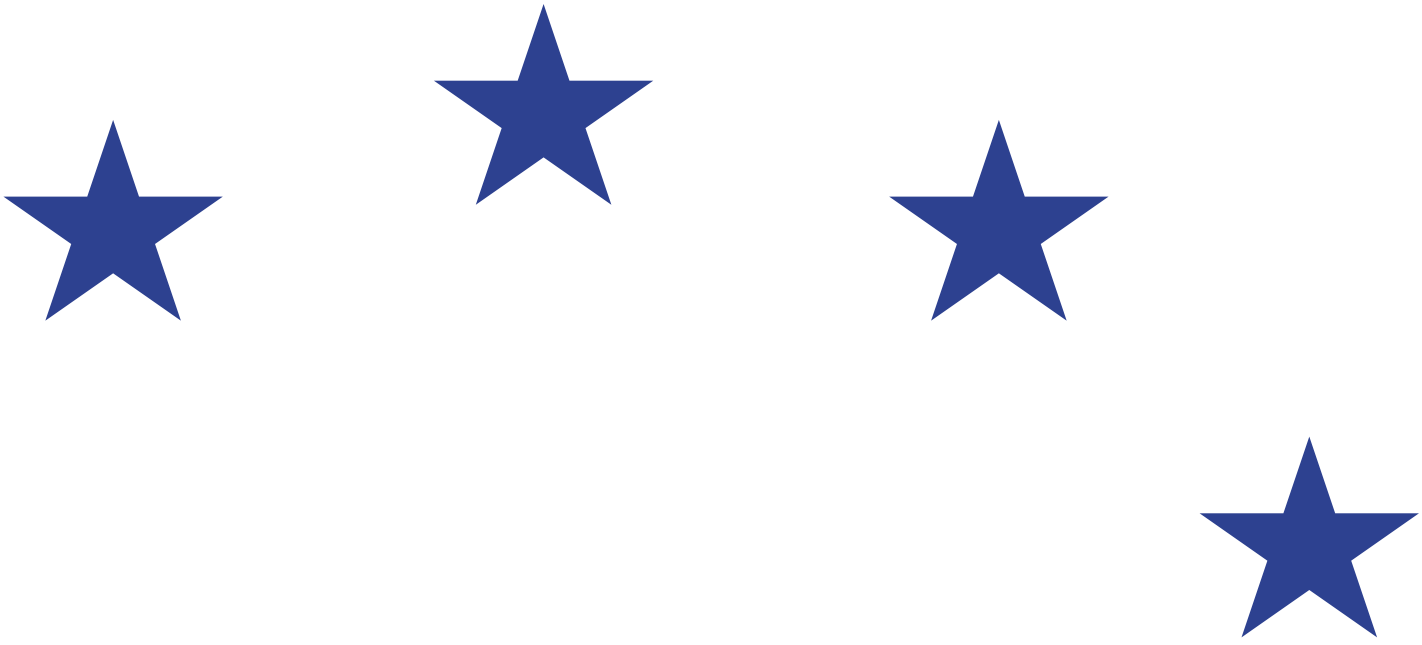
esma European Securities and Markets Authority

TRV Risk Analysis

Parsing prospectuses: A text-mining approach



ESMA document number: 50-164-6327
Publication date: 29 November 2022

ESMA Report on Trends, Risks and Vulnerabilities Risk Analysis

ISBN 978-92-95202-64-1, DOI 10.2856/03284, EK-07-22-983-EN-N
European Securities and Markets Authority (ESMA)
Risk Analysis and Economics Department
201-203 Rue de Bercy
75012 Paris
FRANCE
risk.analysis@esma.europa.eu

Investor protection

Parsing prospectuses: A text-mining approach

Contact: adrien.amzallag@esma.europa.eu[1]

Summary

The EU Prospectus Regulation sets out strict requirements on how issuance prospectuses for securities like shares and bonds should be drafted and presented. Because of the importance of these documents for investors, it is useful to understand how actual prospectuses match up with these specific requirements. We aim to contribute to this question by applying natural language processing techniques to a unique dataset consisting of all prospectuses approved under the Prospectus Regulation between end-November 2020 and January 2022. After evaluating 593,000 pages of text, we find that prospectuses from issuers in the EU can pose challenges for those intending to use them: they contain substantial repetition of text, include broken hyperlinks, may present generic and imprecise risk factors, and may include unclear language regarding availability of working capital. In addition, we find statistical evidence that longer prospectuses, all else being equal, contribute to greater divergence among rating agency assessments of credit risk. This suggests that an abundance of material can present a challenge for even specialised readers to identify information that is key to assessing the product. Our findings are a contribution to assessing the content of issuance prospectuses by means of text mining, i.e. an advanced analytical technique which enables the enormous volumes of text that prospectuses entail to be more comprehensively assessed. Our study also illustrates the effectiveness of text mining as a supervisory technology tool.

Introduction

Today, European investors receive substantially more information than before the 2007-2008 Global Financial Crisis, thanks not least to a plethora of additional disclosure requirements. Much of this information comes in the form of text, via documents like prospectuses, PRIIPs KIDs, and UCITS KIIDs.

A number of EU Directives and Regulations govern the production of these documents, with a view towards striking a balance between flexibility to capture idiosyncratic features of financial products and the benefits of standardisation, such as facilitating comparability and investor comprehension.

This article presents the results of a European Securities and Markets Authority (ESMA) endeavour to apply natural language processing (NLP) methods to a dataset of 3,220 documents retrieved from the recently enhanced Prospectus Register (PR).[2] In total, approximately 593,000 pages of text have been analysed.

The motivation for this article is severalfold. First, prospectus documents are long: an average of approximately 200 pages, but also up to as many as 1,000 pages. In addition to the often-technical language used, the sheer size of prospectuses makes it arduous for investors (especially retail investors) and supervisors to fully process the

1 This article was written by Adrien Amzallag, Giulio Bagattini and Lars Linz. We gratefully acknowledge helpful comments and feedback from Zeno Benetti, Luigi Borrelli, Gregory Frigo, Isabelle Grauer-Gaynor, Claudia Guagliano, Steffen Kern, Elisabeth van Laere, Valerio Novembre, Eoghan O'Neill, Paul-Henri Pruvost, and Ana Maria Rivera Serrano.

2 The PR is hosted by ESMA and began operating in November 2020, when it replaced the former prospectus register under Directive 2010/78/EU implementing Articles 21 and 25 of Regulation (EU) 2017/1129 of the European Parliament and of the Council of 14 June 2017 (Prospectus Regulation). The PR provides a centralised European reporting framework for prospectus documents: national competent authorities submit prospectus documents approved in their jurisdiction paired with a wide range of metadata (i.e. ISIN codes, issuer Legal Entity Identifier, approval date, etc.). The register allows users to search for and download prospectuses, registration documents, universal registration documents, securities notes, summaries, supplements, amendments, and final terms according to a wide range of search criteria on either document or security level.

information therein.[3] With this in mind, the Prospectus Regulation[4], which mostly began to apply in July 2019, contains indications on the style, clarity, and content that prospectuses and accompanying documents should include. In particular, Article 6(2) states that "*The information in a prospectus shall be written and presented in an easily analysable, concise and comprehensible form…*". Textbox 1 provides background on prospectuses, the Regulation, and statistical considerations.

Textbox 1

Prospectuses, the Regulation, and statistics

The Prospectus Regulation is part of an EU-wide arrangement that harmonises requirements for the drafting, approval, and distribution of prospectuses. This regime applies when securities are offered to the public or admitted to trading on a regulated market in an EU Member State. The regime is designed to reinforce investor protection by ensuring that all prospectuses, wherever drawn up in the EU, provide clear and comprehensive information, while at the same time making it easier for companies to raise capital throughout the EU.

Investors can expect to find a substantial amount of information in prospectuses, including a summary of the product, the identity of the key parties involved and information on what is being offered and when. Prospectuses are also required to provide essential information on both the product (such as the reason for offer, use of proceeds, and risk factors) and the issuer (including the history of the company, an overview of its business activities, and its organisational structure). In light of the wide variety of products within the scope of the Regulation, issuers have the flexibility either to provide all material in a single document or to assemble required material in several constituent parts (a tripartite' prospectus). These parts consist of a registration document (including information on the issuer) and a securities note (information specifically on the securities being offered or to be admitted to trading), and may or may not include a summary.

With respect to details on the securities being issued, prospectuses may be drawn up either to contain all necessary information (a standalone prospectus) or to be accompanied by specific information on the individual issuances in additional documents (a base prospectus, accompanied by final terms). In addition, any significant new factor, material mistake or inaccuracy which could influence the assessment of the investment, arising after the prospectus is published but before the offer has closed or trading has begun, requires the approval and dissemination of a supplement to the prospectus. Lastly, the Regulation introduced a simplified disclosure regime subject to certain conditions.

ESMA compiles annual statistics on prospectus approvals—the flow of new prospectuses—in EEA countries, starting in 2007 (ESMA, 2021). Following a peak in 2007, there has been a declining trend in prospectus approvals within the EU27: in 2020, 2,612 prospectuses were approved (29% of the peak of 8,875 in 2007). Given the length of the time period and the number of jurisdictions involved, the drivers behind these declines are likely to be numerous and varied, and would deserve a comprehensive analysis. We leave this topic open for future research.

