UCLouvain

LOUVAIN
School of Management

Louvain School of Management

What is the impact of lexical complexity in white papers on ICO performance and the likelihood of the project to be a scam?

Author: CARELS Lucie
Supervisor: THEWISSEN James
Academic year: 2022-2023
Dissertation for the Master of Management (GEST2M)
Corporate Finance
Daytime schedule

ABSTRACT

An initial coin offering (ICO) is a digital fundraising method which is known to be highly unregulated and thus a source of uncertainty and information asymmetry for investors. The latter is problematic as ICOs are often subject to fraudulent activities, called scams. Prior literature has therefore been applying textual analysis to ICO white papers in order to determine specific textual components, such as thematic content or linguistic errors, which could be used as signals for performance and for potential fraud. This master's thesis aims at completing the existing literature by investigating lexical complexity and its components (lexical diversity, lexical sophistication and lexical density) as such signals through regression analyses. We proceed to our analysis with a database composed of 2.503 initial coin offerings issued between 2015 and 2021. We first find that lexical complexity has a positive effect on the amount raised during the coin offering period, but that this positive marginal impact is getting smaller after reaching some specific level of lexical sophistication and/or density. Moreover, we deepen our analysis and find that the additional amount raised is superior for ICOs coming from English-speaking countries for a given level of lexical complexity in the white paper, except when we consider lexical density. Finally, we find that lexical complexity cannot be used as a signal for scams, as there does not seem to be any relationship between the level of lexical complexity in the white paper and the likelihood of the project to be a scam.

Keywords : Initial coin offering, white paper, lexical complexity, lexical diversity, lexical sophistication, lexical density

First of all, I would like to thank my supervisor, James Thewissen, for his support, his expertise, and his guidance throughout the writing of this thesis.

Then, I would like to express my gratitude to one of my supervisor's PhD students, Diego Barrio Herrezuelo, for providing me with answers to my questions despite being busy with his very own research.

I also would like to say a special thank you to my family for their unconditional support during the writing of this thesis and during my whole academic journey. I would not be the person I am today if it had not been for them.

Finally, I would like to thank my friends, who made these last five years unforgettable.

Contents

1 Introduction

This master's thesis investigates the impact of lexical complexity in ICO white papers on the likelihood of the project to be a scam, and on its performance. As an additional analysis, we look at whether the relationship between lexical complexity and performance, as measured by the amount raised during the coin offering period, is impacted by the fact of being an ICO from an English-speaking country. There has been some prior literature on the analysis of lexical complexity in corporate disclosures. For instance, Joenväärä et al. (2019) find that hedge funds with more lexically diverse descriptions outperform hedge funds that are more homogeneous in their description, in terms of returns and performance, while Humpherys et al. (2011) find that fraudulent corporate disclosures tend to be less lexically diverse and to contain a poorer vocabulary. Overall, prior literature suggests that lexical complexity leads to higher performance and a smaller likelihood of being a fraudulent project.

Yet, there is a particular lack of literature tackling textual analysis in ICO white papers when it comes to analysing lexical complexity. There is evidence showing that a small level of thematic content can be used as a signal for bad project performance and potential scam (Florysiak & Schandlbauer, 2022), and that the presence of linguistic errors affects investor's decision and might lead to a decrease in the amount raised during the coin offering period, and thus to a decrease in the performance of the ICO (Thewissen et al., 2021). Nonetheless, prior literature has not investigated the potential signal lexical complexity could send investors when reading and analysing ICO white papers.

This question is important in today's research agenda for various reasons. First of all, initial coin offerings are an attractive way to raise capital for firms and entrepreneurs, which explains their success during the last few years (Liebau & Schueffel, 2019). However, contrarily to traditional fundraising methods, ICOs are highly decentralized and unregulated, which makes them risky and full of uncertainty for investors mostly because of the information asymmetry on their side. The latter is due to the scarcity of informative sources that they face, and which makes them unable to distinguish between high-quality projects and fraudulent projects (Florysiak & Schandlbauer, 2022). This is the reason why white papers, which are the primary sources of information for investors, are of the greatest importance and must be analysed through signalling theory in order to reduce the existing asymmetry of information for them.

A second reason consists of the trend that has been growing in the last years around the capacity of qualitative data to help predict whether initial coin offerings will be performant, and whether they can potentially be scams. For instance, the presence of thematic content such as technical topics (Shrestha, Thewissen & Pastwa, 2022) or even greater readability (Burke et al., 2022) have been proven to be a signal of ICO performance and of non-fraudulent project. We believe lexical complexity should also be investigated as a potential signal for investors.

A third reason is given by Joenväärä et al. (2019), who affirm that lexical diversity can be used as an estimate of cognitive ability. Indeed, skilled managers are said to use a more varied and lexically complex language, which is the result of a great memory, curiosity and precision, skills that are considered useful to outperform the market. This affirmation seems to indicate that lexical complexity in corporate disclosures could be used as a signal of cognitive ability and thus a signal of truthful and performant ICO, which is one of the reasons why we focus on that specific attribute.

We work with a database composed of 2.503 initial coin offerings issued between 2015 and 2021, of which 35 percent are classified as scams and 26 percent are coming from countries where English is the official language. In order to test our hypotheses, we proceed to three sets of regressions. A first result of our research concerns the impact of a lexically complex white paper on investors' decisions. We find that the amount raised during the coin offering period increases with the level of lexical complexity found in the ICO prospectus. The latter affirmation is also true when we investigate the three components of lexical complexity distinctively. However, we observe that the marginal effect of an increase in lexical sophistication and lexical density is positive and decreasing, meaning that the amount raised increases with our independent variable, but this increase gets smaller the more sophisticated or dense the white paper is.

A second result concerns lexical complexity and its components as signals for potential fraudulent ICOs, called scams. We find that lexical sophistication and lexical density cannot be used for such purpose. However, we report that, after reaching a specific level of lexical complexity or lexical diversity in the ICO white paper, the likelihood of an ICO to be a scam decreases, whereas those variables had no impact before reaching that specific threshold. As for our additional analysis, we find that an increase in lexical complexity has a greater effect on the amount raised for ICOs issued in English-speaking countries than for non-native projects. The

latter is also true for lexical sophistication and lexical diversity, however being an English-native does not impact the relationship between lexical density and the amount raised.

This master's thesis provides a significant contribution to the literature in two ways. First, it extends the research on textual analysis and the value of language skills in finance. There is an increasing amount of literature covering these subjects, such as Joenväärä et al. (2019) who measure linguistic ability in hedge funds' descriptions through the use of lexical diversity. The authors report that more lexically diverse disclosures increase the success of the project and may even lead to positive effects on investment decisions. Another example of textual analysis in finance is given by Boudt and Thewissen (2018), who investigate the use of sentiments and tone in CEO letters, giving some evidence of the fact such impression management has an effect on investors and readers' perception. By investigating the impact of lexical complexity on ICO performance and probability of scam, we thus complete existing literature on textual analysis in financial corporate disclosures.

Secondly, this paper is the first to investigate the role of lexical complexity in ICO white papers as a signal of potential scam and expected performance. Indeed, many studies already investigate other factors such as linguistic errors and informative content. For instance, Thewissen et al. (2021) define the presence of linguistic errors as a signal of the lack of ability of the writer, leading to a poorer performance of the project, while Florysiak and Schandlbauer (2022) investigate the informative content of such disclosures and find that the latter is negatively related with the amount of information asymmetry, meaning a more informative white paper is less likely to be a scam. This paper thus extends the research on textual analysis in ICO white papers, using lexical complexity as a new measure of linguistic ability and as a new signal for ICO scams.

In the following pages, we begin with an overview of prior literature on the themes of lexical complexity, textual analysis in ICO white papers, and regulation issues around ICOs. Based on our theoretical framework, we form our hypotheses and then present the method and database we will exploit in order to answer our research questions. Going forward, we discuss the results of the empirical analysis, as well as our interpretations, and conclude our findings with a short focus on the limitations of our models.

2 Literature review

2.1 Lexical complexity

According to Lu (2012), lexical complexity, which is an equivalent concept of lexical richness, can be measured through three main variables, namely lexical density, lexical sophistication and lexical variation. First, the author refers to lexical density as the "*ratio of the number of lexical words to the total number of words in a text*". Lexical words include nouns, adjectives, verbs except for modal and auxiliary ones, and adverbs. The latter are content words, also known as open class words, and give more information than function words such as prepositions, pronouns and conjunctions, referred to as grammatical words (Johansson, 2008). Lexical density is thus the ratio of the amount of content words to the total number of words in a text. Secondly, lexical sophistication is defined as a measure of rareness (Lu, 2012). Indeed, it relates to the amount of sophisticated, unusual and advanced lexical words in a text, the latter being defined as those that are not part of the list of the 2000 most frequent words by the British National Corpus.

Finally, Lu (2012) defines lexical variety, also known as lexical diversity, as the amount of different words used in a text or in a speech. Another similar definition is given by Jarvis (2013), who defines lexical diversity as "*the variety of words use that can be found in a person's speech or writing*". This concept therefore relates to word repetition and can be measured using two main variables : types and tokens. The type defines the number of words in a sentence or in a text, each distinct word being counted once as one type, whereas a token relates to how often each word is repeated. The latter is thus the measure of repetitiveness and is the total number of words in the text, regardless of repetitions. This definition is in accordance with the one of Joenväärä et al. (2019) who explain lexical diversity as "*the proportion of unique words produced relative to the number of total words that an individual produces in verbal or written discourse*".

However, the type-token ratio has one major issue. Indeed, these two variables do not allow to perfectly measure lexical diversity, as they do not take the length of the text into account (Jarvis, 2013). Indeed, we can consider the two following sentences : "We run every morning" and "We run up and down the slope of that hill every morning". These two sentences do not have the same number of types, as they are composed of four and twelve distinct words respectively.

However, they are both defined as being maximally diverse because the amount of types equals the number of tokens. The two sentences are thus said to be equally diverse, independently from their length. The limitation of the type-token ratio thus relates to the fact that longer texts with higher numbers of tokens result in a lower ratio (Johansson, 2008).

Lexical diversity should not be confused with readability. The latter is defined by Dale and Chall (1949) as the sum of all elements that affect the understanding, the reading speed and the interest readers can have in a specific written document. This concept thus relates to three main elements according to these authors: a typographical aspect, the interest of the reader, and the style of expression. These elements can be found again in Klare's definition (1963), as he defines readability as "*the ease of understanding or comprehension due to the style of writing*". Therefore, lexical diversity and readability are two distinct concepts. The first one relates to the repetitiveness/frequency of words in a text and to the vocabulary richness, whereas the second is a measure of the ease of comprehension of a text. Nevertheless, readability can be influenced by the lexical complexity of a written document, as highlighted by Chen & Meurers (2017). These authors affirm that word frequency is widely used as a measure for readability, as a larger repetition of words would allow readers to make less efforts in order to understand a text.

