal written in either English, Spanish or Portuguese and distributed in PDF file format[d].

The tool is called CFIE-FRSE standing for Corporate Financial Information Environment (CFIE) -Final Report Structure Extractor (FRSE). The tool is available as a web application via `https://cfie.lancaster.`

[c]http://itextpdf.com/api

[d]For now only the Desktop version of the tool can work with multilingual annual reports

ac.uk:8443 or as desktop application, which is freely available on GitHub https://github.com/drelhaj/CFIE-FRSE. The tools detect the structure of annual reports by detecting the key sections, their start and end pages in addition to the narrative contents. This works for all three languages. The tools provide more analysis for reports written in English such as readability metrics, sections classification and tone score. This is because the tool was built to analyse UK annual reports where we have a large dataset to train the system to provide an extra level of analysis.

The extra level of analysis will be made available for Spanish and Portuguese at a later stage. For now we do not have enough reports for both languages to be able to train the system. As explained earlier the aim of this chapter is to show that our extraction methods can be applied to Spanish and Portuguese, a vital step towards fully analysing reports in these two languages and other languages later in the future.

6. Multilingual Extraction

In this section we explain the process we followed to extracting sections from annual reports in each of the three languages: English, Spanish and Portuguese.

6.1. English

As mentioned earlier the work was first designed to analyse UK English annual reports.[24] We automatically harvested more than 10,000 annual reports for firms listed on the London Stock Exchange (LSE). Prior to analysing the annual reports we first worked on sorting them by firm and we created our own identifier which we called "LANCS_ID". Sorting annual reports was done semi-automatically were we used a Java tool to match firm names and extract the reports' years. This was followed by a manual post editing to make sure the matching was correct. Firms that did not match with any of the could be firms that do not exist anymore or firms with a new name due to merging with another firm, those had to be manually matched[e].

[e]PDF filenames do not contain a unique firm identifier. For example, reports collected from Perfect Information use a standard naming convention comprising firm name and publication year. We use filenames as the basis for a fuzzy matching algorithm that pairs firm names extracted from the PDF filename with firm names provided by Thomson Datastream. Matching on name is problematic because firms can change their name over the sample period. The matching procedure must therefore track name changes. To address this problem, we combine firm registration numbers and archived names from the

Annual report structures vary significantly across reporting regimes and therefore to make the initial development task feasible we focus on reports for a single reporting regime. We select the UK due to the LSE's position as largest equity market by capitalisation outside the US. The extraction process is nevertheless generic insofar as reports published in other reporting regimes and languages can be analysed by modifying the language- and regime-dependent aspects of our tool without editing the underlying java source code. Further guidance will be provided in an online appendix, together with full technical details of our method, in due course.

Table 2 shows the structure detection and extraction accuracy for UK annual reports.

Table 2. UK Annual Reports Analysis

Number of downloaded annual reports	11,009
Number of reports analysed	10,820
Percentage of correctly retrieved table of contents	98.28 %
Percentage of correctly retrieved pages	95.00 %
Percentage of correctly retrieved text from sections	95.00 %

As shown in the table the tool analysed more than 98 % of the downloaded annual reports. Firms management in the UK have more discretion over what, where, and how much information on topics such as risk, strategy, performance, etc. is reported, this lead the reports to vary significantly in terms of structure and design. Despite the dissimilarity between the structure of the downloaded annual report, our methods were able to accurately analyse the majority of the reports. Those failing the analysis process were due to one of the following reasons:

1. The file does not allow the text to be extracted (image-based documents). This problem is more common in the early years of our sample (i.e. 2000-2005), as some of the annual reports were poor quality scanned files. These types of reports have virtually disappeared.
2. Reports with a table of contents that could not be read due to the limitation imposed by how the table was designed. For example where

London Share Price Database with Datastream's firm name archive in our fuzzy matching algorithm. For those cases where our algorithm fails to find a sufficiently reliable match, we perform a second round of matching by hand. Further details of the matching procedure, including a copy of the algorithm and a step-by-guide to implementing the matching procedure in SAS are available at `http//cfie.lancaster.ac.uk.8443/`. Licensing restrictions prevent direct publication of proprietary identifiers.

a table of contents is designed with numbers and text in two different columns, or where the table of contents is split into two pages.
3. Absence of page numbers.