It is more challenging, in contrast, to establish the stock of prospectuses that are 'active' at any given time, for use as a benchmark regarding the flow. This is because prospectuses have a validity date of 12 months. Afterwards, they are null and void, and can no longer be used for offers. Prospectuses for single issuances, such as certain equity issuances, can expire much earlier than twelve months. In other cases, such as long-running programmes from which securities are repeatedly issued, the base prospectus is still required to be updated/revised after twelve months (e.g. with the latest financials from the issuer, updated market conditions, and risk factors for the programme). This also means entirely new documents being issued.

The annual prospectus activity statistics produced by ESMA do provide an overall sense of the stock of 'active' prospectuses, but are not an exact indication. For example, a prospectus approved on 1 January 2021 that expires in a few days and a prospectus approved on 31 December 2021 that expires in 12 months would both be counted in the activity statistics for 2021.

The second motivation for this analysis relates to the size of the document set. Thousands of prospectuses and accompanying documents are issued each year. Therefore, it can be challenging for the European supervisory community to determine whether issuers across the EU are meeting the expectations of the legislators who designed the Prospectus Regulation. With this in mind, our research aims to provide novel tools for supervisors and policymakers, to facilitate their assessment of issuers' application of the Prospectus Regulation and to identify themes which potentially warrant closer monitoring.

Against this background, this article explores a number of linguistic features of prospectuses, such as their length, the 'effective' length once all documents referenced by links are included, the extent to which duplication of text occurs, and the complexity of the language used. We also examine the contents of specific sections and required phrases, focusing in particular on the risk factors section and working capital statements.

The remainder of this article is organised as follows. The next section describes the sample. Then, we present findings relating to prospectus documents as a whole (length, duplication, and complexity of language). The subsequent section focuses on specific components such as the risk factors section and working capital statements. Finally, we offer concluding remarks and next steps.

Data: >3,000 prospectuses

Our sample includes all documents – excluding final terms and supplements – submitted to the PR between 30 November 2020 and

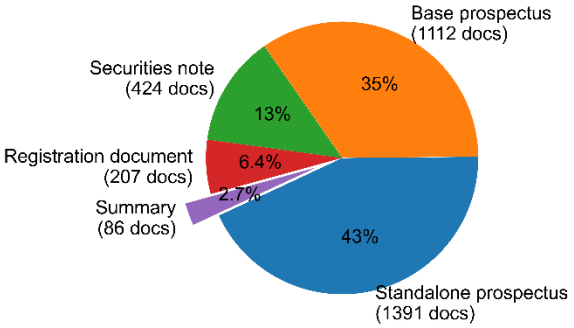
[3] Article 7(3) of the Prospectus Regulation requires a summary limited to 7 pages, to facilitate understanding by retail investors.

[4] Regulation (EU) 2017/1129 of the European Parliament and of the Council of 14 June 2017.

27 January 2022: 3,220 documents in total.[5] This covers all EU prospectuses related to securities offered to the public or admitted to trading on a regulated market during this time period.

Chart 1 below shows a breakdown of the sample by document type. Over 75% of the documents are either standalone or base prospectuses submitted as single documents, which contain a summary, a registration document, and a securities note. The rest of the documents refer to prospectuses which are submitted to the PR as three separate documents.[6]

Chart 1
Document types
Base and standalone prospectuses make up >75% of the sample



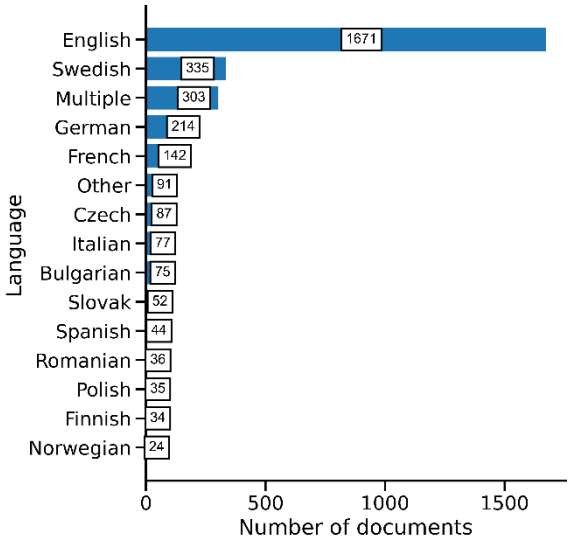
Note: The chart displays the distribution of document types out of the total 3,220 documents.
Source: ESMA

The majority of the instruments in our sample are debt securities (1,229 documents) and shares (744), followed by derivatives (312) and asset-backed securities (265). Our sample also includes relatively fewer common instruments, such as units or shares in closed-end funds, as well as depositary receipts.[7]

In terms of geography, our sample contains prospectuses submitted to 29 different national competent authorities (NCAs) (all of the EEA countries except Lithuania). Most documents in the sample are associated with Ireland (562), closely followed by Luxembourg (520) and thereafter Sweden (485).

Elsewhere, our sample includes documents drafted in 21 different languages. Chart 2 below displays the distribution of languages within our sample. As can be seen, a non-negligible number of documents (303) contain multiple languages.[8]

Chart 2
Document languages
Almost 50% of docs are issued in English



Note: The chart displays the languages of documents in our sample. The category 'Other' includes languages with fewer than 20 documents (Croatian, Danish, Dutch, Estonian, Greek, Hungarian, Icelandic, Portuguese, and Slovenian). We define a document as being drafted in multiple languages if it contains more than 10 pages in at least two languages.
Source: ESMA

Prospectus length: 0.5mn pages to digest

Document length and hyperlinks

Chart 3 below shows the variety of document length – measured in terms of number of pages –

5 The dataset also includes 5,364,697 final terms and 2,532 supplements provided separately. Our analysis excludes this information for ease of processing, insofar as these items relate only to base prospectuses and the metrics that we focus on in this study (length, duplication, risk factor similarity, and working capital statements) are expected to be mainly driven by information contained in the base prospectus, rather than in final terms or supplements.

6 We choose to preserve this distinction in the dataset and not to merge summaries, registration documents, and securities notes into single prospectuses, since registration documents can be used for more than one prospectus. Additionally, some related summaries had yet to be submitted to the register at the data cut-off date.

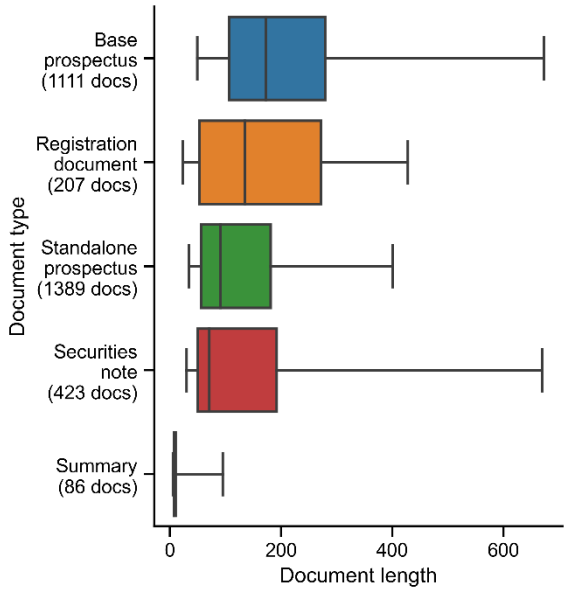
7 An alternative presentation of the data sample would be to distinguish between the motivation of issuances, such as Initial Product Offering (i.e. a new equity issuer), general admission to trading in a regulated market under a well-established programme, recapitalisations, or complex transactions involving a variety of securities. Such descriptive information is not directly available in the PR.

8 To avoid the risk of languages being identified incorrectly by our software, we define a document as being drafted in multiple languages if it contains more than 10 pages in at least two languages (i.e. more than 10 pages in the first language and more than 10 pages in the second language, and so forth in case of a third language).

observed across each document type.[9] Base prospectuses tend to be the longest document type – sometimes exceeding 650 pages – and also display a large variety in length, as is also the case for securities notes.

Unsurprisingly, summaries tend to be the shortest documents. At first glance, some appear to exceed the maximum length of seven pages required in Article 7(3) of the Prospectus Regulation. However, upon further inspection this appears to be explained by the fact that documents submitted to the PR under the label 'summary' can include both a summary and a securities note.

Chart 3
Document length
Base prospectuses are the longest documents



Note: Each box shows the range of the number of pages for documents of a specific type. The vertical line in each box indicates the median. Box edges are the 25th and 75th percentiles, while the whiskers represent the 5th and 95th percentiles.
Source: ESMA

We also investigate the extent to which prospectus documents refer to other documents via URLs. The use of hyperlinks is relevant because it yields a more precise measure of the amount of information that an issuer makes available via the prospectus. In addition, a large number of sources that are external to the prospectus may make it more difficult for investors and supervisors to retrieve all content relevant to their understanding of the product.

Table 1 below provides some summary statistics. We consider hyperlinks from several perspectives, including the total use of links in a document and the number of distinct links (i.e. ignoring duplicate links). In addition, we follow the trail of links and test how many hyperlinks in prospectuses are broken. Finally, we distinguish between links to documents and links to webpages. Each of these categorisations enables us to draw insights on readers' likely experience, as discussed in the following paragraphs.