Lexical diversity has already been investigated in corporate disclosures such as hedge fund descriptions. Indeed, Joenväärä et al. (2019) use it as a measure of linguistic ability and investigate the impact of the latter on hedge funds' success and investment decisions. The authors assume that lexical diversity is a sign of cognitive ability, meaning that the more a manager employs a rich variety of vocabulary, the more he/she is expected to be skilled. Indeed, they affirm that using a richer and more varied vocabulary requires from a person to have a good memory, curiosity and a sense of precision.

Throughout their empirical study, they find that higher levels of lexical diversity in hedge funds descriptions are positively related to higher levels of returns, meaning projects with lexically diverse descriptions outperform hedge funds that lack of lexical complexity. Furthermore, these authors provide evidence that investors tend to invest more in funds that have more lexically diverse descriptions than in others. Looking in the same direction, other authors such as Humpherys et al. (2011) try to identify fraudulent financial statements through linguistic credibility analysis. They find that managers crafting fraudulent corporate disclosures tend to use less lexical diversity and thus a poorer vocabulary in their disclosures.

2.2 Prior literature on textual analysis in ICO white papers

Informative sources for investors are quite scarce concerning initial coin offerings, which is the reason why white papers matter. The latter are the primary source of information for potential investors and can be seen as an attempt at reducing the existing asymmetry of information caused by the lack of regulation of those modern fundraising (Florysiak & Schandlbauer, 2022). This prospectus is also defined as the sole document on which investors can rely for ICO investment decisions (Thewissen et al., 2022). Therefore, many prior literature has investigated whether the quality of white papers can determine the project's quality and performance.

First, Florysiak and Schandlbauer (2022) investigate the content of ICO disclosures, to find whether ICO issuers can signal their project's quality by improving the white paper's informative content, and whether it is successful to do so. They find that there is a negative relationship between information content and information asymmetry, whereas standard content is positively correlated to the latter. This means that the more informative content there is in a white paper, the more information asymmetry is reduced, thus reducing the probability of facing a scam. Shrestha, Thewissen and Pastwa (2022) also investigate white papers' content with a focus on thematic content and their findings confirm the previous observations. Indeed, through the use of topic modeling, these authors provide evidence that thematic content can be used as a signal of ICO performance. They go further by looking at each topic's impact on an ICO success and find that technical topics are positively related to the ICO success, meaning the presence of these topics increases the probability of the project to be successful. Yen, Wang & Chen (2021) confirm once more the previous findings, as they show that ICO disclosures with more unique and less common content have a higher success in terms of amount raised, and their tokens are still actively traded after the ICO campaign.

One question remains concerning the incentive that scammers would have to offer highly informative ICO disclosures in order to appear as honest projects issuers. However, Florysiak and Schandlbauer (2022) affirm such low-quality ICO issuers would be unable to replicate such information content, as it would require them to disclose more information which would reveal the true nature of their activity. There is also some evidence showing that informative content does not have the same importance at every stage of the ICO process. Indeed, Shrestha, Thewissen and Pastwa (2022) go further by investigating the importance of white papers depending on the stage of the ICO process. They show that the impact of white paper thematic content differs

depending on the period during which the white paper is read. The authors suggest that ICO prospectus are interpreted as a signal for project's quality and performance before and during the ICO process, whereas it seems to be no significant signal anymore after the fundraising, due to the fact that investors gained access to a larger pool of external sources.

Secondly, language accuracy and the absence of linguistic errors are another quality feature that has been investigated as a signal of successful ICOs and a determinant of investment decisions. Thewissen et al. (2021) empirically explore the impact of linguistic errors on investor's behaviour and investment decisions. Their methodology implies the application of language expectancy theory (LET) to ICO white papers. The latter assumes that everyone has expectations when it comes to the writing of a document, and the non-respect of those might have an impact on the message the document conveys. Indeed, this theory states that the presence of formal or grammatical mistakes influences the persuasiveness of the message and may even go as far as affecting the reader's behaviour. This is the reason why language accuracy is one of the most expected characteristics of ICO disclosures. The authors thus provide evidence that linguistic typos affect the performance of an ICO, because of the fact investors see them as a "red flag". Indeed, the authors report some evidence of a negative relationship between linguistic errors in ICO white papers and the amount the investors are willing to invest in the project. The presence of such linguistic mistakes is interpreted as a lack of skills, ability and intelligence of the writers, and this has a negative impact on the reader's perception of the quality of the information provided (Thewissen et al., 2021). Moreover, according to the authors, this penalty varies for different error types, grammatical and multiple errors seem to have a larger negative influence on the invested amount for instance. The impact is stronger for native English-speaking countries, and countries without ICO regulation.

Readability is also investigated as a quality feature of ICO white papers. Zhang et al. (2019) investigate the relationship between the readability of white papers and the ICO first-day return. They find that ICO disclosures that are more readable are more likely to result in higher first-day returns for investors. They conclude that an increase in readability decreases uncertainty and thus information asymmetry. Burke et al. (2022) confirm this finding while investigating the impact of ICO white papers' readability on investors investment decisions. Their results provide some evidence of a positive relationship between readability and ICO success. The authors give two reasons to their findings. First, knowing that ICO issuers can use white papers to hide adverse business prospect and lie about the feasibility of their business plan, an ICO white

paper that is more readable leads to less investor deception and to a decrease in information asymmetry. Then, a prospectus which is easier to read induces a higher willingness to invest on the side of the investors, because the latter needs less cognitive effort to process and comprehend the information of the white paper. Therefore, disclosures that are more readable lead to a higher likelihood of the project to be a success and a smaller likelihood of it being a scam.

Linguistic styles also can be considered as a component of white papers' quality. Monaco et al. (2021) investigate the impact of linguistic styles on ICO success, measured by the amount raised by the project. Their findings suggest a positive relationship between the use of precise language and the amount raised, and a negative relationship between the use of concrete language and more numerical terms and the amount raised. The precise linguistic style is described as highly analytical with complex and organized concepts. They also find that white papers tend to convey more positive than negative sentiments.

Boudt and Thewissen (2018) investigate impression and sentiment management in CEO letters. The paper starts from the assumption that management uses some forms of impression management to influence investors' perception of the firm's future performance in a self-serving manner. Indeed, the authors state that the information's order in corporate disclosures influences the sentiments perceived by investors, as first and last information will be remembered more easily than middle information. Their results are consistent with these assumptions, as they provide some evidence that net sentiments are on average positive, as CEOs have an interest in depicting a positive image of the firm. Negative sentiments tend to be concentrated at the beginning of the text, whereas positive ones follow a U-shaped distribution. The latter finding suggest that CEOs first begin with negative news and tend to immerse them in a large number of positive words. Overall, the net sentiment becomes more and more positive throughout the text. This study thus shows that strategic communication is a tool that managers can use to influence investors' expectations and perception.

Tone and sentiment analysis was applied to ICO white papers by Zhang et al. (2021). A common factor between ICO white paper and CEO letters is that they are both unregulated and unaudited disclosures, which means the issuer can shape the message as he/she sees fit (Boudt & Thewissen, 2018). These authors investigate whether sentiment management can be used as a signal affecting ICO success on the first trading day. They measure sentiment management through the tone used in the ICO disclosure. Their findings support the evidence

of a positive relationship between net positive tone in ICO prospectus and the first-day return of the project, meaning that sentiment management has an impact on investors' investment decisions. However, the authors also express the need for causal arguments in the white paper in order for it to stay credible and persuasive. These findings, if misused could lead to ICO issuers using sentiment management as a tool to manipulate investors' decisions and maximize the proceeds in a self-serving manner.

2.3 Initial coin offerings, information asymmetry and scams

An initial coin offering (ICO) is a digital fundraising method that involves the issuance of tokens to investors, in exchange for financing (Tiwari, Gepp & Kumar, 2020). The tokens can then either be used to access some services provided by the project issuer, or can be used as an independent cryptocurrency. ICOs are considered to be cheaper, easier and quicker to implement than traditional public offerings, such as IPOs, which is one of the reason why they are attractive (Florysiak & Schandlbauer, 2022). Momtaz (2020) affirms that initial coin offerings are attractive for entrepreneurs because they can be used to raise funds at any stage of the company, and with few transaction costs. The phenomenon is attractive for investors as well, as it gives them rapid and efficient exit options, through tokens' liquidity. However, traditional public offerings grant an ownership share to investors whereas ICOs do not (Tiwari, Gepp & Kumar, 2020).

The very first initial coin offering was conducted in 2013 by Mastercoin. The use of such a way to raise financing has then been increasing, reaching 764 realized ICO projects between January and August 2018, and more than 18 billion of US dollars collected during that same period (Fisch, 2019). However, the number of ICOs has since been decreasing. Some causes for this last observation could be the poor economic performance of those projects and the preponderant presence of scams (Liebau & Schueffel, 2019). Indeed, an important issue surrounding initial coin offerings is that they are highly decentralized and unregulated (Florysiak & Schandlbauer, 2022). Fisch (2019) affirms that there is neither a regulated platform upon which ICOs must be conducted, nor some compulsory registration of the project and its issuer. The latter and the absence of any formal disclosure requirements lead to a low amount of objective information passed from the issuer to the investor. The consequences of such lack of regulation are a large amount of uncertainty and a high degree of information asymmetry. Indeed, ICOs' lack

of regulation leads to a lack of information on the investors' side, meaning they cannot easily distinguish high-quality projects from low-quality ones (Florysiak & Schandlbauer, 2022).

This new fundraising method is thus very subject to fraudulent activities, called scams. A scam can be described as "*an act throughout which the scammer purposely deprives the trustful investor of his or her funds to advantage to the scammer*" (Liebau & Schueffel, 2019). ICO issuers have a large incentive to engage in fraudulent activities, due to the fact that they exert considerable discretionary power over capital and resources allocation. According to Fisch (2019), it is thus necessary for high-quality ICO projects to help investors distinguish them through signals. Therefore, there is a great need for signaling theory, which would consist for project issuers to send signals related to the quality of their ICO in order to reduce the information asymmetry which investors are suffering of.