6.2. Spanish

The analysis of the Spanish reports will deal mainly with four challenges: (1) the difficulty of retrieving the documents from the companies; (2) the lack of a standard and common structure; (3) the high amount of variation in the headers of the tables of contents, due to the previous point as well as linguistic reasons; and (4), the amount of noise in the PDFs.

We manually collected 100 annual reports from the biggest firms in Spain for the year 2015. The selection criteria was made using the Orbis international Database[f], focusing only on the very large businesses and excluding those with no recent financial data as well as banks and public authorities/governments. The query returned a list of 9,126 firms. The next step was to manually retrieve the annual reports of the top 100 firms from their public web pages.

This initial step proved problematic since not every company made available the annual reports on their websites. Spanish legislation obliges firms to display their financial statements[g] and corporate governance reports, but not the annual reports[h]. Therefore, the decision to provide a report relies entirely on the company, which will either present it freely at their website along with the financial statements, display it in a restricted section only available to shareholders, or not show it at all. This issue becomes more frequent among subsidiaries or branch companies that belong to a bigger group, which, if international, may only present its annual report in English. Overall, reports were found in nearly 1 of each 7 companies from the Orbis Database. In order to obtain 100 documents, 638 web pages were accessed.

Following this, we manually extracted all the sections from the tables of contents of the reports, creating a keyword list of 1,503 tokens. This will be fed to the CFIE-FRSE program later on in order to extract and process the text.

The second challenge brought by Spanish reports that will undoubtedly influence their automatic processing has to do with their lack of standar-

[f]https://www.bvdinfo.com/en-gb/our-products/company-information/international-products/orbis
[g]According to Title 7 of Corporation Law.
[h]Art. 538 and 540 of Corporation Law, available in the link above.

dised structure. In other words, Spanish legislation does not provide guidelines for the development of these documents, leaving each company to design them as they see fit. The structure of today's reports are the result of an evolution over the years, adapted to the shareholders' demands and information the companies have considered relevant to the public, some of them based on overseas countries[i].

There are, nevertheless, some common points regarding their composition. The typical sections include a letter from the CEO, lineup of the owners of the company, a business review of the last year and an analysis of the future including GRI parameters. Table 3 shows the 50 most frequent items from our table of contents keyword list:

Table 3. Most frequent sections in Spanish annual reports.

Section (Spanish)	Section (English Translation)	Frequency
Carta del Presidente	Letter from the Chairman	58
Gobierno Corporativo	Corporate Governance	26
Anexos	Appendix	19
Consejo de Administración	Board of Directors	14
Sociedad	Society	12
Principales Magnitudes	Main Events	12
Modelo de Negocio	Business Model	12
Estrategia	Strategy	12
Responsabilidad Social Corporativa	Corporate Responsibility	11
Recursos Humanos	Human Resources	11
Medio Ambiente	Environment	11
Informe de Gestión	Financial Statement	11
Órganos de Gobierno	Board of Directors	10
Informe de Auditoría	Auditor's Report	10
Clientes	Clients	10
Acerca de este Informe	About this Report	10

The most frequent items seem to show these aspects. Carta del Presidente, the 'letter of the chairman', appears in 58 documents. In second place, in 29 documents, is the 'board of directors' (gobierno corporativo) followed by the annexes (which seldom contain the GRI indicators) and the main events ('principales magnitudes').

The frequencies, however, lead us to the third problem of Spanish reports: a great variation in the sections. For example, the letter of the CEO, 'carta del presidente', appears in 60 documents. This does not mean that 40

[i]https://www.icsa.org.uk/assets/files/free-guidance-notes/contents-list-for-the-annual-report-of-a-uk-company.pdf

reports lack this section, but that the sections are written differently. For example, in the complete frequency list, the letter from the chairman/CEO appears as the following (Table 4 translation in English added by the authors), summing up a total of 100. Instead of 'letter' we may find 'message', 'report' or 'greetings', or sometimes including the name of the company CEO. We can also find the variation of 'chairman' in its female inflection, presidenta. The same happens in the rest of the sections which may oblige us in the future to lemmatise and unify the sections in order to make the tool quicker and more efficient.

Table 4. Chairman and CEO Letters Synonyms.