Table 1
Hyperlinks in documents
Many links are not functioning

Category	Value	Share
Total links	126,177	39 per doc
Distinct links	31,799	10 per doc
Links to webpages	23,330	73%
Links to documents	8,469	27%
of which functioning	7,189	85%

Note: All numbers in the 'Value' column are totals over the entire data sample. The percentages displayed in the 'Share' column are percentages out of distinct hyperlinks. 'Distinct' refers to within a document, i.e. links which are repeated within a document are only counted once. The same link appearing in different documents is counted once per document.
Source: ESMA

In terms of document type, base prospectuses tend to use links most heavily (median of c. 40 links per document), followed by standalone prospectuses (c. 25 links per document). Summaries contain the least number of hyperlinks, which is expected given that the Regulation requires summaries to be self-contained. Links to documents beyond the prospectuses (e.g. financial statements) account for 27% of all hyperlinks found in prospectuses, with the rest pointing to web pages.[10] We also test whether the files linked in each prospectus can be retrieved and find this to be the case for 85% of the links. In other words, out of 8,469 distinct links, 7,189 links could be confirmed to function.[11]

[9] An interesting extension would be to compare the different sections within each document category with each other, such as the length of the risk factors section. In order to be efficient over such a large sample of documents, this analysis would require machine-readable documents with each section precisely delimited. We address this topic in the concluding remarks of this article, and leave open this specific analysis for future research.

[10] Linked documents are mainly PDF files, but also include Word, Excel, and .txt files.

[11] We run two distinct algorithms (using two different Python modules) to automatically access the links and return an output signalling whether the link is valid or not. These algorithms have two limitations. First, some links that are

The ‘effective’ document length

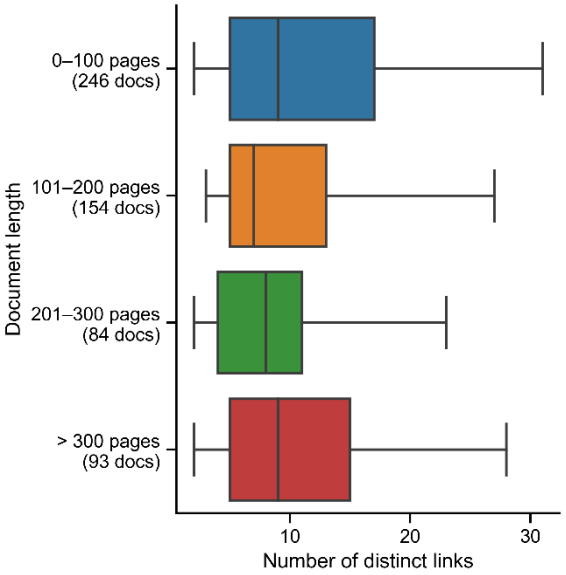
Given the pervasive use of links, which point the reader to additional resources, the length of prospectus documents presented in Chart 3 above arguably does not give the full picture when it comes to the amount of information made available by issuers via prospectuses. Issuers might draft short prospectuses while actually confining relevant information to linked documents.[12] Such a practice would be of interest relative to certain provisions in the Prospectus Regulation, including the requirements to be “easily analysable” and “concise” in Article 6(2).

To begin with, Chart 4 shows that, in the case of equity securities, there is indeed a tendency for shorter prospectuses to use more links: documents with less than 100 pages often include more links than longer documents. This suggests that information is increasingly ‘outsourced’ as documents get shorter.[13]

To further shed light on the amount of information which issuers provide by referring to external sources, we download the documents linked to via each prospectus. This results in an additional 943,841 pages of outsourced text – or 293 additional pages per prospectus. These extra texts include additional rules, marketing material, information on the issuer, periodic (e.g. annual) reports, and financial accounts.

We then augment the actual length of the prospectus documents in our sample with the length of these extra materials, in effect producing the ‘effective’ length of a prospectus.[14] This measure appears relevant to examining how “easily analysable” and “concise” the prospectus documents are in practice.

Chart 4
Number of links
Shorter equity prospectuses contain more links



Note: Each box shows the range in the number of distinct hyperlinks in (standalone) prospectuses of equity securities, grouped by document length. The sample excludes 16 “EU Recovery” prospectuses, which are markedly shorter documents drafted according to simplified requirements. These prospectus types had extremely low numbers of hyperlinks, and appear less comparable to the wider universe insofar as, due to their special disclosure regime, issuers might not need to refer to as much external information. The vertical line in each box indicates the median number of links. Box edges are the 25th and 75th percentiles, while the whiskers represent the 5th and 95th percentiles.
Source: ESMA

Chart 5 compares the range in ‘effective’ length with the length of the original document, grouping prospectuses according to national jurisdiction. Not surprisingly, this combined length often markedly exceeds the length of the original document, in some cases more than doubling the number of pages that a reader must go through to view all of the information presented by the issuer. Linked documents are often several hundred pages long, which results in the 75th percentile of the ‘effective’ document length (marked by the upper end of the black lines in

deemed to be invalid are actually functioning when inserted by a human into a web browser, as some webpages have ‘firewalls’ that prevent automatic applications to access them. Second, we cannot rule out errors in the extraction of the hyperlink text from the PDF documents.

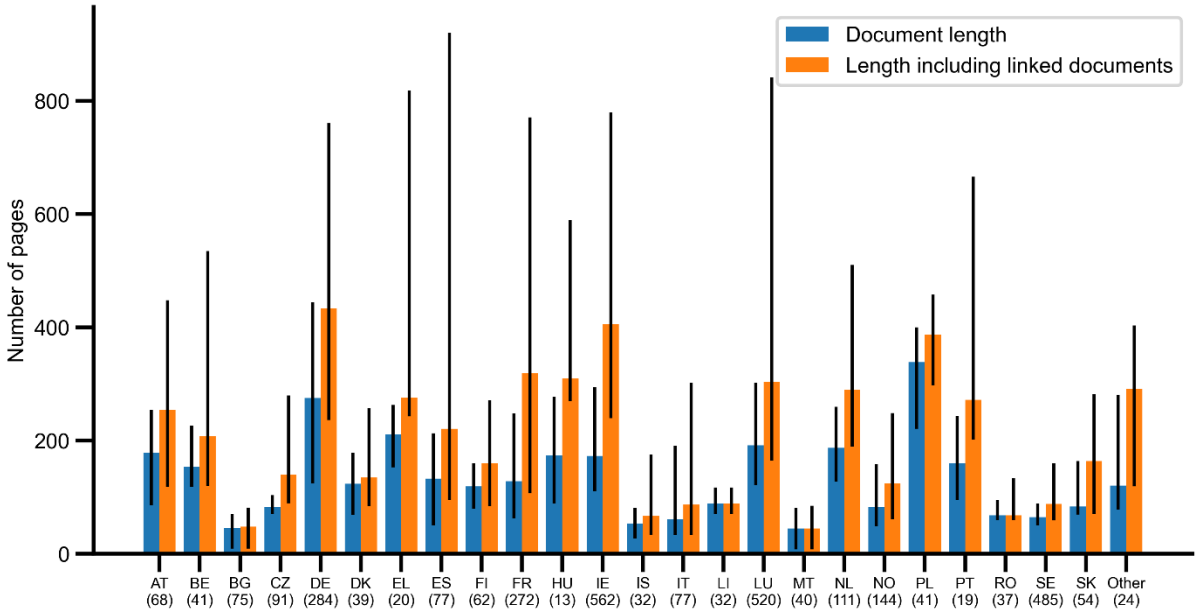
12 In fact, the Prospectus Regulation allows certain mandatory information to be incorporated in the prospectus by referencing other documents. See in particular Article 19 of the Prospectus Regulation and also footnote 14.

13 In prospectuses of securities other than equities, we do not observe a clear pattern.

14 Some information included—such as marketing material—is not expected to fall under the provisions of incorporating required information by reference, as per Article 19 of the Prospectus Regulation. Where text contained in a referenced document does not represent part of the text of a prospectus in a strict sense, we nevertheless retrieve this information as well, insofar as readers may also wish to go through this material before determining its usefulness or for understanding the financial product and any accompanying risks. It is also challenging to automatically distinguish external material that is merely being volunteered by issuers rather than being provided in accordance with the Prospectus Regulation. Further advances in machine-readability of prospectuses would facilitate this identification task and enable a more precise assessment.

Chart 5) exceeding 800 pages in some jurisdictions.

Chart 5
Median document length by NCA
Effective length often larger than original length



Note: The chart compares the length of prospectuses in each jurisdiction with the 'effective' length obtained by combining the original document with those it links to. The height of each bar indicates the respective medians of the length and effective length in each jurisdiction. The black lines indicate the 25th and 75th percentiles. Countries with fewer than 10 documents (Cyprus, Estonia, Croatia, Latvia, and Slovenia) are grouped in the category 'Other'.
Source: ESMA

Impact of length on rating consensus

The fact that some prospectuses are very long may affect how transparently and effectively information is communicated to financial markets. However, the specific impact of a longer document is not evident *ex ante*. On one hand, more content suggests that, all else being equal, more potentially valuable information is being made available to the reader. On the other hand, and in line with prior academic findings, an abundance of written material might render these resources less digestible, by making it more difficult for readers to identify key information for understanding the specific product, its issuer, and the broader operating environment.[15]

Against this backdrop, we investigate whether the amount of written material in the prospectus affects credit rating agencies' (CRAs) divergence of opinion on a product's credit risk. We focus on CRAs on the assumption that they are professional and experienced readers of prospectuses, with extensive background in assessing products of the type contained in the PR. In this respect, prospectuses are a key – though by no means unique – source of information for CRAs' credit assessments.[16] Therefore, if longer prospectuses and longer associated documents are conducive to more effective and exhaustive disclosure of key product and issuer information, this should be reflected in more consistent credit ratings among different CRAs.