3 Hypotheses development

3.1 The impact of lexical complexity on ICO performance

We can formulate our hypotheses based on prior research on lexical complexity in corporate disclosures. Based on the findings of Joenväärä et al. (2019), we can hypothesize that higher levels of lexical complexity in ICO white papers should be positively related to the success of the project. Indeed, these authors provide evidence that lexical diversity can be used as a measure of linguistic ability in hedge funds' descriptions. They find that hedge funds with more lexically complex disclosures, and thus a richer vocabulary, outperform projects without such linguistic features, leading to higher returns and thus higher success. We expect to provide some evidence that this statement is also true when it comes to initial coin offerings and white papers.

Hypothesis 1 – There is a positive curvilinear relationship between lexical complexity in ICO white papers and the amount invested in the project.

3.2 The signal lexical complexity sends to investors

Then, Humpherys et al. (2011) apply the same kind of linguistic credibility analysis to financial statements in order to detect fraudulent activities. Their initial expectation is to find

that fraudulent companies include more positive words, as well as more imagery and affect in their 10-K reports compared with truthful actors. Moreover, they assume that scammers are more likely to make their statements more complex by using longer sentences and terms. Finally, the authors expect fraudulent 10-K statements to be less lexically diverse and contain less language-specificities, while being composed of a higher number of words. In line with their hypotheses, they find that there is a lower level of lexical diversity and a poorer vocabulary in fraudulent corporate disclosures. An assumption the authors made based on their findings is that scammers try to distract readers from their fraudulent activity and from their bad results by including an excessive quantity of non-qualitative information, which would have the effect of increasing the word count in the text while diluting its diversity and complexity.

Based on these findings, we hypothesize that fraudulent ICO project issuers craft less lexically complex white papers than honest project issuers. We thus expect to find a negative relationship between lexical complexity in ICO white papers and the likelihood of the project to be a scam, meaning we assume that the likelihood of a project to be a scam decreases with an increase in the white paper's lexical complexity. Nevertheless, we decided to go further in our analysis and to include an additional hypothesis which assumes a curvilinear relationship between the two variables.

Hypothesis 2a – There is a negative relationship between the lexical complexity displayed in ICO white papers and the likelihood of this same project to be a scam.

Hypothesis 2b – There is a negative curvilinear relationship (U-shape relationship) between the lexical complexity displayed in ICO white papers and the likelihood of this same project to be a scam.

3.3 ICOs from English-speaking countries

We can go further into our analysis and question the impact of the project issuer being a native speaker on investors' perception of the project credibility, as the latter could impact their investment decision. According to Volz, Reinhard and Müller (2019), there is empirical evidence of a higher mistrust towards non-native speakers, as they are perceived as being less credible than native speakers. One reason would be the limited ability of people to express themselves in non-native languages, as they get limited in the information they give about a

specific idea. The authors add that non-native speakers also are said to take more time when it comes to lexical decisions.

Given the fact that non-native speakers have a more limited ability to express themselves, we expect a higher tolerance towards a less lexically complex writing coming from a non-native writer rather than towards a native one. This is what Rubin and William-James (1997) investigate, as they look at the difference between a teacher's evaluation of a writing coming from a native writer and from a non-native one. Contrarily to what they expect, they find that teachers do not react more strictly to non-native speakers' writings. An assumption the authors make is that this phenomenon is explained by the fact native speakers should have known better, thus leading to being more lenient with non-native writers as this is not their first language. Rubin and Willim-James' paper might not particularly refer to lexical complexity, but we nevertheless would like to apply the same reasoning to our research question. Based on these findings, we expect investors to react more harshly towards issuers coming from English-speaking countries.

Hypothesis 3 – The curvilinear relationship between lexical complexity in ICO white papers and the amount raised is stronger for ICOs from English-speaking countries.

4 Methodology and data description

4.1 Research design and data selection

The purpose of this master's thesis is to look at the signal that the presence of lexical complexity in white papers sends to investors in terms of ICO performance and success. We therefore distinguish between three components of lexical complexity, namely lexical sophistication, lexical density and lexical diversity. As our database is composed of a large number of measures for each of these three components, we will use summary indices obtained through principal component analysis in order to analyze our results in a more straightforward way. According to Ringnér (2008), principal component analysis is a mathematical algorithm that creates principal components, the latter being new variables representing linear combinations of the initial variables. A principal component is typically going in the direction which showed the sample's largest variation. In other words, if we take the example of *PC1_S*, which is a lexical sophisti-

cation principal component in our database, we know that the latter is a linear combination of variables such as *ls1* and *vs1*, which are specific measures of lexical sophistication and verb sophistication. Thus, *PC1_S* can be seen as a summary of all the lexical sophistication measures in our database, which allows us an easier analysis and manipulation of the huge number of measures that we have.

Our upcoming correlation analysis of Table 3 confirms multivariate regressions to be an appropriate tool for our research, because of the fact our variables are highly correlated between them. But before going into the regression models we build for our research, we want to replace extreme values of our main variables of interest by less extreme substitutes. We thus winsorize *pc1_all, PC1_S, PC1_D, lexical.density* and *AmountRaised*, with an interval of [0,05-0,95]. This means that we do not just trim extreme values from our database, we replace them with the 5th percentile of the sample in the case of an extremely small value and with the 95th percentile in the case of an extremely high value. This method allows us to minimize the influence of extreme values, without removing data points from our database. We will thus be working with winsorized variables, named *pc1_all_wins, PC1_S_wins, PC1_D_wins, lexical.density_wins* and *AmountRaised_wins*.

Hypothesis 1 investigates the impact of lexical complexity and its three components on the ICO performance. We test the latter through our first regression model:

$$log(1 + \textit{AmountRaised\_wins}) = \beta_0 + \beta_1 \textit{ComplexityMeasure} + \beta_2 \textit{Square} + \beta_3 \textit{Controls} + \varepsilon$$

The logarithm of *AmountRaised_wins* represents the amount collected during the coin offering period and is used as a measure of ICO performance. *ComplexityMeasure* is the independent variable and is either represented by *pc1_all_wins, PC1_S_wins, PC1_D_wins* or *lexical.density_wins*, which are the component analysis variables for lexical complexity, lexical sophistication, lexical diversity and lexical density respectively. Our second independent variable, *Square*, is equal to the square of *ComplexityMeasure*, and is included in order to test for curvilinear relationships. Finally, *Controls* is a set of control variables, which are variables that have an impact on the amount raised during the coin offering period, such as *WP_Readability* and *WP_Sentiment*. If the level of lexical complexity found in ICO white papers has a positive curvilinear impact on the ICO performance as we assume with hypothesis 1, we expect β_1 to be positive and statistically significant and β_2 to be negative and statistically significant.

Hypothesis 2 investigates the signal that the level of lexical complexity in ICO white papers sends investors when considering the likelihood of the project to be a scam. We test the latter through our second regression model:

$$\textit{Scam\_update} = \beta_0 + \beta_1 \textit{ComplexityMeasure} + \beta_2 \textit{Square} + \beta_3 \textit{Controls} + \varepsilon$$

This second model is similar to the first one apart for the dependent variable. Indeed, the latter is *Scam_update*, which is a binary variable that takes the value of one when the ICO is recognized as a fraudulent project and takes the value of zero otherwise. The set of independent and control variables is the exact same as for model 1. If we assume that hypothesis 2 is true, we expect to find that β_1 has a negative and statistically significant coefficient, while β_2 has a positive and significant coefficient.

Hypothesis 3 is a cross sectional analysis and thus requires the introduction of an interaction term, as it questions the amplification effect of the ICO being from an English-speaking country on hypothesis 1. We test this hypothesis through the following regression model:

$$\begin{aligned} log(1 + \textit{AmountRaised\_wins}) = \beta_0 &+ \beta_1 \textit{ComplexityMeasure*English} \\ &+ \beta_2 \textit{ComplexityMeasure} + \beta_3 \textit{English} \\ &+ \beta_4 \textit{Controls} + \varepsilon \end{aligned}$$

ComplexityMeasure and *Controls* still represent the same variables as for model 1 and 2. *English* is a binary variable that takes the value of 1 when the ICO is coming from an English-speaking country and 0 otherwise. If the impact of lexical complexity is stronger for ICOs coming from English-speaking countries, we expect β_1 to be positive and statistically significant.

4.2 Control variables

In order to limit the influence of confounding factors and endogeneity effects, we include control variables in our regressions. Confounding factors are variables which are not in our model but still affect both the dependent and independent variables, and which may therefore cause a lack of internal validity in our models. For instance, if our dependent variable is the amount raised and our independent variable is lexical diversity, the number of pages in the

white paper is defined as a confound because it affects both the degree of lexical diversity, and the amount investors decide to invest. Including control variables in our models is thus a solution in order to assure that we present unbiased estimates of the effect of lexical complexity measures on the performance of ICOs and on the likelihood of these ICOs to be scams. We rely on Fisch (2019), Thewissen et al. (2021) and Thewissen et al. (2022) to select a list of control variables that fit our research goals.

A first set of three control variables that we include in our models is associated with ICO white papers' attributes. First, we include *WP_pages* in order to control for the number of pages that ICO prospectuses contain. Secondly, we control for the tone of the document by including *WP_Sentiment* to our models. The latter gives a degree of net sentiment for the text, a positive variable meaning that it contains a majority of positive words. It is computed by subtracting the proportion of negative words in the text from the proportion of positive words in the text (Thewissen et al., 2022). Thirdly, we include *WP_Readability*, which quantifies the ease of reading each white paper.

The second set of control variables is based on project characteristics, as defined by Thewissen et al. (2021). First, we include *Fin_TaxHaven*, a binary variable that indicates whether the project is located in a tax haven. Secondly, we control for the number of accounts the project has on social networks, through the inclusion of *SocialMediaCount*. A third control variable that we include in this set is *Fin_Team*, which indicates the number of members working behind the ICO project. Fourthly, *Year* is used to control for the date of the project's issuance. Then, we control for the institution score as defined by Worldwide Governance Indicators (Thewissen et al., 2022), through *Fin_Institution*. Finally, we include *Fin_Rating*, which indicates a rating score attributed to the project by external experts.

Our third and last set of control variables is associated with ICO-specific characteristics and attributes, as defined by Thewissen et al. (2021). We first control for whether the ICO blockchain was built on the Ethereum platform through *Fin_EthereumBlockDummy*. Then, we also consider that compliance with a minimal set of regulations as to be controlled for, which is why we add *Fin_WhitelistKYCDummy* as a control variable. According to Thewissen et al. (2021), the simple fact of implementing a Whitelist as well as a Know Your Customer policy is a sign of regulatory compliance which may reassure investors and impact their decision. We also decided to include several dummy variables which indicate whether a minimal investment

amount is specified for investors (*Fin_MinInvestDummy*), whether the maximal amount of tokens put in circulation is communicated (*Fin_hardcapDummy*), whether a minimal amount to be raised was specified (*Fin_softcapDummy*), and whether a bonus scheme was offered to investors during the ICO (*Fin_BonusDummy*). Finally, we control for the number of currencies in which investors can exchange the project's tokens through *Fin_NumCurrencies*.