Section (Spanish)	Section (English Translation)	Frequency
Carta del presidente	Letter from the chairman	58
Carta del consejero delegado	Letter from the CEO	9
Mensaje del presidente	Message from the chairman	3
Mensaje del consejero delegado	Message from the CEO	3
Informe del presidente	Report from the chairman	3
Carta del presidente a los accionistas	Letter from the chairman to the shareholders	3
Carta de la presidenta	Letter from the chairman (female)	3
Carta al accionista	Letter to the shareholder	3
Cartas	Letters	2
Carta de Xabier Etxeberria, director general ejecutivo	Letter from Xabier Etxeberria, CEO	2
Carta del presidente y consejero delegado	Letter from the chairman and CEO	2
Carta del CEO	Letter from the CEO	2
Carta de Ignacio Martín, presidente ejecutivo	Letter from Ignacio Martín, CEO	2

The following step was to prepare the tool to perform the analysis or, in other words, to adapt the English algorithms to work for Spanish. The most important feature was to include alphabetic characters missing in English. Spanish uses the standard Latin writing system but with a series of additional characters such as the stressed vowels with the diacritic acute and umlaut over the letter *u*, the *ñ* (*eñe*) letter and the cedilla letter *ç*. The second was to include a Spanish list of stop-words to deal with line breaking.

Lastly, we performed the analysis of the reports with the CFIE-FRSE program and observed the results, summarised in Table 5.

The program successfully processes the documents and Spanish characters, albeit having some problems that are caused, mainly, by the composition of the table of contents page, the fourth and final challenge encountered in this language. Firstly, it analysed only 65 documents, skipping the re-

Table 5. Spanish Annual Reports Analysis

Number of downloaded annual reports	100
Number of reports analysed	73
Percentage of reports completely analysed	46.60 %
Percentage of correctly retrieved table of contents	91.70 %
Percentage of correctly retrieved pages	53.40 %
Percentage of correctly retrieved text from sections	58.90 %

maining 35. This appears to be caused by a series of reasons: the nature of the generation of the PDF that does not allow the text to be accessed; the table of contents page is not in a good format (e.g. one report featured the pages and the sections in different lines); or it is divided into several pages.

Secondly, almost 70 % of the table of contents page was retrieved properly, but only around 34 % of the documents were analysed completely (good retrieval of pages and text per section). No analysis that failed the table of contents retrieval was correctly analysed afterwards, which indicates the key importance of this first step. These failures occurred in reports that contained a lot of noise or a complex array of the text: pages and sections in different lines, table of contents spanning several pages, sections scattered among the page joined with images, etc. Automatically extracting the raw text from these cases fails and therefore the program cannot proceed. Finally, some PDF documents contain double paging or blank pages that affect the correct automatic retrieval of the pages.

Overall, the CFIE-FRSE program appears to work well for Spanish. However the irregularity of the composition of the annual reports makes it difficult to succeed in every document. In addition to this, the keyword list of sections should be updated and improved in order to deal with the high amount of variation in the sections and make the analysis faster.

6.3. Portuguese

Firms listed in the Portuguese Stock Exchange[j] are required to submit an annual report in Portuguese[k] and, as in most countries, Portuguese

[j]The Portuguese Stock Exchange is a relatively small stock exchange currently listing 52 firms totalling a market value of 60,763 M Euros (53,352 M Sterling Pounds), as of 30th June 2017. Research on the Portuguese market is scarce and tends to focus on disclosures on specific aspects of the annual report such as CEO letters,[25] governance[26] or intangibles.[27]

[k]We found that approximately two thirds of the firms opt to disclose both a Portuguese and an English version of their annual reports. When considering the historical time

market regulations and the Companies Act impose very few requirements concerning the contents of the annual reports and there is no mandatory or recommended structure for the Portuguese annual reports[l]. In addition, firms can voluntarily disclose other information, such as sustainability related information, however firms tend to disclose this type of information via their websites.

Our sample includes 627 reports, issued by 77 firms, for the period 2005-2015. All reports were submitted in an unstructured PDF format (table 6)[m].