In order to test this hypothesis, we retrieve the credit ratings and accompanying information

[15] For discussions illustrating the link between length and textual complexity, see Szmrecsanyi (2004), Amadxarif, Brookes, Garbarino, Patel, and Walczak (2021), deHaan, Song, Xie, and Zhu (2021), and the additional extensive references cited therein.

[16] Our analysis relates to a broader strand of literature examining the impact of transparency and textual complexity on investor and rating agency behaviour. For example, regarding asset-backed securities, see Ghent, Torous, and Valkanov (2019), Zhang, Zhao, and Zhao (2020), and Neilson (2022). See Celerier and Vallée (2017) regarding structured retail products. Some of these papers also explore complexity measures besides simple document length.

issued by CRAs for bonds and structured finance instruments in our sample. We use data from the European Ratings Platform and Refinitiv, and obtain ratings for over EUR 368bn worth of securities.[17] For products assessed by more than one CRA, we then calculate the standard deviation of the credit ratings.[18] A low standard deviation means that the different CRAs' credit risk assessments led to similar conclusions, and hence indicates consensus around the product's viability. Conversely, a larger standard deviation indicates higher disagreement among CRAs, entailing higher uncertainty around a product's prospects.[19] According to this measure, on average, the ratings of structured finance products are characterised by a higher disagreement than the ratings of bonds (0.88 versus 0.62).

We then test whether disagreement in credit ratings is influenced by the amount of content contained in the prospectus. We do so by regressing our ratings disagreement measure on the length of the prospectus and the total length of the documents linked to via the prospectus (discussed in the previous subsection).[20] To pin down the role of these two quantities, we control for other factors that may also affect the degree of uncertainty around an instrument's credit risk.

Table 2 shows the results of the estimations. In line with the hypothesis that longer prospectuses hinder CRAs' assessments, longer documents are associated with a larger standard deviation ('st. dev.') in ratings and thus more disagreement around the instrument's credit risk. This is maintained in all of the different specifications of the regression model.

Interestingly, the length (number of pages) of the linked documents tends to have the opposite effect, with more external material estimated to *increase* CRAs' consensus. At the same time, this effect appears smaller than the overall impact of prospectus length on credit ratings disagreement, for the same amount of additional pages.

As shown in column 1, these results hold even if we account for the effect of the level of the rating (ratings around the middle end of the scale tend to be associated with markedly more disagreement – see also footnote 20). They also hold after controlling for the standard deviation of an instrument's rating dates (in case ratings issued at slightly different times rely on different sets of information), the instrument's time to maturity (the minimum is approximately twelve months), and the company's market capitalisation.

We also account for the possibility that the different security type (bond vs. ABS), issuer/originator's industry sector, country of domicile, and issuance date may, respectively, explain the different levels of disagreement.[21] We do this by progressively adding to the estimation fixed effects for the respective categories (columns 2, 3 and 4). Conversely, we drop the time to maturity and market capitalisation control variables, which do not seem to play a significant explanatory role, as their values are missing for a large number of observations. Beyond slight changes in the magnitude of the estimated coefficients, our findings are robust to including these additional control variables. Interestingly, solicited ratings tend to be associated with less divergence, as opposed to those cases where the instrument receives both solicited and unsolicited ratings. This relationship appears sensible: for solicited ratings agencies typically also leverage

[17] We focus on bonds and structured finance products as we could retrieve only very few credit ratings associated with other security types or the issuer itself. The CRAs for which we could retrieve rating information are Moody's (1,227 instruments rated), S&P (1,179), Fitch (754), DBRS (278), Creditreform (226), Scope Ratings (49), Nordic Credit Rating (33), KBRA (29), and JCR (9). We consider credit ratings issued at the earliest 60 days before the approval date of the instrument's prospectus. The final sample includes bonds worth at least EUR 263bn and ABS worth at least EUR 105bn. The outstanding amount at issuance was not available for 403 out of 1,030 securities retrieved. Therefore, the actual outstanding amount is likely to be far higher than these figures.

[18] We map the agencies' ratings from excellent to poor on a scale from 1 to 20, using the information provided as per Annex I: Table 6, row 9 in Commission Delegated Regulation (EU) 2015/2. The final sample includes 1,030 instruments rated by at least two CRAs. Most instruments (430 bonds and 311 ABS) are rated by two CRAs, while some have three ratings (207 bonds and 22 ABS) or four ratings (60 bonds).

[19] This also relates to the literature on opacity – where risks are hard to observe for an outsider – and its (positive) association with dispersion in ratings, either in the form of credit ratings (Morgan 2002) or analyst forecasts (Güntay and Hackbarth 2010). See also references in the subsequent footnotes.

[20] We conduct this analysis at the prospectus level, combining securities notes, registration documents and summaries for those prospectuses submitted as multiple documents.

[21] See also Morgan (2002) and Livingston, Naranjo, and Zhou (2007) on the effect of industry-specific opacity on rating divergences. Vu, Alsakka, and ap Gwilym (2017) examine the drivers of divergences in credit ratings of sovereign bonds, and conclude that opaque sovereigns – in terms of information disclosure and political risk – are more likely to receive split ratings. Although our focus is on corporate issuances, sovereign ratings typically act as a floor for the ratings of both private entities that are domiciled in that sovereign and their issuances.

information collected directly via discussions with the issuer.

Table 2

Effect of prospectus length on credit rating consensus

Longer prospectuses lead to more diverse ratings

	(1) Rating st.dev.	(2) Rating st.dev.	(3) Rating st.dev.	(4) Rating st.dev.	(5) Rating st.dev.	(6) Rating st.dev.
Prospectus length	.13***	.091***	.054**	.047**	.06**	.073***
	(5.74)	(3.43)	(2.13)	(2.03)	(2.19)	(2.89)
Linked docs length	– .015***	– .0062**	– .0073***	– .0075***	– .0013	– .00083
	(– 3.89)	(– 2.24)	(– 3.17)	(– 3.31)	(– 0.25)	(0.14)
Average rating	.14***	.2***	.21***	.22***	.23***	.23***
	(5.35)	(8.38)	(8.30)	(8.51)	(5.84)	(5.72)
Average rating squared	– .0041*	– .0085***	– .0089***	– .0093***	– .0087***	– .0087***
	(– 1.93)	(– 4.46)	(– 4.54)	(– 4.67)	(– 3.14)	(– 3.11)
Rating date st.dev.	– 7.4e–09***	– 5.6e–09**	– 5.0e–09*	– 5.0e–09	4.8e–09	6.5e–09
	(– 3.60)	(– 2.13)	(– 1.78)	(– 1.32)	(0.81)	(1.04)
Solicited only	– .0037	– .14*	– .19***	– .20***	– .17**	
	(– 0.05)	(– 1.77)	(– 2.98)	(– 3.07)	(– 2.31)	
Time to maturity	– .00019					
	(– 0.41)					
Log (Market cap.)	.029					
	(1.07)					
Constant	Yes	No	No	No	No	No
Security type F.E.	No	Yes	Yes	Yes	Yes	Yes
Industry F.E.	No	No	Yes	Yes	No	No
Country F.E.	No	No	Yes	Yes	No	No
Quarter F.E.	No	No	No	Yes	No	No
Issuer F.E.	No	No	No	No	Yes	Yes
CRA dummies	No	No	No	No	No	Yes
Observations	579	1030	1013	1013	936	936
R-squared	0.27	0.29	0.34	0.36	0.6	0.61

Note: The table presents the results of the estimation of a linear regression model where the unit of observation is represented by a security and its respective prospectus document. The dependent variable is the standard deviation of the credit ratings issued by CRAs for the respective security. Only ratings issued at the earliest 60 days before the date of approval of the security's prospectus are considered. The sample includes only debt instruments (bonds and ABS) and only prospectuses which are not drafted in multiple languages. "Prospectus length" and "Linked docs length" are expressed in hundreds of pages and winsorised at the 99th percentile; "Rating date st. dev." and "Time to maturity" are expressed in years. "Solicited only" is a dummy variable which takes the value of 1 if the instrument receives only solicited ratings (83% of the observations) and 0 if the instrument receives both solicited and unsolicited ratings (17% of the observations). In column 6, the "Solicited only" dummy is absorbed by the CRA-level dummy variables. In "Log (Market cap.)", market capitalisation refers to the next available parent company (e.g. if not available for an issuer, such as an ABS special purpose entity, then we take the market capitalisation for the next-available parent along the hierarchy of ownership, using Refinitiv data). Quarter fixed effects refer to the quarter and year in which the prospectus was approved. Issuer fixed effects are at the issuer Legal Entity Identifier (LEI) level, after mapping some LEIs of special purpose vehicles to the LEI of their parent company. Standard errors are clustered at the level of the ultimate parent of the instrument's issuer, in order to account for correlation in the residuals for related instruments and prospectuses. T-statistics are in parentheses. * indicates significance at the 90th percent confidence level, ** significance at the 95th percent level, and *** significance at the 99th percent level.
Sources: ESMA, Refinitiv

In addition, in column 5 of Table 2 we take a more restrictive approach and replace all of these categories (except the security type) with fixed effects at the issuer level. This technique allows us to study the determinants of a higher or lower disagreement in ratings for different

prospectuses issued by the same company.[22] This effectively rules out the possibility that the observed disagreement in ratings is driven by unobserved (and non-prospectus related) differences between companies. While the estimated effect of linked document length becomes insignificant in column 5, this restrictive identification does not attenuate the relationship between prospectus length and rating disagreement.