4.3 Descriptive statistics

Our database is composed of 2.503 initial coin offerings issued between 2015 and 2021, of which 876 are classified as scams by the binary variable *Scam_update*. Our sample is thus composed of 35 percent of fraudulent projects. The average amount raised for an ICO project in the sample reaches 8.178.242 dollars, with a minimum of 52.892 dollars and a maximum of 35.000.000 dollars.

In terms of lexical complexity in the white paper, the average project reaches a level of 0,032 on a scale going from -6,2 to 4,9. This means that, on average, white papers are said to be lexically complex. If we look at the three components of lexical complexity, beginning with lexical sophistication, we see that the latter has a range going from -2,8 to 2,9 in our database, with an average of -0,012. Thus, there are not much sophisticated words in our sample's prospectuses on average. The average lexical diversity in ICO white papers reaches 0,049, whereas the range goes from -5,6 to 4,3, which means that the average white paper is quite lexically diverse. Finally, lexical density in our sample goes from -1,5 to 1,7, with a mean of 0,0003. On average, white papers in our database are thus quite low on lexical density.

When looking at the descriptive statistics of our control variables in Table 2, the first observation is that 26 percent of the ICOs in our sample come from English-speaking countries, meaning that 1.841 out of 2.503 white papers have been written by non-native issuers. On average, the prospectuses of our database have a positive tone, as depicted by the positive mean of *WP_Sentiment*. The latter means that the average white paper contains more positive words than negative ones. The number of pages goes from 3 to 167 with a mean of 33 pages, and the readability of those pages is positive on average, meaning they are not too difficult to read and understand. When it comes to the localisation of our sample's projects, 34 percent are located in tax havens. The average number of social network accounts for an ICO project is a little

over 6, for a minimum of 0 and a maximum of 12. Furthermore, the team behind the project is composed of an average of 12 members.

Concerning the ICO-specific characteristics of our sample, we can observe that 85 percent of the ICO blockchains were built on the Ethereum platform, and 66 percent implement Whitelisting and Know Your Customer policies. The average number of currency options per project is slightly over 2, with a minimum of 1 and a maximum of 30. Moreover, we find that 43 percent of our sample's ICOs specify a minimum amount to invest per investors, while 86 percent communicate the maximum number of tokens to be distributed during the coin offering period, and 74 percent indicate a minimum amount to be raised for the fundraising to be validated. Finally, we find that 66 percent of our sample offered bonus schemes to investors.

Table 3 represents the correlation matrix of our dependent, independent and control variables. We can observe that the amount raised and the level of lexical complexity in white papers are positively correlated with a coefficient of 0,18, which is significant at a 99% confidence level. Moreover, correlations between the amount raised and the three components of lexical complexity, namely lexical sophistication, lexical diversity and lexical density, are equal to 0,25, 0,15 and 0,13 respectively and are all significant at a 99% confidence level. These first observations give us a first positive appreciation of hypothesis 1. If we then look at the correlation between the likelihood of being a scam and the four independent variables, we observe something a bit different. Indeed, the likelihood of being a scam seems to be positively correlated with both lexical complexity and lexical diversity, with corresponding correlations of 0,05 and 0,05 (significant at a 95% confidence level). However, lexical sophistication and lexical density do not seem to be correlated with the probability of the ICO to be a scam, as we observe insignificant correlation factors. Thus, hypothesis 2 does not seem as straightforward as hypothesis 1, as it looks like only 2 independent variables out of 4 have an impact on our dependent variable.

Finally, we find some positive and statistically significant correlations between *English* and the level of lexical sophistication and lexical density in ICO white papers. In both cases, the correlation factors are significant at a 99% confidence level. This means that white papers from English-speaking countries tend to be more lexically sophisticated and dense than white papers from non-native issuers. However, we observe a negative and statistically significant (at a 90% confidence level) correlation between *English* and the level of lexical diversity in the white

paper, meaning they are negatively related, while we find no significant correlation between *English* and lexical complexity as a whole.

5 Empiric results

5.1 Hypothesis 1 - Lexical complexity as an indicator of performance

Our first regression tests the impact of lexical complexity (integrating the three components) on the amount raised during the ICO. We observe in Table 4 that the coefficient for *pc1_all_wins* is positive and significant at a 99% confidence level, which means that the performance of an initial coin offering, and thus the amount raised for that project, is positively related to the degree of lexical complexity we find in its white paper. However, *PC1_all_square* has a negative coefficient which is not significant, meaning that the relationship between the amount raised and the complexity is not curvilinear, but strictly linear. Therefore, model 1 indicates that a higher level of lexical complexity in white papers is strictly associated with a higher amount raised during the ICO.

After testing the impact of lexical complexity as a whole, we want to investigate the impact of its three distinct components. Model 2 reports the effect of a higher level of lexical sophistication in an ICO white paper on the amount raised during that same project. We observe in Table 4 that the coefficient for *PC1_S_wins* is positive and significant at a 99% confidence level. Therefore, there seems to be a positive relationship between the level of lexical sophistication and the performance of an ICO. Moreover, *PC1_S_square* has a negative and statistically significant coefficient (at a 90% confidence level), which means that the relationship between the level of lexical sophistication and the performance is described by a reversed U-shape relationship in our model. The interpretation is that a higher level of lexical sophistication in an ICO white paper generates an additional amount raised during the coin offering period, but after reaching a specific level of sophistication, this additional amount raised starts to slightly decrease as the white paper gets more lexically sophisticated.

In order to better visualize the interpretation of our results, we plot the linear and curvilinear relationships between the level of lexical sophistication and the logarithm of the amount raised. In Figure 1, the linear relationship is represented by the red curve, whereas the curvi-

linear relationship is represented by the blue curve. We can observe that, when we add the quadratic variable in our model, the relationship between the independent variable and the ICO performance becomes curvilinear, meaning that the amount raised increases with the level of lexical sophistication, but as *PC1_S_wins* increases, its marginal effect on performance slightly decreases. Therefore, we conclude that the marginal impact of lexical sophistication on the amount raised is positive, as a more lexically sophisticated white paper leads to a higher amount raised. However, after reaching a specific threshold, the additional amount raised that relates to a small increase in lexical sophistication gets smaller.

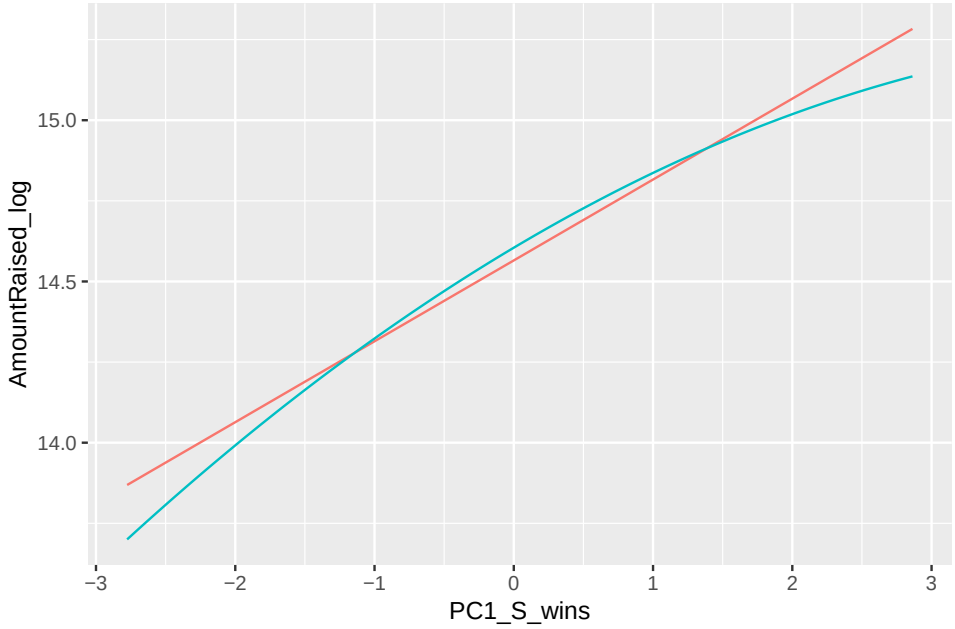

Figure 1: The linear and curvilinear impact of lexical sophistication in ICO white papers on the amount raised during the coin offering period

(*) This plot was produced with RStudio. *AmountRaised_log* is a logarithm, while *PC1_S_wins* is the degree of lexical sophistication in the white paper. The red line represents the linear relationship between the level of lexical sophistication in the white paper and the amount raised during the coin offering period, whereas the blue curve represents the curvilinear relationship between the two variables.

The third model investigates the impact of lexical diversity on the amount raised during the coin offering period. Table 4 reports that *PC1_D_wins* has a statistically significant (at a 99% confidence level) and positive coefficient, which attests of a positive relationship between the level of lexical diversity in a white paper and the performance of the ICO. However, the coefficient of *PC1_D_square* is not statistically significant, which means we reject the hypothesis of a curvilinear relationship between an ICO performance and the level of lexical diversity found in its white paper. Model 3 thus indicates that the amount raised during an ICO strictly increases when the white paper's lexical diversity increases.

Model 4 tests the impact of lexical density on ICO performance. We can observe in Table 4 that the coefficient of *lexical.density_wins* is positive and statistically significant at a 99% confidence level, which means that there is a positive relationship between the degree of lexical density found in an ICO white paper and the performance of that same project. Furthermore, *lexical.density_square* has a negative and significant coefficient (at a 90% confidence level), which attests of a curvilinear relationship between our two variables, taking the form of a reversed U-shape curve. Our interpretation is that, at first, the additional amount raised increases with a higher lexical density level, but after reaching a specific level of lexical density in the ICO white paper, the additional amount collected during the coin offering period tends to slightly decrease.

In order to visualize this result, we plot the linear (red curve) and curvilinear (blue curve) relationships between lexical density and the amount raised during the coin offering period. In Figure 2, we find something similar as for lexical sophistication. Indeed, the curve that includes the quadratic variable, namely the curvilinear curve, shows the variation in the marginal effect of an increase in lexical density on performance. We find a positive and decreasing effect of lexical density on the amount raised during the coin offering period, which means that after reaching a specific level of density, the more lexically dense the white paper gets the smaller the additional amount collected becomes. Lexical density thus has a positive but decreasing marginal impact on investors' behaviour and investment decision.