Table 6. Number of Reports Per Year

Language	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	Total
Downloaded	51	52	60	61	61	62	64	62	58	59	37	627
Processed	23	26	38	37	38	40	44	43	42	36	29	396
	(45 %)	(50 %)	(63 %)	(61 %)	(62 %)	(65 %)	(69 %)	(69 %)	(72 %)	(61 %)	(78 %)	(63 %)

The tool was able to process 373 (63 %) of the annual reports. An analysis of some of the non-processed reports seems to indicate that the most common problems are related to the following:

1. The file does not allow the text to be extracted (image-based documents). This problem is more common in the first years of our sample, as some of the annual reports were of poor quality scanned files. These types of reports have virtually disappeared.
2. Reports with a table of contents that could not be read due to the limitation imposed by how the table was designed. For example where a table of contents is designed with numbers and text in two different columns, or where the table of contents is split into two pages.

The process of adapting the CFIE-FRSE Desktop tool to the Portuguese language was divided into several steps:

series of the firms currently included in the main index – PSI-20 – the percentage of firms reporting in both languages increases to 93 %.

[l]The Companies Act requires the annual report to include, amongst other items, a review of the firm's activities, performance and financial position, a description of the main risks and uncertainties, a description of subsequent events, the expected evolution of the firm and a proposed net income allocation and dividends.

[m]The Portuguese Stock Exchange Website has a repository of filings where all annual reports are available. However, the interface is not designed to download batches of reports and, therefore, we retrieved all available in Perfect Information in October 2016.

- The detection and extraction of the table of contents
- Aggregate sections into a standard set of pre-defined sections (Chairman's Statement, Performance Review, etc)
- Create dictionaries according to the research question

At this stage, we dealt with the first two steps as the last one varies with the research objectives. The detection and extraction of the table of contents was based on a specific algorithm that looks for expressions that commonly appear as headings in an annual report. For this purpose, we created a list of gold–standards by collecting 67 Reports in Portuguese from different firms and listing all sections in those reports. The initial list contained 2,053 section headers. We then cleaned the list for firm-specific sections and duplicates. The final list had 694 entries, including a considerable number of variations for the same section. For the second step, we analysed the final list and assigned each entry to a pre-defined section: Chairman, CEO, Performance, Auditor, Financial Statements and Other. This standardisation is necessary to deal with the lack of structure in the report and the numerous alternative names for a given section. The Portuguese reports presented some additional challenges, some of which we would like to highlight in this section.

Firstly and as with Spanish, the Portuguese alphabet includes additional characters in comparison to the English language. These additional characters are phonetic modifications of common characters, such as "À", "Á", "Ã", "Â" and "Ç".

Secondly, we had to develop a list of stop-words to deal with the line breaking. In addition to the common problems, we had to deal with the specificity of the Portuguese language that includes different variations for masculine and feminine and singular and plural words. One such example is the proposition "of", which can be translated as "de", "do", "da", "dos" and "das".

Thirdly, in 2008 the Portuguese Language Orthographic Agreement was signed into law with a transition period, ending on December 31st, 2015. During the transition period, different firms used different spelling variations for some words and, therefore, the algorithm has to recognise all spelling variations. For instance the Portuguese word for shareholders changed from "Accionistas" to "Acionistas". On the other hand, the double spelling of some words was kept, such as the word "Sector", which can also be spelled "Setor". This imposes an additional layer of difficulty when analysing narratives.

Fourthly, virtually all reports include an Auditor's Report. However, the Portuguese Companies Act imposed additional monitoring mechanisms. Firms have to have a Fiscal Committee and an Audit Committee[n]. Some firms, have separate headings for each report, others have a section for the Auditor's Report and one for the remaining reports. Other firms, have one single section for all these reports. For this reason, we subdivided the Auditor section into three subsections: External Auditor, Audit Commission and Fiscal Committee. In the absence of any indication, we assume that it is the External Auditor's report.

Finally, the Portuguese Companies Act allows different governance structures to have different names, which correspond to the Chairman and CEO roles. This in turn involves aligning the different names to the role.

As an overall view of the whole process, the software is capable of processing the annual reports well in Portuguese and we believe that the adaptation process will be similar for other languages. Out of the 396 reports processed, we analysed a random sample of 100 reports and based on these analyses (Table 7), we conclude that the software was able to correctly process 71 of these reports. The main problems arise from: i) table of contents (TOC) split into more than one page; ii) table of contents without page numbers; iii) reports without a table of contents; iv) or reports with more than one table of contents.