Finally, we make sure that our results are not affected by which CRAs specifically rate which instruments. For example, some CRAs may, due to structural/long-term differences across CRA rating methodologies, consistently disagree with certain other CRAs. In column 6, we control for this via a set of dummy variables – one for each CRA – which take the value of 1 if an instrument is rated by the respective CRA and 0 otherwise. The inclusion of these variables does not have a significant impact on the estimation.

Overall, these findings suggest that longer prospectuses may not only fail to provide additional clarity, but can also be associated with increased ambiguity. Our econometric approach allows us to exclude a number of alternative explanations. In terms of economic significance, an additional length of 100 pages for a prospectus by the same issuer increases the rating disagreement by 10% of its average value.[23] Given the wide range of prospectus lengths that we observed in Chart 3, this is a substantial effect. Our findings appear to be in line with numerous studies that examine the link between opacity and divergences across credit ratings (see footnotes 18 and 21).

Conversely, external documents incorporated by reference in the prospectus, or provided on top of the legally required information, may increase readers' ability to discriminate among companies, suggesting that they could offer valuable insights. However, the moderate economic significance, as well as the lack of conclusive evidence from the most restrictive estimations displayed in Table 2, suggest treating this finding with caution.

Further interesting extensions to this work include assessing whether other complexity metrics used in the academic literature also play a role in credit rating divergences, including those explored in this article. This would reflect the fact that complexity is a rich concept, with many different interpretations. We leave such statistical explorations open for future research endeavours and, in the next section, illustrate and investigate several possible metrics for inspiration.

Linguistic assessment: volume versus content

Duplication

The extent to which duplication occurs is of particular relevance from an investor protection perspective, insofar as duplication of language can lead to reader fatigue and the risk of key provisions being overlooked. To explore the extent to which duplication of text occurs in prospectuses, we examine to what degree identical sentences appear more than once within each document.[24]

As shown in Chart 6 below, securities notes tend to have the greatest share of duplicate sentences. Sometimes more than half of the sentences in a document are duplicated. Interestingly, summaries also contain duplicate sentences within themselves to varying degrees. In a similar spirit, we also find many instances of longer blocks of text being repeated (such as blocks of two or three consecutive sentences), especially among base prospectuses and securities notes. This suggests that more deliberate instances of duplication are also taking place in prospectus documents.

Elsewhere (not shown), we count the occurrence of duplicates relative to each unique sentence. We find that securities notes tend to repeat the same sentence more times than any other prospectus document type. At the extreme, 10% of securities notes (196) in the sample used in this section of the article repeat sentences four times or more.[25]

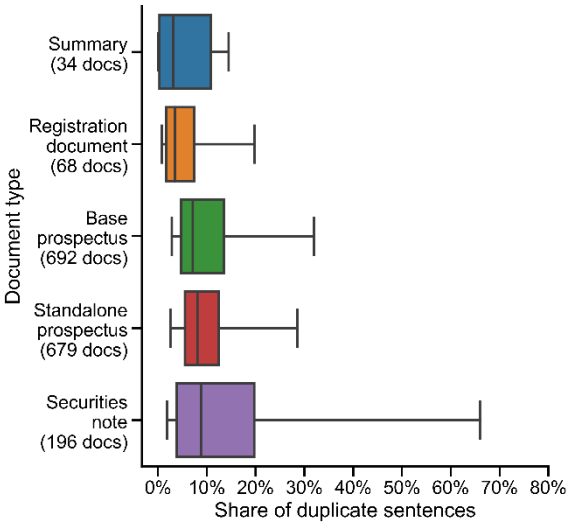
[22] For example, Hyytinen and Pajarinen (2008) find that credit rating divergences are more likely to occur among younger firms. Morgan (2002) and Livingston, Naranjo, and Zhou (2008) also demonstrate how firm-specific effects may lead to rating divergences.

[23] This is obtained based on the average rating disagreement in the sample (0.72) and the regression coefficient for the prospectus length in column 6 of Table 2, in hundreds of pages (0.073). A similar conclusion applies if we take the standard deviation of the rating disagreement (0.73) as a benchmark.

[24] Prior to this, we remove logical sources of duplication, such as the same website or document name being repeated at top or bottom of a page. However, to the extent that a summary repeats the same phrases included elsewhere in the prospectus document, this information would also be included in the duplication statistics presented in this section. This appears valid to consider in an analysis of duplication, in light of recital 30 of the Prospectus Regulation, notably that the summary "…should not be a mere compilation of excerpts from the prospectus."

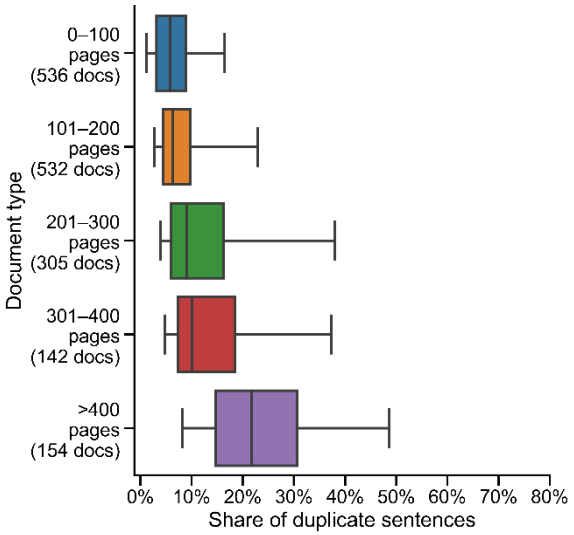
[25] Only English-language documents are included.

Chart 6
Share of duplicate sentences by document type
Greatest duplication in securities notes



Note: For each document, we calculated the number of sentences that are identical to each other, divided by the total number of sentences in the document. The chart above displays the range of that percentage of duplicate sentences across all of the documents of a specific type. By 'sentences' we mean full sentences, i.e. a set of words that is complete in itself, typically containing a subject and predicate, conveying a statement, question, exclamation, or command, and consisting of a main clause and sometimes one or more subordinate clauses (Oxford Languages 2022). Thus, we removed all tables, section headings, page numbers, etc. The vertical line in each box indicates the median. Box edges are the 25th and 75th percentiles, while the whiskers represent the 5th and 95th percentiles. Only English-language documents are included.
Source: ESMA

Chart 7
Share of duplicate sentences by document length
Longer documents have more duplication



Note: For each document, we calculated the number of sentences that are identical to each other, divided by the total number of sentences in the document. The chart above displays the range of that percentage of duplicate sentences across all of the documents grouped by page length. By 'sentences' we mean full sentences, i.e. a set of words that is complete in itself, typically containing a subject and predicate, conveying a statement, question, exclamation, or command, and consisting of a main clause and sometimes one or more subordinate clauses (Oxford Languages 2022). Thus, we removed all tables, section headings, page numbers, etc. Each box shows the range for documents of a specific length bucket. The vertical line in each box is the median for that respective document type. Box edges are the 25th and 75th percentiles, while the whiskers represent the 5th and 95th percentiles. Only English-language documents are included.
Source: ESMA

Overall, as shown in Chart 7, longer documents tend to have a greater amount of duplicated text. On this basis, it appears that longer documents do not always include correspondingly greater amounts of information. From another perspective, any new information that a longer document might include would appear to be increasingly diluted as it grows longer. However, one reason may also be that there is a lot of repetition for individual issuances/securities within a programme, such as for securities notes, which seem to be the longest documents (see Chart 3 above).

It is challenging to assess the extent to which duplication of text can be deemed excessive. To help benchmark our findings, we conduct similar duplicate sentences checks on a sample of approximately 1,000 publicly available investment fund prospectuses obtained in July 2021. These prospectuses each refer to an investment fund (multi-fund prospectuses are excluded). The variety of duplication across these fund prospectuses compared with standalone prospectuses[26] is shown in Chart 8 below.