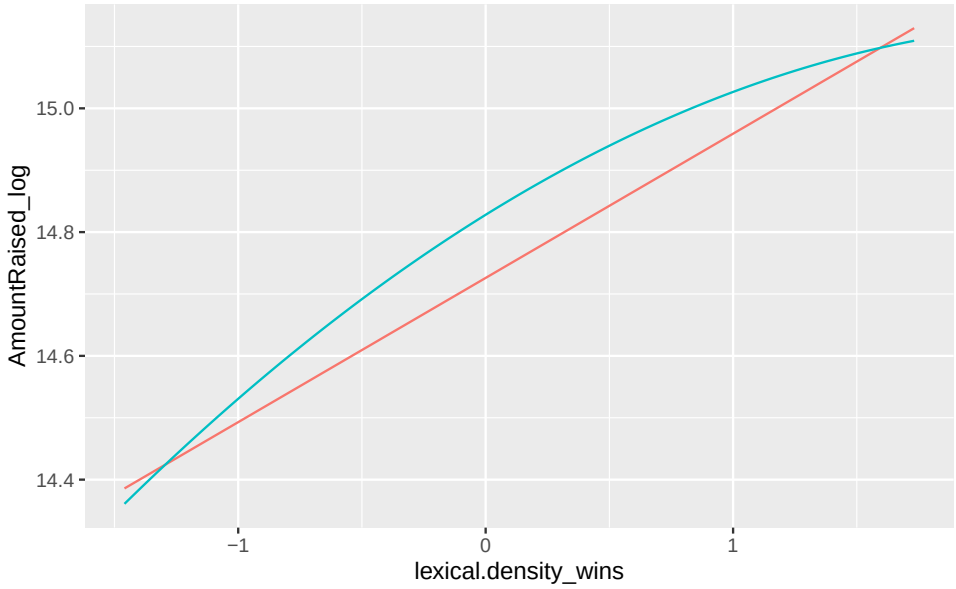

Figure 2: The linear and curvilinear impact of lexical density in ICO white papers on the amount raised during the coin offering period

(*) This plot was produced with RStudio. *AmountRaised_log* is a logarithm, while *lexical.density_wins* is the degree of lexical density in the white paper. The red line represents the linear relationship between the level of lexical density in the white paper and the amount raised during the coin offering period, whereas the blue curve represents the curvilinear relationship between the two variables.

In conclusion, our first hypothesis is that we expect to find a positive curvilinear relationship between the measure of lexical complexity we investigate, and the amount raised during the coin offering period. Model 1 to 4 report that this assumption is true when it comes to lexical sophistication and lexical density, however it must be rejected when it comes to lexical complexity and lexical diversity, as there seems to be a strictly positive linear relationship between these variables and the amount raised during the coin offering period.

5.2 Hypothesis 2 - Lexical complexity as a signal for scams

In this section, we test our second hypothesis, which is divided into two assumptions. We begin with hypothesis 2a, which assumes a negative linear relationship between the lexical complexity we find in ICO white papers and the likelihood of those ICOs to be scams. We observe in Table 5 that none of our interest variables has a significant coefficient, which is not in line with our initial expectations. We thus conclude that hypothesis 2a is rejected, as there does not seem to be a linear negative relationship between lexical complexity and the likelihood of a project to be a scam.

We can then look into hypothesis 2b, which assumes a curvilinear relationship between lexical complexity and the likelihood of an ICO project to be a scam. We expect to find that, at first, the probability of an ICO to be fraudulent decreases as the lexical complexity increases, but after a specific level of the latter, this same probability should start to increase. We start with model 5b, which looks at the impact of lexical complexity as a whole. In Table 6, we can observe that the coefficient of *pc1_all_wins* is negative and insignificant, meaning that an increase in the level of lexical complexity in white papers is said to have no direct impact on the likelihood of the ICO to be a scam. Nevertheless, if we look at the coefficient of *PC1_all_square*, we find that it is negative and significant at a 95% confidence level. Our interpretation is that, at first, an increase in lexical complexity does not have any impact on the probability of an initial coin offering to be a scam, but after reaching a specific level of lexical complexity in the white paper, the likelihood of being fraudulent slightly decreases.

Model 7b leads to very similar observations than model 5b. Indeed, in Table 6 we can observe that the coefficient for *PC1_D_wins* is negative and insignificant, which once more means that adding more lexical diversity to the white paper does not impact the likelihood of the project to be fraudulent. However, the coefficient of *PC1_D_square* is negative and significant at a 99% confidence level, which confirms our hypothesis of a curvilinear between lexical diversity and the likelihood of being a scam. The interpretation is that, under a specific level of lexical diversity in the ICO white paper, the likelihood of being a scam is not impacted, whereas after reaching that level, the likelihood of the project to be a scam decreases the more lexically diverse the white paper gets.

An assumption when it comes to the reason why we observe such results for model 5b and 7b may be that scammers excessively increase the number of irrelevant words in the white paper's text in order to hide their fraudulent activity from readers and to convince them to invest in their fake project. The latter assumption is in line with what Humpherys et al. (2011) assume when finding a negative relationship between the fact of a 10-K statement to be fraudulent and the lexical diversity found within it. Indeed, they hypothesize that fraudulent managers' strategy is to generate excessive amounts of irrelevant content to distract any reader from their bad results, which has the impact of diluting the level of lexical diversity. This is the exact same strategy that we believe fraudulent ICO issuers may be using, which in turn might be decreasing the level of lexical complexity in their disclosures. However, even when writing a document without paying attention to lexical complexity, we can assume that, on average, people tend to write

texts that are quite lexically diverse, sophisticated and dense. If we find this level of complexity to be normal for average writers, we can expect the latter to not even see that the white paper is lexically complex unless they make a real and calculated effort to reach a high enough level of it. This would explain why lexical diversity and complexity cannot be used as a signal for scams until reaching a specific level of it.

We can discuss model 6b and 8b together, as they will lead to similar observations. These two models respectively investigate the impact of an increase in lexical sophistication and lexical density levels in white papers on the likelihood of being a fraudulent project. In Table 6, we examine that both *PC1_S_wins* and *lexical.density_wins* have insignificant coefficients, such as for model 5b and 7b, which means that lexical sophistication and lexical density in white papers do not seem to be related to the likelihood of ICOs to be scams. What differs from models 5b and 7b is that *PC1_S_square* and *lexical.density_square* also have insignificant coefficients. The interpretation is that lexical sophistication and density cannot be used as a signal for potential scams at all. We thus reject hypothesis 2 for these two variables but we confirm it when it comes to lexical complexity and lexical diversity. Indeed, above-the-average-levels of lexical complexity and lexical diversity in ICO white papers can be used as a signal of non-fraudulent ICO projects.

5.3 Hypothesis 3 - ICOs from English-speaking countries

Models 9 to 12 are cross-sectional regression models, as they investigate the impact of lexical complexity, coupled with the fact of coming from an English-speaking country, on the amount raised during the coin offering period. Therefore, we use a specific interaction term for each of our model: *pc1_all_wins*English, PC1_S_wins*English, PC1_D_wins*English*, and *lexical.density_wins*English*. *English* is a binary variable, which takes the value of 1 in the case the ICO is issued in a country where English is the official language and the value of 0 otherwise. The assumption is that the relationship we found between lexical complexity and the performance of the ICO in our first four models will be even stronger in the case of ICOs coming from English-speaking countries, meaning we expect to find a positive and significant coefficient for our interaction terms.

The ninth model tests the impact of lexical complexity in ICO prospectuses coming from countries where English is the official language on the amount raised. A first observation we can

make in Table 7 is that *pc1_all_wins* still has a positive and significant coefficient, confirming the positive relationship we found between complexity and performance. However, if we look at the coefficient of *English*, we find it to be insignificant, meaning that the fact of being an ICO from an English-speaking country does not directly impact the amount raised during the coin-offering period. Our variable of interest, *pc1_all_wins*English* has a positive and significant coefficient at a 95% confidence level, which means that the additional amount raised due to more lexical complexity is even higher if the ICO has been issued in an English-speaking country. Model 9 thus confirms our third hypothesis for the case of lexical complexity as a whole.

We can then go further into the analysis and investigate the same assumption but while replacing lexical complexity with each of its three components. Model 10 reports the impact of lexical sophistication in white papers from an English-speaking country on the amount raised. We observe in Table 7 that *English* still has an insignificant coefficient, meaning being from a country where English is an official language does not make the ICO more performant. Nevertheless, if we look at our variable of interest, *PC1_S_wins*English*, we find a positive and significant coefficient at a 95% confidence level. The interpretation is that issuers having English as their native language raise a greater additional amount than non-native ones when they get more lexically sophisticated. Model 11 reports the exact same observation as for model 10, but for the case of lexical diversity. As the coefficient for *PC1_D_wins*English* is positive and significant at a 95% confidence level, the positive effect of a more lexically diverse white paper on the amount raised will be greater for ICOs in an English-speaking country than for ICOs coming from countries where English is not the native language.

The last model investigates the impact of the ICO being issued from an English-speaking country on the effect lexical density has on the performance of that ICO. The observation we can make in Table 7 is different from the ones we made for the previous three models. Indeed, the coefficient of *lexical.density_wins*English* is insignificant, which means that the amount raised during the coin offering period is equivalent for ICOs that come from an English-speaking country and those that do not, given a specific level of lexical density in the project's white paper. We thus conclude that hypothesis 3 is confirmed when it comes to lexical sophistication, lexical diversity and lexical complexity in general, whereas it is rejected for lexical density. Indeed, ICOs coming from English-speaking countries that increase their level of complexity, sophistication or diversity seem to be getting a bonus in terms of amount raised compared to non-natives ICOs with the exact same level of lexical complexity. Investors thus tend to reward

English-speaking issuers that integrate a greater level of complexity in their white paper, rather than to punish them compared to non-native speakers when the level is not high enough. The conclusion is that, when they get more lexically complex, diverse or sophisticated, native issuers get a bonus in terms of amount raised compared with non-native issuers.

6 Conclusion

This master's thesis builds on prior literature which has been striving to determine which textual elements in ICO white papers can be used to signal performant projects and fraudulent activities. Thematic content (Florysiak & Schandlbauer, 2022), readability (Zhang et al, 2019) and linguistic errors (Thewissen et al., 2021) are some elements that have already been investigated as such. This paper aims at completing the existing literature by investigating whether lexical complexity and its three components, namely lexical sophistication, lexical diversity, and lexical density, can be used as signals of ICO performance and of scam.