Table 7. Summary of Portuguese Reports Analysis

Reports	Count	Percentage
Correctly processed	71	71 %
Errors		
TOC with more than 1 page	9	9 %
TOC without page numbers	8	8 %
No TOC	1	1 %
More than one TOC	1	1 %
Incorrectly processed	10	10 %
Total	100	100 %

[n]The Companies Act allows for some variation in this requirement, which are not relevant for this discussion.

7. Conclusion

The work reported in this chapter work demonstrates the adaptability of our extraction and classification procedures to non-English annual reports published in regulatory settings other than the UK. This work develops, describes and evaluates the first procedure for automatically extracting and analysing qualitative information in digital PDF annual report from three languages showing that our extraction methods could be expanded to be applied to other languages with minor tweaks to regular expressions and through providing a hand crafted gold–standard for each language. The results show that our extraction methods were able to process more than 98 % of UK annual reports despite the non-standard format and layout. Running the same methods over Spanish and Portuguese annual reports shows that the methods are capable of processing more than 63 % of the tested annual reports. The program successfully processes the documents and Spanish characters, albeit having some problems that are caused, mainly, by the composition of the table of contents page, the fourth and final challenge encountered in this language. The reported work makes it easier for investors, firms and analysts to access and acquire information automatically from a large volume of annual reports. Around 70 % of the table of contents page was retrieved properly, but only around 34 % of the documents were analysed completely. Overall, the CFIE-FRSE program appears to work well for Spanish. However the irregularity of the composition of the annual reports makes it difficult to succeed in every document. For Portuguese annual reports the tool was able to process 373 (63 %) of the annual reports.

Acknowledgements

The work described in this paper has been undertaken as part of three projects. We acknowledge the support of the Economic and Social Research Council (ESRC) (grant references ES/J012394/1 and ES/K002155/1), the Institute of Chartered Accountants in England and Wales (ICAEW) and the International Centre for Research in Accounting (ICRA) at Lancaster University. We would also like to thank Professor Ana Gisbert, Universidad Autónoma de Madrid, for her help with the Spanish annual reports.

References

1. F. LI, The information content of forward-looking statements in corporate filings—a naïve bayesian machine learning approach, *Journal of Accounting Research.* **48**(5), 1049–1102 (2010). ISSN 1475-679X. doi: 10.1111/j.1475-679X.2010.00382.x. URL `http://dx.doi.org/10.1111/j.1475-679X.2010.00382.x`.
2. T. LOUGHRAN and B. MCDONALD, When is a liability not a liability? textual analysis, dictionaries, and 10-ks, *The Journal of Finance.* **66**(1), 35–65 (2011). ISSN 1540-6261. doi: 10.1111/j.1540-6261.2010.01625.x. URL `http://dx.doi.org/10.1111/j.1540-6261.2010.01625.x`.
3. T. Schleicher and M. Walker, Bias in the tone of forward-looking narratives, *Accounting and Business Research.* **40**(4), 371–390 (2010). doi: 10.1080/00014788.2010.9995318. URL `http://dx.doi.org/10.1080/00014788.2010.9995318`.
4. T. LOUGHRAN and B. MCDONALD, Textual analysis in accounting and finance: A survey, *Journal of Accounting Research.* **54**(4), 1187–1230 (2016). ISSN 1475-679X. doi: 10.1111/1475-679X.12123. URL `http://dx.doi.org/10.1111/1475-679X.12123`.
5. A. Doucet, G. Kazai, B. Dresevic, A. Uzelac, B. Radakovic, and N. Todic. Icdar 2009 book structure extraction competition. In *Proceedings of the Tenth International Conference on Document Analysis and Recognition (ICDAR'2009)*, pp. 1408–1412, Barcelona, Spain (July, 2009).
6. S. Teufel, *The structure of scientific articles: Applications to Citation Indexing and Summarization.* CSLI Studies in Computational Linguistics, Center for the Study of Language and Information, Stanford, California (2010). ISBN 9781575865560.
7. L. McConnaughey, J. Dai, and D. Bamman. The labeled segmentation of printed books. In *Proceedings of the EMNLP 2017 conference* (2017).
8. A. Devitt and K. Ahmad. Sentiment polarity identification in financial news: A cohesion-based approach. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pp. 984–991 (2007).
9. R. P. Schumaker. An analysis of verbs in financial news articles and their impact on stock price. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics in a World of Social Media*, WSA '10, pp. 3–4, Association for Computational Linguistics, Stroudsburg, PA, USA (2010).
10. T. L. Im, P. W. San, C. K. On, R. Alfred, and P. Anthony. Analysing market sentiment in financial news using lexical approach. In *Open Systems (ICOS), 2013 IEEE Conference on*, pp. 145–149 (Dec, 2013).
11. J. Z. Ferreira, J. Rodrigues, M. Cristo, and D. F. de Oliveira. Multi-entity polarity analysis in financial documents. In *Proceedings of the 20th Brazilian Symposium on Multimedia and the Web*, WebMedia '14, pp. 115–122, ACM, New York, NY, USA (2014). ISBN 978-1-4503-3230-9. doi: 10.1145/2664551.2664574. URL `http://doi.acm.org/10.1145/2664551.2664574`.
12. B. Neuenschwander, A. C. Pereira, W. Meira, and D. Barbosa. Sentiment analysis for streams of web data: A case study of brazilian financial markets.