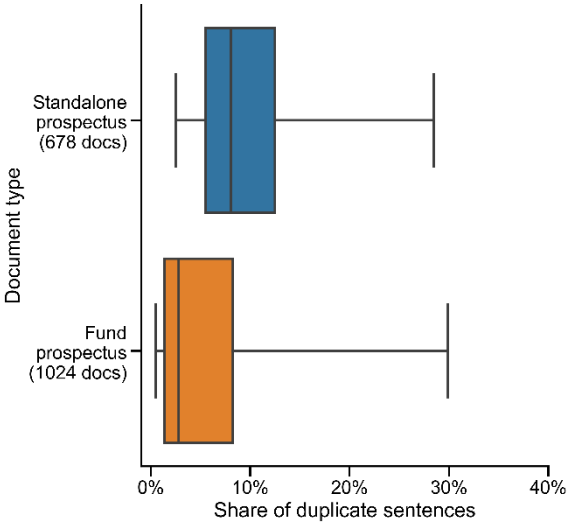
This exercise reveals that investment fund prospectuses tend to have far fewer instances of duplication than standalone prospectuses produced under the Prospectus Regulation. However, there appears to be no a-priori reason for this.[27] This benchmarking exercise indicates

[26] Base prospectuses with or without final terms would not appear ideal to be compared against fund prospectuses, as the former include final term schedules (either complete – for base prospectuses with final terms – or not completed – for base prospectuses without final terms) that are of a different structure than fund prospectuses. We consider that standalone prospectuses are the most appropriate document type in the PR to compare against fund prospectuses.

[27] The fund prospectuses in our sample are shorter (77 pages on average; 73 pages median) than the standalone security prospectuses in our comparison sample (167 pages on average; 121 pages median). With this in mind, one might argue that longer documents generally include

that the current extent of duplication among prospectuses under the Prospectus Regulation could deserve further attention.

Chart 8
Benchmarking duplication of text in prospectuses
More duplication in standalone vs. fund prospectuses



Note: For each document, we calculated the number of sentences that are identical to each other, divided by the total number of sentences in the document. The chart above displays the range of that percentage of duplicate sentences across all the documents in the chosen group. "Fund prospectus" refers to EU investment fund prospectuses made publicly available and that refer to a single fund (i.e. multi-fund prospectuses are not included). By 'sentences' we mean full sentences, i.e. a set of words that is complete in itself, typically containing a subject and predicate, conveying a statement, question, exclamation, or command, and consisting of a main clause and sometimes one or more subordinate clauses (Oxford Languages 2022). Thus, we removed all tables, section headings, page numbers, etc. Each box shows the range for documents of a specific type. The vertical line in each box is the median for that respective document type. Box edges are the 25th and 75th percentiles, while the whiskers represent the 5th and 95th percentiles. Only English-language documents are included.
Source: ESMA

Linguistic complexity

According to Article 6(2) of the Prospectus Regulation, prospectuses must be written in an "easily analysable" and "comprehensible form". Clearly, the complexity of the language used to describe a financial product likely affects the ease with which a reader is able to analyse and comprehend the product. In this way, the choice of language can affect whether one of the main aims of the Prospectus Regulation ("*to protect investors by removing asymmetries of information between them and issuers*" – recital 3) is achieved.

The field of linguistics has developed a number of ways to assess the complexity of a text, which range from basic metrics, such as sentence length, to more complicated econometric-based methods. However, it is important to choose carefully among these measures – here we need to be certain that the complexity measure will not also be driven by differences in document length.[28] Moreover, although we only analyse English-language documents in this subsection, it is preferable to focus on measures that have not been calibrated solely in majority English-speaking countries, since we focus on documents produced across the EU with many different cultural orientations.[29]

With these considerations in mind, we employ another measure of complexity that is both popular and invariant to document length: Yule's I (short for 'index of diversity'), which measures the uniformity of vocabulary in a text.[30] Lower values of Yule's I indicate less diversity of language, while higher values of Yule's I suggest that the vocabulary of a text is more diverse (i.e. richer). From the perspective of prospectuses and financial products, Yule's I can be interpreted as indicating the variety of words a reader can expect to find. Given that prospectuses describe specialised products, we would expect that more

greater relative amounts of duplication and thus that comparing standalone prospectuses with fund prospectuses is not a valid exercise. Therefore, as a robustness check we compared only standalone prospectuses whose length ranges from 50 to 100 pages (183 documents) with fund prospectuses whose length is also in this range (545 documents). We find identical results: standalone prospectuses appear to contain a significantly greater share of duplicated text than fund prospectuses.

28 This would exclude common metrics like the Type-Token ratio (the ratio of the number of unique words to the total number of words in the document), Hapax richness (the number of words that appear only once in the document relative to the total number of words), and semantic entropy (how likely it is that a reader can predict the next word after the word they have just read in the text). See Shannon, Weaver, and Burks (1963), Dale, Moisl, and Somers (2000), McCarthy and Jarvis (2010), Tolochko and Boomgaarden (2019), and Amzallag (2021).

29 This rules out some popular metrics, such as average word and sentence length, the Flesch–Kincaid readability test (Kincaid, Fishburne, Rogers, and Chissom 1975), the Automated Readability Index (Senter and Smith, 1967), and the FOG Index (Gunning, 1952). Many such metrics are specifically calibrated to reference specific situations, for example, the reading level of American high school students. This reference point does not seem appropriate for assessing prospectuses produced in Europe.

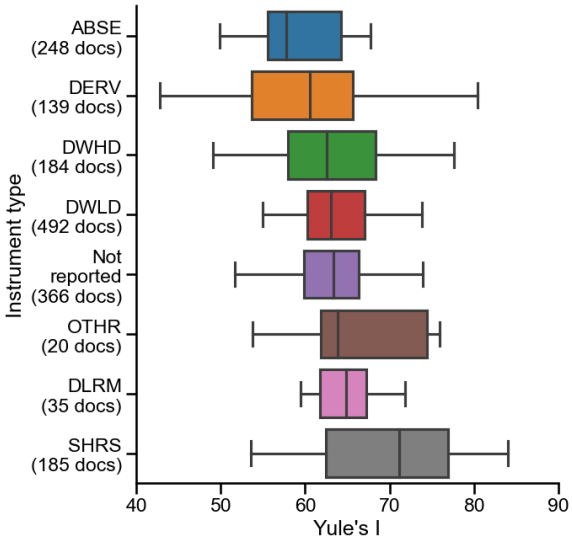
30 Yule's I metric is calculated as $\frac{M_1 \times M_1}{M_2 - M_1}$, where M_1 is the total number of words in the document, and M_2 is the sum, across all distinct words in the document, of the squared frequency of each distinct word. See Yule (1944, pp. 54–60), Williams (1970), and Oakes (1998).

complex products would have richer vocabulary (i.e. a higher Yule's I).

Chart 9 illustrates the range in Yule's I across prospectuses, grouped by instrument type. Surprisingly, prospectuses describing instruments with more complex payoffs and features, such as asset-backed securities (ABSE) and derivatives (DERV), tend to have the lowest Yule's I (although derivatives display the widest range, which reflects the variety of products pooled in this category). In other words, asset-backed securities, convertibles, and derivatives tend to have the least diverse (i.e. most uniform) vocabulary of all instrument types in our sample.

In contrast, typically more straightforward instruments like equities (SHRS) and debt instruments with high denomination (DWHD) tend to have a relatively higher Yule's I, and thus greater variety in language compared with other instrument types in our sample.[31] This is counterintuitive from our perspective: more complex products are expected to need additional terms to describe their features. Instead, it may be that more complex products use simpler language in order not to drive away potential customers, or perhaps that more simple products use a greater variety of words in order to stand out relative to competitors – we leave these considerations for further research. In any case, as a robustness check for our results, we find similar results (not shown) using an alternative measure: the Measure of Textual Lexical Diversity.[32]

Chart 9
Linguistic diversity
Complex products use less specialised language



Note: Each box shows the range in Yule's I, grouped across documents discussing the same type of instrument. The vertical line in each box is the median Yule's I for that instrument type. Box edges are the 25th and 75th percentiles, while the whiskers represent the 5th and 95th percentiles. The sample only includes English-language documents. ABSE = asset-backed securities, DERV = Derivatives, DLRM = Debt with denomination < EUR 100,000 available only to qualified investors, DWHD = Debt with denomination < EUR 100,000, DWLD = Debt with denomination ≥ EUR 100,000, SHRS = Shares, and Not reported = No security type identifiable. OTHR includes units or shares in closed end funds, depository receipts, and convertible instruments. Although prospectuses can be related to more than one instrument type, we have assigned each prospectus to one instrument type only, giving priority to more complex instruments (e.g. any prospectus including a classification to ABSE is treated as ABSE).
Source: ESMA

Content: Key information, with gaps

Risk factors

Prospectuses are mandated to include a dedicated section which describes the main sources of risk ('risk factors') linked to the instrument offered and to the issuer in general. The Prospectus Regulation has extensive requirements on the structure and contents of the risk factors, complemented by the ESMA Guidelines on Risk Factors. For example, Article 16(1) of the Regulation, supported by Recital 54, stipulates that risk factors must be specific to the

[31] An outlier perhaps is debt instruments that are only available to qualified investors (DLRM), i.e. not deemed suitable by EU policymakers for all investors. These also seem to have a relatively higher Yule's I, compared with other instrument types in the PR, although the smaller sample size (36 documents) suggests this finding should be treated with caution and reviewed once more data become available.