Based on previous literature, we formulate three main hypotheses relating to our research question. First, we expect to find a positive curvilinear relationship between lexical complexity and ICO performance, proxied by the amount raised during the coin offering period. Secondly, we assume this positive relationship to be even stronger when it comes to ICOs coming from English-speaking countries. Third, we expect to find a negative curvilinear relationship between the level of lexical complexity found in the white paper and the likelihood of the project to be a scam, meaning we hypothesize that the probability of an ICO to be a scam decreases as the level of lexical complexity goes up to a specific threshold, and then it tends to slightly increase.

We run three main sets of regressions based on a database composed of 2.503 ICOs. Our first relevant finding is that the level of lexical complexity present in the ICO white paper can be used as a signal of performance. Indeed, in this paper, we use the total amount raised during the coin offering period as a proxy for project performance. Our models report a positive relationship between lexical complexity and the amount raised, meaning that the more lexically complex the ICO white paper is, the more successful the ICO project is in terms of amount collected. The latter is also true when it comes to lexical diversity. Furthermore, we find some evidence that this relationship is even more complex when it comes to lexical sophistication and lexical density. Indeed, these two variables are positively related to performance as well, however their

marginal impact decreases as their level increases. The latter means that the additional amount raised during the coin offering period is increasing until reaching a specific threshold of lexical sophistication/density, after which the additional amount starts to slightly decrease. Thus, the level of lexical complexity (and its components) found in the white paper seems to have a direct impact on the investor's behaviour regarding the project.

We then proceed to a cross-sectional analysis in order to determine whether the linguistic origin of ICO projects has an impact on the relationship we find between lexical complexity and performance. We expect to find that ICOs coming from English-speaking countries benefit from a stronger impact of a high level of lexical complexity on the amount raised during the coin offering period. In other terms, given a specific level of lexical complexity, ICO projects coming from English-speaking countries raise a bigger additional amount than non-native ICOs. This finding is true when it comes to lexical complexity, lexical sophistication and lexical diversity, whereas it is not true for lexical density, for which the level does not seem to lead to any discrimination between native and non-native ICOs.

Finally, we investigate lexical complexity and its components as potential signals of fraudulent activities. We find that all our variables of interest have unsignificant coefficients, meaning that none of them can be used to directly determine whether an ICO project is a potential scam. However, we also find that reaching a specifically high level of lexical complexity or lexical diversity in the white paper decreases the likelihood of the project to be fraudulent. An assumption we make concerning this finding relates to the observation of Humpherys et al. (2011), who hypothesize that fraudulent managers tend to generate excessive amounts of irrelevant content to distract investors from their fraudulent activity, which may have the impact of decreasing the lexical diversity found in their written statements. We thus believe that what we consider to be a "normal" level of lexical complexity/diversity is not impactful, but that reaching specifically high levels of these two variables might only be possible for honest ICO issuers that do not try to distract their readers from any dishonest behaviour.

However, our research is subject to some limitations. First, in this paper, we assume that our independent variables are exogenous. The latter means that we do not take into consideration the fact that some omitted variables may have an impact on the relationship between our independent and dependent variables. For instance, it could be that the linguistic abilities of the project's issuer influence the level of lexical complexity found in their white papers, thus indi-

rectly influencing the dependent variable. Such an endogeneity issue could lead to difficulties in validating the direction and magnitude of our causal relationships. However, the likelihood of facing any endogeneity issue in this paper is largely reduced by our extensive choice of control variables. Indeed, we include a total of 16 relevant control variables in our models, which aim at reducing the impact of a potential omitted variable bias.

Secondly, we are exposed to a lack of longitudinal data. Indeed, our paper is based on cross-sectional data, meaning that the data we use is a snapshot of the state of ICO fundraising projects after closure. Thus, this means that the variable representing the amount raised during the ICO coin offering period is the total amount collected at the end of the fundraising, and we are missing data that relates to intermediates phases of the fundraising process. Longitudinal data would be a way to have a more complete understanding of the impact of lexical complexity on the performance of an ICO during all the stages that are composing its process, from pre-ICO to post-ICO stages. Based on this second limitation, we suggest that future researches in this field analyse the impact of investigated textual signals in ICO white papers, such as linguistic errors, readability and lexical complexity, on ICO performance at different stages of the projects. This would result in having an idea of what stages of this fundraising method are more likely to be impacted by these textual elements. The latter has already been done with thematic content. Indeed, Thewissen et al. (2022) find that the impact of thematic content on the project's performance decreases and almost disappears in the post-ICO period because of the fact investors are gaining access to a larger range of information. Extending this paper with a research using longitudinal data would thus be useful in completing the existing knowledge about ICO processes and performance.

7 References

Boudt, K., & Thewissen, J. (2018). Jockeying for position in CEO letters: impression management and sentiment analysis. *Financial management, 48*(1), 77-115. doi: 10.1111/fima.12219

Burke, Q., Li, B., Wan, C., & Wang, Y. (2022). *Disclosure readability in unregulated capital markets: Evidence from initial coin offerings*. Manuscript submitted for publication. doi: 10.2139/ssrn.4173578

Chen, X., & Meurers, D. (2017). Word frequency and readability: Predicting the text-level readability with a lexical-level attribute. *Journal of Research in Reading, 41*(3), 486-510. doi: 10.1111/1467-9817.12121

Dale, E, & Chall, J.S. (1949). The concept of readability. *Elementary English, 26*(1), 19-28. https://www.jstor.org/stable/41383594

Fisch, C. (2019). Initial coin offerings (ICOs) to finance new ventures. *Journal of business Venturing, 34*(1), 1-22. doi: 10.1016/j.jbusvent.2018.09.007

Florysiak, D., & Schandlbauer, A. (2022). Experts of charlatans? ICO analysts and white paper informativeness. *Journal of Banking and Finance, 139*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3265007

Humpherys, S.L., Moffitt, K.C., Burns, M.B., Burgoon, J.K., & Felix, W. (2011). Identification of fraudulent financial statements using linguistic credibility analysis. *Decision Support Systems, 50*(3), 585-594. doi: 10.1016/j.dss.2010.08.009

Jarvis, S. (2013). Capturing the diversity in lexical diversity. *Language Learning, 63*(1), 87-106. doi: 10.1111/j.1467-9922.2012.00739.x

Joenväärä, J., Karppinen, J., Teo, M., & Tiu, C.I. (2019). The vocabulary of hedge funds. *Capital Markets: Capital Market Efficiency eJournal*. doi: 10.2139/ssrn.3438758

Johansson, V. (2008). Lexical diversity and lexical density in speech and writing: a developmental perspective. *Lund University Department of Linguistics Working Papers, 53*, 61-79. https://fr.scribd.com/document/498869410/Johansson-2008-Lexical-Diversity-an-Lexical-Density-in-Speech#

Klare, G.R. (1963). The measurement of readability. *Iowa State University Press*.

Liebau, D., & Schueffel, P. (2019). Cryptocurrencies & Initial Coin Offerings : Are they scams ? – An empirical study. *The JBBA, 2*(1). doi: 10.31585/jbba-2-1-(5)2019

Lu, X. (2012). The relationship of lexical richness to the quality of ESL learners' oral narratives. *The Modern Language Journal, 96*(2), 190-208. doi: 10.1111/j.1540-4781.2011.01232

Momtaz, P.P. (2020). Initial Coin Offerings. *Plos One*. doi : 10.1371/journal.pone.0233018

Monaco, E., Onesti, G., Cruz, D., & Rosati, P. (2021). It's not only what you say but "How" you say it: linguistic styles and ICOs success. In Spagnoletti, P., De Marco, M., Pouloudi, N., Te'eni, D., Vom Brocke, J., Winter, R., & Baskerville, R. (Eds), *Lecture Notes in Information Systems and Organisation*, 109-121. Springer. doi: 10.1007/978-3-030-87842-9

Ringnér, M. (2008). What is principal component analysis ?. *Nature Biotechnology, 26*, 303-304. https://www.nature.com/articles/nbt0308-303

Rubin, D.L., & Williams-James, M. (1997). The impact of writer nationality on mainstream teacher's judgments of composition quality. *Journal of Second Language Writing, 6*(2), 139-154. https://doi.org/10.1016/S1060-3743(97)90031-X

Thewissen, J., Shrestha, P., Torsin, W. & Pastwa, A.M. (2022). Unpacking the black box of ICO white papers: a topic modeling approach. *Journal of Corporate finance, 75*. doi: 10.1016/j.jcorpfin.2022.102225

Thewissen, J., Thewissen, J., Torsin, W., & Arslan-Ayaydin, Ö. (2021). Linguistic errors and investment decisions : The case of ICO white papers. *The European Journal of Finance*. doi: 10.1080/1351847X.2022.2075780

Tiwari, M., Gepp, A., & Kumar, K. (2020). The future of raising finance – a new opportunity to commit fraud : a review of initial coin offering (ICOs) scams. *Crime, Law and Social Change, 73*(4), 417-441. doi: 10.1007/s10611-019-09873-2

Volz, S., Reinhard, M-A., & Müller, P. (2019). Why don't you believe me? Detecting deception in messages written by nonnative and native speakers. *Applied Cognitive Psychology, 34*(1), 256-269. https://doi.org/10.1002/acp.3615

Yen, J-C., Wang, T., & Chen, Y-H. (2021). Different is better : How unique initial coin offering language in white papers enhances success. *Accounting & Finance, 61*(4), 5309-5340. doi: 10.1111/acfi.12760

Zhang, S., Aerts, W., Lu, L., & Pan, H. (2019). Readability of token whitepaper and ICO first-day return. *Economic Letters, 180*, 58-61. doi: 10.1016/j.econlet.2019.04.010

Zhang, S., Aerts, W., Zhang, D., & Chen, Z. (2021). Positive tone and initial coin offering. *Accounting & Finance, 6*(2), 2237-2266. doi: 10.1111/acfi.12860

8 Appendixes

Table 1: Variable Definition

Dependent variables	
AmountRaised	Amount raised during the coin offering period in US dollars
Scam_update	Binary variable that takes the value of 1 if the project is a scam, and 0 otherwise
Independent variables	
pc1_all	Measure of lexical complexity integrating lexical sophistication, lexical diversity and lexical density
PC1_all_square	Square of *pc1_all*
PC1_S	Measure of lexical sophistication
PC1_S_square	Square of *pc1_S*
PC1_D	Measure of lexical diversity
PC1_D_square	Square of *pc1_D*
lexical.density	Measure of lexical density
lexical.density_square	Square of *lexical.density*
English	Binary variable that takes the value of 1 when the project comes from an English-speaking country, and 0 otherwise
Control variables	
WP_pages	The number of pages in the white paper
WP_Sentiment	Measure of the white paper's tone, which is given by the proportion of positive words minus the proportion of negative words
WP_Readability	Measure that quantifies the ease of reading the white paper
Fin_TaxHaven	Binary variable that takes the value of 1 if the project is located in a tax haven, and 0 otherwise
SocialMediaCount	The number of accounts the ICO project has on social networks
Fin_Team	The number of members behind the ICO project