In *Proceedings of the 20th Brazilian Symposium on Multimedia and the Web*, WebMedia '14, pp. 167–170, ACM, New York, NY, USA (2014). ISBN 978-1-4503-3230-9.

13. D. Merkl-Davies and V. Koller, 'metaphoring' people out of this world: a critical discourse analysis of a chairman's statement of a uk defence firm, *Accounting Forum*. **36**(3), 178–193 (9, 2012). ISSN 0155-9982. doi: 10.1016/j.accfor.2012.02.005.
14. N. M. Brennan and D. M. Merkl-Davies. Accounting narratives and impression management. In *The Routledge Companion to Communication in Accounting* URL `https://ssrn.com/abstract=1873188`.
15. A. Brooking, *El capital intelectual: el principal activo de las empresas del tercer milenio*. Paidós Empresa, Barcelona (1997).
16. L. Edvinsson and M. Malone, *El capital intelectual: Cómo identificar y calcular el valor de los recursos intangibles de su empresa*. Gestión, Barcelona (1999).
17. E. García-Meca, I. Parra, M. Larrán, and I. Martínez, The explanatory factors of intellectual capital disclosure to financial analysts, *European Accounting Review*. **14**(1), 63–94 (2005).
18. F. Tejedo Romero, Información del conocimiento organizacional a través de los informes anuales publicados en las páginas web de las empresas, *Revista Española de Documentación Científica*. **37**(1) (2014). doi: http://dx.doi.org/10.3989/redc.2014.1.1068.
19. F. Tejedo Romero, Información de los recursos intangibles ocultos: ¿memorias de sostenibilidad o informe anual?, *European Research on Management and Business Economics*. **22**(2), 101–109 (2016).
20. E. Oliveras, C. Gowthorpe, Y. Kasperskaya, and J. Perramon, Reporting intellectual capital in Spain, *Corporate Communications: An International Journal*. **13**(2), 168–181 (2008). doi: http://dx.doi.org/10.1108/.
21. M. A. Villacorta, Revelation of the voluntary information about Human Capital in the Annual Reports, *Intangible Capital*. **2**(1) (2006). doi: http://dx.doi.org/10.3926/ic.43.
22. V. I. Levenshtein, Binary codes capable of correcting deletions, insertions, and reversals, *Soviet Physics Doklady*. **10**(8), 707–710 (1966).
23. B. Lowagie, *iText in Action*. Covers iText 5, Manning Publications Company (2010). ISBN 9781935182610.
24. M. El-Haj, P. Rayson, S. Young, and M. Walker. Detecting document structure in a very large corpus of UK financial reports. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014, Reykjavik, Iceland, May 26-31, 2014.*, pp. 1335–1338 (2014). URL `http://www.lrec-conf.org/proceedings/lrec2014/summaries/402.html`.
25. L. Costa, L. Rodrigues, and R. Craig, Factors associated with the publication of a ceo letter, *Corporate Communications: An International Journal*. **18**(4), 432–450 (2013).
26. R. L, F. Romero, and R. Craig, Corporate governance and intellectual capital reporting in a period of financial crisis: Evidence from portugal, *International Journal of Disclosure and Governance*. **14**(1), 1–29 (2017).

27. M. Marques, Os activos intangíveis nas contas das empresas do psi 20: uma evidência empírica, *Pecvnia: Revista de la Facultad de Ciencias Económicas y Empresariales*. **8** (2009).