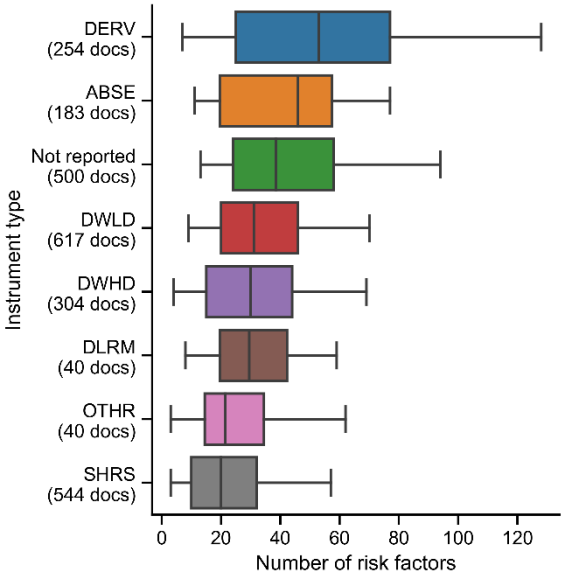
[32] The Measure of Textual Lexical Diversity is derived from the ratio of the number of unique words to the total number of words in the document, corrected for differences in length. The standard threshold of 0.72 was used. See McCarthy and Jarvis (2010) and Tolochko and Boomgaarden (2019).

issuer and the securities being described.[33] Against this backdrop, we designed several NLP tools to analyse this specific section, focusing on documents drafted in English, French, German, and Swedish.[34]

We identify the different risk factors on the basis of some assumptions. Issuers typically group risks in categories and sub-categories. However, they neither follow a specific classification nor explicitly declare which risks qualify as distinct 'factors', as there is no ex-ante definition for these concepts. Hence, we consider each of the most granular sub-categories within the section as a distinct risk factor. This data extraction exercise leads us to identify a total of 90,701 risk factors across 2,482 documents.

Chart 10 shows the distribution of the number of risk factors in a document, grouping documents by type of instrument. As expected, structured finance products like derivatives and asset-backed securities tend to have greater numbers of risk factors mentioned than more straightforward products like shares and debt instruments.

Chart 10
Number of risk factors by instrument type
Share prospectuses contain least risk factors



Note: Each box shows the range of the number of risk factors found in a document, grouped across documents discussing the same type of instrument. The vertical line in each box is the median number of risk factors. Box edges are the 25th and 75th percentiles, while the whiskers represent the 5th and 95th percentiles. The sample includes only documents with a risk factor section in English, French, German, or Swedish. ABSE = asset-backed securities, DERV = Derivatives, DLRM = Debt with denomination < EUR 100,000 available only to qualified investors, DWHD = Debt with denomination < EUR 100,000, DWLD = Debt with denomination ≥ EUR 100,000, SHRS = Shares, and Not reported = No security type identifiable. OTHR includes units or shares in closed end funds, depository receipts, and convertible instruments.
Source: ESMA

Arguably, the risk factors section carries meaningful informational content if it describes risks *specifically*, and is transparent on the idiosyncratic, company-specific sources of risks and tying macroeconomic risk factors to the specific circumstances faced by the company. In fact, from an economic perspective, a well-functioning risk factors section helps to reduce the information asymmetry between the issuer and investors by disclosing information about the company issuing the securities. This information is valuable because it cannot be easily retrieved elsewhere. The more generic/unspecific the description of the sources of risk faced by investors, the less relevant it is – since

[33] Article 16(1): "*The risk factors featured in a prospectus shall be limited to risks which are specific to the issuer and/or to the securities and which are material for taking an informed investment decision.*"

Recital 54 (extract): "*… A prospectus should not contain risk factors which are generic and only serve as disclaimers, as those could obscure more specific risk factors that investors should be aware of, thereby preventing the prospectus from presenting information in an easily analysable, concise and comprehensible form…*"

[34] English, French, German, and Swedish are the four most common languages in our sample. We examined all document types except for summaries (where a complete risk factors section is not expected). We dropped 126 documents where either the risk factors section could not be extracted, or – more frequently – it referenced another document (and therefore it did not spell out the risk factors). Following this step, we are left with 1,775 documents in English, 336 in Swedish, 238 in German, and 133 in French.

information presented in such a way is likely to already be in the public domain.

Against this backdrop, we investigate the extent to which the same, or highly similar, language is adopted when multiple issuers describe a common risk factor. This could happen by chance, implying that the motivation for specifying that risk is brief and generic. Alternatively, companies may re-use past text from other issuers as a basis for their drafting. This could be the product of different issuers outsourcing the drafting of the risk factors section to the same law firm, but also a deliberate choice by the company to save time and resources, or to use competitors as a benchmark.

We extract the sentences that follow and describe the risk factor heading, and focus our assessment on risk factors related to the following seven topics: interest rate risk, credit risk, market risk, operational risk, liquidity risk, environmental social and governance (ESG) factors, and the COVID-19 pandemic.[35] We then compare these sentences across different issuers mentioning similar risk factors, using a measure of language similarity for all pairs of sentences.[36] This results in 94 million total pairs of risk factors that were assessed.

We find a large number of highly similar risk factors. Table 3 gives an overview of the number of English-language documents where we found risk factors similar to others contained in a prospectus from a different issuer. It is clear that there is a non-negligible amount of recycling of language for certain risk factors, particularly regarding interest rate risk and liquidity risk (respectively 63% and 48% of the documents mentioning these topics use language which is highly similar across issuers). We also assess risk factors in French, German, and Swedish, and find similar results (not shown).

Table 3
Similar risk factors across prospectuses
Many cases of highly similar language

Risk factor category	Docs containing similar risk factors	%
Interest rate risk	705	63%
Liquidity risk	548	48%
Credit risk	209	30%
COVID-19	201	21%
Market risk	59	18%
ESG	67	16%
Operational risk	67	13%

Note: The first column of the table shows the number of documents containing at least one risk factor highly similar to a risk factor by a different issuer, for different sets of risk factors grouped by topic. Only English-language documents are included in this table. The second column shows this figure as a percentage of the total number of documents mentioning that topic. ESG refers to Environmental, Social and Governance.
Source: ESMA

One limitation of our findings is that our sample may still contain subsidiaries which are part of the same conglomerate or could be traced back to the same holding company. This makes it challenging – based on our dataset – to assess to what extent this seemingly widespread recycling of legal text takes place between genuinely unrelated companies rather than within conglomerates.[37]

Working capital statements

Article 14(3) and Annex III of the Prospectus Regulation require issuers to include a working capital statement in the securities notes of equity prospectuses. Articles 12, 14, and 19(2) specify that this is also expected in prospectuses of units issued by closed-end funds, depository receipts issued over shares, and – in some cases[38] – securities convertible into shares. Bearing these categories in mind, and focusing on English-language prospectuses, there are 208 documents in our sample that are expected to contain such statements, most of which for equity instruments.

As shown in Chart 11 below, in 157 documents we identified a working capital statement that adheres to the conventional wording for “clean”

[35] We allocate risk factors to these topics based on the presence of the respective keywords in the title of the risk factor.

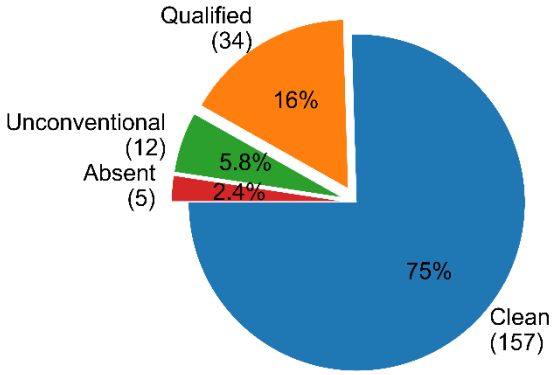
[36] We use a cosine similarity measure based on a function which compares two texts and gives as output a number between 0 (for texts with no degree of similarity detected) and 1 (if the two texts are identical). Any pair of risk factors across issuers with a similarity score greater than 0.7 is deemed to be ‘similar’. After excluding pairs of documents issued by the same entity or by different legal entities that can be attributed to the same company, this exercise resulted in 25,570 pairs of similar risk factors.

[37] A more precise analysis of these cases, which requires tools to identify all inter-group relationships, is left for future research.

[38] Specifically, when the security is convertible into shares which are not admitted to trading on a regulated market.

statements.[39] Furthermore, in 34 documents the statement is "qualified".[40] However, a number of documents do not seem to feature a working capital statement in line with one of the above formats. Specifically, in 12 documents the working capital statement uses language which deviates from the wording for either a "clean" or a "qualified" statement. In 5 further cases, we could not find a statement that adheres to one of the "clean" or "qualified" wordings, or the related paragraph heading.

Chart 11
Working capital statements
Some statements unclear or absent



Note: The pie chart shows the type of working capital statement identified in 208 English-language prospectuses. 'Unconventional' statements use wording compatible with neither a clean nor a qualified statement.
Source: ESMA

Concluding remarks

The application of NLP methods to roughly 593,000 pages of documents retrieved from the PR allowed us to assess a wide range of unexplored policy-relevant themes.

Our results yield a number of insights regarding the accessibility of prospectuses for investors and suggest that these documents may not always convey key investor information optimally. We found that prospectuses and accompanying documentation often include a significant number of references to external documents, which sometimes doubles the length of the original prospectus. Some documents contain substantial duplication of text, which may make the documents difficult to digest. In addition, more complicated instrument types tend to use less variety in language, a counterintuitive finding that suggests that the descriptions of intricate financial products may not necessarily reflect their greater complexity.