Year	Year of the ICO project's issuance
Fin_Rating	Rating attributed to the ICO project by external experts
Fin_Institution	Aggregated institution score based on Worldwide Governance Indicators (Thewissen et al., 2021)
Fin_EthereumBlockDummy	Dummy variable that takes the value of 1 if the project blockchain is built on the Ethereum platform (Thewissen et al., 2021)
Fin_WhitelistKYCDummy	Dummy variable that takes the value of 1 when the ICO implements whitelisting and KYC (Know Your Customer) Compliances, and 0 otherwise (Thewissen et al., 2021)
Fin_MinInvestDummy	Dummy variable that takes the value of 1 when a minimum amount to invest is specified, and 0 otherwise (Thewissen et al., 2021)
Fin_NumCurrencies	The number of currencies in which the project's tokens can be exchanged
Fin_hardcapDummy	Dummy variable that takes the value of 1 when a maximal amount of tokens put into circulation is specified, and 0 otherwise
Fin_softcapDummy	Dummy variable that takes the value of 1 when a minimum amount to raise for the project is specified, and 0 otherwise (Thewissen et al., 2021)
Fin_BonusDummy	Dummy variable that takes the value of 1 if investors were given a bonus scheme during the ICO, and 0 otherwise (Thewissen et al., 2021)

Table 2: Summary Statistics

Variable	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max
AmountRaised_wins	8178242	9992461	52892	1000000	12000000	35000000
Scam_update	0.35	0.48	0	0	1	1
pc1_all_wins	0.032	3.1	-6.2	-2.1	2.5	4.9
PC1_S_wins	-0.012	1.6	-2.8	-1.3	1.2	2.9
PC1_D_wins	0.049	2.7	-5.6	-1.8	2.2	4.3
lexical.density_wins	0.0003	0.88	-1.5	-0.75	0.67	1.7
English	0.26	0.44	0	0	1	1
WP_Sentiment	0.0025	0.0065	-0.025	-0.0012	0.0065	0.03
WP_pages	33	17	3	21	42	167
WP_Readability	0.00013	1	-3.7	-0.68	0.64	3.4
Fin_TaxHaven	0.34	0.47	0	0	1	1
SocialMediaCount	6.4	2.1	0	5	8	12
Fin_Team	12	7.5	1	7	16	69
Year	2018	0.59	2015	2018	2018	2021
Fin_Rating	0.13	0.77	-2.3	-0.44	0.68	2.2
Fin_Institution	2.4	1.8	-4.3	1.7	3.7	4.4
Fin_EthereumBlockDummy	0.85	0.35	0	1	1	1
Fin_WhitelistKYCDummy	0.66	0.47	0	0	1	1
Fin_MinInvestDummy	0.43	0.5	0	0	1	1
Fin_NumCurrencies	2.3	1.8	1	1	3	30
Fin_hardcapDummy	0.86	0.35	0	1	1	1
Fin_softcapDummy	0.74	0.44	0	0	1	1
Fin_BonusDummy	0.66	0.47	0	0	1	1

** This table provides the summary statistics of the dependent variables, independent variables and control variables discussed in this paper. AmountRaised_wins is expressed in dollars, while Scam_update, English, Fin_TaxHaven and all the dummy variables are binary variables. Those same variables are defined in Table 1. This table has been produced with RStudio.*

Table 3: Correlation table between independent, dependent and control variables

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
(1)AmountRaised_wins											
(2)Scam_update	-0.13***										
(3)pc1_all_wins	0.18***	0.05**									
(4)PC1_S_wins	0.25***	0.02	0.78***								
(5)PC1_D_wins	0.15***	0.05**	0.98***	0.65***							
(6)lexical.density_wins	0.13***	0.01	0.11***	0.24***	0.06***						
(7)English.	0.02	0.01	-0.01	0.06***	-0.03*	0.05***					
(8)WP_Sentiment	-0.06**	-0.01	-0.13***	-0.15***	-0.12***	0.13***	-0.01				
(9)WP_pages	0.08***	0.04*	0.72***	0.55***	0.70***	0.08***	-0.05**	-0.07***			
(10)WP_Readability	0.07**	0.00	0.23***	0.33***	0.18***	0.14***	0.06***	-0.05***	0.15***		
(11)Fin_TaxHaven	0.11***	-0.01	0.17***	0.19***	0.15***	0.11***	-0.24***	-0.04*	0.13***	0.08***	
(12)SocialMediaCount	-0.06**	0.06***	0.17***	0.11***	0.18***	0.03	-0.03	-0.03*	0.18***	0.00	0.05**
(13)Fin_Team	0.12***	0.08***	0.39***	0.35***	0.37***	0.13***	-0.05**	0.02	0.37***	0.11***	0.13***
(14)Year	-0.07**	-0.08***	0.05**	0.08***	0.04*	0.03	0.01	-0.04*	0.12***	0.05**	0.05**
(15)Fin_Rating	0.08***	0.10***	0.31***	0.25***	0.30***	0.07***	-0.01	-0.02	0.27***	0.05***	0.10***
(16)Fin_Institution	0.14***	0.01	0.16***	0.20***	0.13***	0.10***	0.22***	-0.05**	0.07***	0.04**	0.31***
(17)Fin_EthereumBlockDummy	-0.02	0.02	0.05***	0.01	0.06***	0.04*	-0.03	-0.08***	0.05***	-0.02	0.05**
(18)Fin_WhitelistKYCDummy	0.07**	0.04**	0.26***	0.24***	0.25***	0.12***	-0.02	-0.01	0.22***	0.09***	0.14***
(19)Fin_MinInvestDummy	-0.04	0.02	0.05***	0.03	0.06***	0.01	-0.02	-0.05***	0.09***	0.00	0.03
(20)Fin_NumCurrencies	0.03	-0.01	0.05**	0.04*	0.04**	0.00	0.02	0.05**	0.08***	0.02	-0.03
(21)Fin_hardcapDummy	0.00	0.08***	0.12***	0.09***	0.12***	0.04*	-0.03	-0.04**	0.13***	0.05**	0.08***
(22)Fin_softcapDummy	-0.01	0.06***	0.10***	0.04**	0.11***	-0.01	-0.05***	-0.04**	0.10***	0.00	0.06***
(23)Fin_BonusDummy	-0.04	0.12***	0.08***	0.01	0.08***	-0.02	-0.01	0.00	0.09***	-0.05**	-0.02

	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)	(21)	(22)
(13)Fin_Team	0.30***										
(14)Year	0.09***	-0.01									
(15)Fin_Rating	0.59***	0.41***	0.10***								
(16)Fin_Institution	0.01	0.08***	0.05***	0.08***							
(17)Fin_EthereumBlockDummy	0.15***	0.03	0.00	0.10***	0.03*						
(18)Fin_WhitelistKYCDummy	0.24***	0.22***	0.19***	0.32***	0.15***	0.06***					
(19)Fin_MinInvestDummy	0.24***	0.10***	0.11***	0.20***	0.02	0.13***	0.15***				
(20)Fin_NumCurrencies	0.09***	0.07***	0.08***	0.12***	-0.04*	-0.14***	0.06***	0.06***			
(21)Fin_hardcapDummy	0.25***	0.14***	0.12***	0.29***	0.06***	0.13***	0.24***	0.21***	0.07***		
(22)Fin_softcapDummy	0.25***	0.12***	0.12***	0.27***	0.01	0.13***	0.19***	0.22***	0.08***	0.48***	
(23)Fin_BonusDummy	0.21***	0.13***	0.00	0.17***	0.00	0.06***	0.11***	0.16***	0.12***	0.17***	0.19***

** This table provides the correlation matrix between the dependent variables, independent variables and control variables discussed in this paper. It shows the Pearson correlation coefficients with significance levels of 99% (***), 95% (**) and 90% (*). This table has been produced with RStudio.*

Table 4: Regression results - Hypothesis 1

	Dependent variable:			
	log(1 + AmountRaised_wins)			
	(1)	(2)	(3)	(4)
pc1_all_wins	0.104^{***}			
	(0.026)			
PC1_all_square	-0.003			
	(0.005)			
PC1_S_wins		0.257^{***}		
		(0.040)		
PC1_S_square		-0.025^{*}		
		(0.013)		
PC1_D_wins			0.080^{***}	
			(0.029)	
PC1_D_square			-0.006	
			(0.004)	
lexical.density_wins				0.248^{***}
				(0.059)
lexical.density_square				-0.049^{*}
				(0.029)
WP_Sentiment	-4.218	-1.858	-5.803	-13.036^{*}
	(7.680)	(7.616)	(7.678)	(7.734)
WP_pages	-0.005	-0.003	-0.001	0.005^{*}
	(0.004)	(0.004)	(0.004)	(0.003)
WP_Readability	0.104^{**}	0.048	0.126^{**}	0.108^{**}
	(0.051)	(0.051)	(0.050)	(0.050)
Fin_TaxHaven	0.167	0.158	0.176	0.165
	(0.110)	(0.109)	(0.110)	(0.110)
SocialMediaCount	-0.097^{***}	-0.090^{***}	-0.099^{***}	-0.106^{***}
	(0.033)	(0.033)	(0.033)	(0.033)
Fin_Team	0.031^{***}	0.027^{***}	0.032^{***}	0.032^{***}
	(0.007)	(0.007)	(0.007)	(0.007)

as.factor(Year)	Yes	Yes	Yes	Yes
Fin_Rating	0.383***	0.368***	0.397***	0.419***
	(0.096)	(0.095)	(0.097)	(0.096)
Fin_Institution	0.131***	0.111***	0.137***	0.131***
	(0.028)	(0.028)	(0.028)	(0.028)
Fin_EthereumBlockDummy	−0.086	−0.039	−0.097	−0.102
	(0.160)	(0.158)	(0.160)	(0.159)
Fin_WhitelistKYCDummy	0.256**	0.235**	0.267**	0.267**
	(0.119)	(0.118)	(0.120)	(0.120)
Fin_MinInvestDummy	−0.082	−0.072	−0.086	−0.076
	(0.104)	(0.103)	(0.104)	(0.104)
Fin_NumCurrencies	0.042	0.040	0.043*	0.038
	(0.026)	(0.026)	(0.026)	(0.026)
Fin_hardcapDummy	−0.069	−0.032	−0.081	−0.132
	(0.199)	(0.197)	(0.200)	(0.199)
Fin_softcapDummy	−0.085	−0.103	−0.077	−0.051
	(0.142)	(0.141)	(0.143)	(0.142)
Fin_BonusDummy	−0.062	−0.045	−0.068	−0.046
	(0.115)	(0.114)	(0.115)	(0.115)
Constant	15.209***	14.605***	15.081***	14.828***
	(1.719)	(1.692)	(1.723)	(1.712)
Observations	1,202	1,202	1,202	1,202
R^2	0.149	0.165	0.144	0.148
Adjusted R^2	0.132	0.149	0.128	0.132
Residual Std. Error (df = 1178)	1.680	1.664	1.685	1.681
F Statistic (df = 23; 1178)	8.941***	10.126***	8.634***	8.914***