We also showed that the number of risk factors included in prospectuses can vary dramatically. Furthermore, in some cases, different issuers use the same or very similar language to explain risk factors, suggesting that the informational value of this content may not always be high. Conversely, at times issuers use a variety of language, rather than strictly standard wording, to describe the availability of working capital. Our findings bring to light a substantial degree of heterogeneity across types of documents, instruments, and countries with respect to these elements.

Moreover, we found statistical evidence that longer prospectuses, all else being equal, contribute to greater divergence among rating agency assessments of credit risk. This suggests that an abundance of material can present a challenge for even specialised readers to identify information that is key to assessing the product. Given the widespread use of credit ratings throughout EU financial markets, the fact that longer documents contribute to greater divergence across ratings of the same instrument may have further downstream impacts on users of these ratings. This includes, for example, investment decisions by retail investors, capital requirements calculations (where external ratings-based approaches are used), or the risk premia of these instruments (Hsueh and Kidwell 1988; Liu and Moore 1987; Jewell and Livingston 1998).

Taken together, these findings point to a number of areas where readers may have difficulty in both gathering all of the necessary information about a financial product, and in understanding the content of the document itself.

To the extent that retail investors are the focus – for example as per Recital 83 of the Prospectus Regulation[41] – the findings discussed in this article can provide indications on the

[39] "Clean" is defined as a statement asserting clearly and concisely that sufficient working capital is available to the company over the following 12 months.

[40] "Qualified" is defined as a statement that working capital is not sufficient under the current circumstances, either including or excluding the proceeds of the offering.

[41] "(83) *In exercising its delegated and implementing powers in accordance with this Regulation, the Commission should respect the following principles:*

- *the need to ensure confidence in financial markets among retail investors and SMEs by promoting high standards of transparency in financial markets,*

…"

transparency that prospectuses bring to retail investors.

More generally, the findings in this article provide further evidence on the impact of stylistic and drafting choices (e.g. length, external references) on the interpretation of the information in the prospectus.

Importantly, our analysis illustrates the usefulness of text-mining as a supervisory technology tool. The development of algorithms capable of analysing the content of prospectuses opens new possibilities for supporting supervisory assessments, as key information which would be time intensive for the human eye to find can be extracted in seconds from lengthy documents. Such information can also be used for supervisory convergence activities, for example in peer reviews (for an example see ESMA, 2022). This facilitates the detection of anomalies, which supervisors may subsequently prioritise for manual inspection. Our study also helps illustrate the extent to which the Prospectus Regulation is implemented by issuers, and whether there is still room for improvement in certain areas. Finally, the methodologies developed can assist supervisors' ability to systematically monitor risks faced by investors in relation to specific financial instruments and how clearly and thoroughly these risks are presented in the prospectus.

Finally, the analysis proves once more the limitations of documentation requirements that do not mandate that documents be submitted in machine-readable formats. Indeed, further improvements to the format of documents collected by supervisors would unlock additional benefits (for supervisors and policymakers) of automated techniques such as the ones illustrated in this article. The widespread use of .pdf formats requires 'defrosting' to take place prior to being able to analyse a document using a computer. This step leads to substantial loss of information, as well as time-consuming efforts to re-create the structure of a document. Requiring documents to be submitted to the Prospectus Register using machine-readable formats (as defined in Article 2(13) of the Open-Data Directive) would substantially alleviate this situation.

Of course, NLP-based analyses, including the one presented here, also come with limitations. Importantly, the choice of linguistic metrics used as criteria in the text analysis determines the final outcome. This choice of terms is subject to prior selection and judgement. While this inevitably limits the objectivity of the outcome, it can best be mitigated by transparency about which choices have been made, which we provide in this article. More broadly, text length, repetitiveness, and complexity may be viewed differently by different readers. Long texts can be interpreted as unfocused, but also as comprehensive. Repetitiveness can imply low information content, but also consistency of substance across a document. Again, transparency about our analytical criteria is a key mitigant. Finally, we provide a starting point of applying NLP to issuance prospectuses. We hope future analyses will provide complementary and even competing findings and thereby further enrich our understanding of this important market.

References

Amadxarif, Z., Brookes, J., Garbarino, N., Patel, R. and Walczak, E. (2021), 'The language of rules: textual complexity in banking reforms', Bank of England Staff Working Paper, No 834.

Amzallag, A. (2021), '54 000 PRIIPs KIDs – How to read them (all)', ESMA Risk Analysis Article, February.

Celerier, C. and Vallee, B. (2017), 'Catering to investors through security design: Headline rate and complexity', *Quarterly Journal of Economics*, Vol. 132, No 2, pp. 1469–1508.

Dale, R., Moisl, H. and Somers, H. (2000), *Handbook of Natural Language Processing*, CRC Press, New York.

deHaan, E., Song, Y., Xie, C. and Zhu, C. (2021), 'Obfuscation in mutual funds', *Journal of Accounting and Economics*, Vol. 72, No 2–3.

ESMA (2021), Report: EEA prospectus activity and sanctions in 2020, July.

ESMA (2022), *Peer review of the scrutiny and approval procedures of prospectuses by competent authorities*, ESMA42-111-7170.

Ghent, A., Torous, W. N. and Valkanov, R. (2019), 'Complexity in structured finance', *Review of Economic Studies*, Vol. 86, No 2, pp. 694–722.

Gunning, R. (1952), *The Technique of Clear Writing*, McGraw-Hill, New York.

Güntay, L. and Hackbarth, D. (2010), 'Corporate bond credit spreads and forecast dispersion', *Journal of Banking and Finance*, Vol. 34, pp. 2328–2345.

Hsueh, P.L. and Kidwell, D.S. (1988), 'Bond ratings: Are two better than one?' *Financial Management*, Vol. 17, pp. 46–53.

Hyytinen, A. and Pajarinen, M. (2008), 'Evidence of Young Firm Opacity: Evidence from Split Ratings', *Journal of Banking and Finance*, Vol. 32, pp. 1234–1241.

Jewell, J. and Livingston, M. (1998), 'Split ratings, bond yields, and underwriter spreads', *Journal of Financial Research*, Vol. 21, pp. 185–204.

Kincaid, J. P., Fishburne, R. P., Rogers, R. L. and Chissom, B. S. (1975), 'Derivation of new readability formulas (automated readability index, Fog count, and Flesch reading ease formula) for Navy enlisted personnel', *Research Branch Reports*, Naval Technical Training Command, Millington TN, Research Branch.

Livingston, M., Naranjo, A. and Zhou, L. (2007), 'Asset Opaqueness and Split Bond Ratings', *Financial Management*, Vol. 36, pp. 49–62.

Livingston, M., Naranjo, A. and Zhou, L. (2008), 'Split Bond Ratings and Rating Migration', *Journal of Banking and Finance*, Vol. 32, pp. 1613–1624.

Liu, P. and Moore, W. (1987), 'The Impact of Split Bond Ratings on Risk Premia', *Financial Review*, Vol. 22, pp. 71–85.

McCarthy, P. M. and Jarvis, S. (2010), 'MTLD, VOCD-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment', *Behavior Research Methods*, Vol. 42, No 2, pp. 381–392.

Morgan, D.P. (2002), 'Rating Banks: Risk and Uncertainty in an Opaque Industry', *American Economic Review*, Vol. 92, pp. 874–888.

Neilson, J. J., Ryan, S. G., Wang, K.P. and Xie, B. (2022), 'Asset-level transparency and the (e)valuation of asset-backed securities', *Journal of Accounting Research*, Vol. 60, No 3, pp. 1131–1183.

Oakes, M.P. (1998), *Statistics for Corpus Linguistics*, Edinburgh University Press, Edinburgh.

Oxford Languages (2002), https://languages.oup.com/google-dictionary-en/ accessed on 24 October 2022.

Senter, R. J. and Smith, E. A. (1967), *Automated Readability Index*, University of Cincinnati and Aeromedical Research Laboratories, Wright-Patterson Air Force Base, AMRL-TR-6620.

Shannon, C. E., Weaver, W. and Burks, A. W. (1963), *The Mathematical Theory of Communication*, University of Illinois Press, Chicago.

Szmrecsanyi, B. (2004), 'On operationalizing syntactic complexity', in *Le Poids des Mots, Proceedings of the 7th international conference on textual data statistical analysis*, Louvain-la-Neuve, Vol. 2, pp. 1032–1039.

Tolochko, P. and Boomgaarden, H. (2019), 'Determining political text complexity: Conceptualizations, measurements, and application', International Journal of Communication, Vol. 13.

Vu, H., Alsakka, R. and ap Gwilym, O. (2017), 'What drives the differences of opinion in sovereign ratings? The roles of information disclosure and political risk', *International Journal of Finance & Economics*, Vol 22:3, pp. 216–233.

Williams, C. B. (1970), *Style and Vocabulary*, Griffin, London.

Yule, C. U. (1944), *The Statistical Study of Literary Vocabulary*, Cambridge University Press, Cambridge.

Zhang, H. H., F. Zhao and X. Zhao (2020), Working Paper, 'Neglected risks in the communication of mortgage-backed security offerings', *Proceedings of the 5th Biennial Real Estate Conference (12–13 December 2019),* Federal Reserve Bank of Atlanta.

esma European Securities and Markets Authority