** This table provides the results of the set of regressions answering hypothesis 1, which assumes a positive curvilinear relationship between the level of lexical complexity found in the white paper and the amount raised during the coin offering period. Model 1 takes lexical complexity as independent variable, while model 2, 3 and 4 take lexical sophistication, lexical diversity and lexical density as respective independent variables. The 4 models include a quadratic variable. ***, **, * represent the significance level of the resulting coefficient, at a 99%, 95% and 90% confidence level respectively. The significance test applied in this table is the two-sided t-test. This table has been produced with RStudio.*

Table 5: Regression results - Hypothesis 2a

	Dependent variable:			
	Scam_update			
	(5a)	(6a)	(7a)	(8a)
pc1_all_wins	0.004			
	(0.014)			
PC1_S_wins		−0.008		
		(0.022)		
PC1_D_wins			0.003	
			(0.015)	
lexical.density_wins				0.005
				(0.032)
WP_Sentiment	−1.603	−1.986	−1.692	−1.890
	(4.199)	(4.204)	(4.188)	(4.214)
WP_pages	−0.0005	0.0003	−0.0003	0.00002
	(0.002)	(0.002)	(0.002)	(0.002)
WP_Readability	−0.010	−0.006	−0.009	−0.009
	(0.028)	(0.029)	(0.027)	(0.028)
Fin_TaxHaven	−0.099	−0.097	−0.098	−0.099
	(0.060)	(0.060)	(0.060)	(0.060)
SocialMediaCount	−0.005	−0.005	−0.005	−0.005
	(0.017)	(0.017)	(0.017)	(0.017)
Fin_Team	0.006	0.006	0.006	0.006
	(0.004)	(0.004)	(0.004)	(0.004)
Year	Yes	Yes	Yes	Yes
Fin_Rating	0.122***	0.125***	0.123***	0.124***
	(0.047)	(0.047)	(0.047)	(0.047)
Fin_Institution	0.004	0.005	0.004	0.004
	(0.016)	(0.016)	(0.016)	(0.016)
Fin_EthereumBlockDummy	0.019	0.018	0.018	0.018
	(0.080)	(0.080)	(0.080)	(0.080)

Fin_WhitelistKYCDummy	−0.083	−0.078	−0.082	−0.081
	(0.064)	(0.064)	(0.064)	(0.064)
Fin_MinInvestDummy	−0.090	−0.092	−0.090	−0.091
	(0.057)	(0.057)	(0.057)	(0.057)
Fin_NumCurrencies	−0.021	−0.021	−0.021	−0.021
	(0.015)	(0.015)	(0.015)	(0.015)
Fin_hardcapDummy	0.173*	0.171*	0.172*	0.172*
	(0.092)	(0.092)	(0.092)	(0.092)
Fin_softcapDummy	0.030	0.030	0.030	0.031
	(0.072)	(0.072)	(0.072)	(0.072)
Fin_BonusDummy	0.230***	0.230***	0.230***	0.231***
	(0.060)	(0.061)	(0.060)	(0.060)
Constant	−5.236	−5.274	−5.248	−5.261
	(146.954)	(146.954)	(146.954)	(146.954)
Observations	2,503	2,503	2,503	2,503
Log Likelihood	−1,497.865	−1,497.858	−1,497.898	−1,497.904
Akaike Inf. Crit.	3,041.729	3,041.715	3,041.797	3,041.808

** This table provides the results of the set of regressions answering hypothesis 2a, which assumes a negative linear relationship between the level of lexical complexity found in the white paper and the likelihood of the ICO to be a scam. Model 5a takes lexical complexity as independent variable, while model 6a, 7a and 8a take lexical sophistication, lexical diversity and lexical density as respective independent variables. ***, **, * represent the significance level of the resulting coefficient, at a 99%, 95% and 90% confidence level respectively. The significance test applied in this table is the two-sided t-test. This table has been produced with RStudio.*

Table 6: Regression results - Hypothesis 2b

	Dependent variable:			
	Scam_update			
	(5b)	(6b)	(7b)	(8b)
pc1_all_wins	−0.007			
	(0.014)			
PC1_all_square	−0.007**			
	(0.003)			
PC1_S_wins		−0.007		
		(0.022)		
PC1_S_square		−0.004		
		(0.007)		
PC1_D_wins			−0.015	
			(0.016)	
PC1_D_square			−0.007***	
			(0.002)	
lexical.density_wins				0.010
				(0.032)
lexical.density_square				−0.025
				(0.017)
WP_Sentiment	−1.954	−1.945	−1.966	−1.637
	(4.204)	(4.205)	(4.198)	(4.217)
WP_pages	0.002	0.0005	0.002	−0.0001
	(0.003)	(0.002)	(0.003)	(0.002)
WP_Readability	−0.008	−0.007	−0.010	−0.010
	(0.028)	(0.029)	(0.028)	(0.028)
Fin_TaxHaven	−0.093	−0.096	−0.088	−0.096
	(0.060)	(0.060)	(0.060)	(0.060)
SocialMediaCount	−0.005	−0.005	−0.006	−0.005
	(0.017)	(0.017)	(0.017)	(0.017)
Fin_Team	0.006	0.006	0.006	0.006
	(0.004)	(0.004)	(0.004)	(0.004)

Year	Yes	Yes	Yes	Yes
Fin_Rating	0.119**	0.124***	0.121**	0.119**
	(0.047)	(0.047)	(0.047)	(0.047)
Fin_Institution	0.004	0.005	0.004	0.004
	(0.016)	(0.016)	(0.016)	(0.016)
Fin_EthereumBlockDummy	0.009	0.016	0.006	0.017
	(0.080)	(0.080)	(0.080)	(0.080)
Fin_WhitelistKYCDummy	−0.085	−0.079	−0.084	−0.077
	(0.064)	(0.064)	(0.064)	(0.064)
Fin_MinInvestDummy	−0.085	−0.091	−0.084	−0.085
	(0.058)	(0.057)	(0.058)	(0.057)
Fin_NumCurrencies	−0.020	−0.021	−0.020	−0.021
	(0.015)	(0.015)	(0.015)	(0.015)
Fin_hardcapDummy	0.169*	0.172*	0.169*	0.168*
	(0.093)	(0.093)	(0.093)	(0.093)
Fin_softcapDummy	0.028	0.029	0.026	0.031
	(0.072)	(0.072)	(0.072)	(0.072)
Fin_BonusDummy	0.225***	0.228***	0.224***	0.229***
	(0.061)	(0.061)	(0.061)	(0.061)
Constant	−5.154	−5.269	−5.183	−5.212
	(146.954)	(146.954)	(146.954)	(146.954)
Observations	2,503	2,503	2,503	2,503
Log Likelihood	−1,494.674	−1,497.707	−1,493.059	−1,496.830
Akaike Inf. Crit.	3,037.347	3,043.414	3,034.119	3,041.660

** This table provides the results of the set of regressions answering hypothesis 2b, which assumes a negative curvilinear relationship (U-shape relationship) between the level of lexical complexity found in the white paper and the likelihood of the ICO to be a scam. Model 5b takes lexical complexity as independent variable, while model 6b, 7b and 8b take lexical sophistication, lexical diversity and lexical density as respective independent variables. The 4 models include a quadratic variable. ***, **, * represent the significance level of the resulting coefficient, at a 99%, 95% and 90% confidence level respectively. The significance test applied in this table is the two-sided t-test. This table has been produced with RStudio.*

Table 7: Regression results - Hypothesis 3

	Dependent variable:			
	log(1 + AmountRaised_wins)			
	(9)	(10)	(11)	(12)
pc1_all_wins	0.084***			
	(0.027)			
PC1_S_wins		0.210***		
		(0.044)		
PC1_D_wins			0.066**	
			(0.029)	
lexical.density_wins				0.243***
				(0.065)
English1	0.037	0.012	0.045	0.043
	(0.124)	(0.123)	(0.125)	(0.125)
WP_Sentiment	−4.662	−2.060	−6.164	−13.558*
	(7.670)	(7.615)	(7.673)	(7.742)
WP_pages	−0.006	−0.004	−0.003	0.006*
	(0.004)	(0.004)	(0.004)	(0.003)
WP_Readability	0.097*	0.049	0.119**	0.108**
	(0.051)	(0.051)	(0.050)	(0.051)
Fin_TaxHaven	0.201*	0.190*	0.206*	0.166
	(0.116)	(0.115)	(0.116)	(0.117)
SocialMediaCount	−0.097***	−0.090***	−0.100***	−0.109***
	(0.033)	(0.033)	(0.033)	(0.033)
Fin_Team	0.031***	0.028***	0.033***	0.033***
	(0.007)	(0.007)	(0.007)	(0.007)
as.factor(Year)	Yes	Yes	Yes	Yes
Fin_Rating	0.384***	0.373***	0.396***	0.435***
	(0.096)	(0.095)	(0.097)	(0.096)
Fin_Institution	0.129***	0.113***	0.133***	0.129***
	(0.030)	(0.030)	(0.030)	(0.030)

Fin_EthereumBlockDummy	−0.099	−0.050	−0.099	−0.094
	(0.159)	(0.158)	(0.160)	(0.160)
Fin_WhitelistKYCDummy	0.263**	0.251**	0.270**	0.254**
	(0.119)	(0.118)	(0.120)	(0.120)
Fin_MinInvestDummy	−0.081	−0.069	−0.086	−0.091
	(0.104)	(0.103)	(0.104)	(0.104)
Fin_NumCurrencies	0.037	0.037	0.039	0.038
	(0.026)	(0.026)	(0.026)	(0.026)
Fin_hardcapDummy	−0.035	−0.010	−0.048	−0.121
	(0.199)	(0.198)	(0.200)	(0.199)
Fin_softcapDummy	−0.085	−0.101	−0.076	−0.050
	(0.142)	(0.141)	(0.142)	(0.143)
Fin_BonusDummy	−0.062	−0.035	−0.067	−0.041
	(0.115)	(0.114)	(0.115)	(0.115)
pc1_all_wins:English1	0.091**			
	(0.039)			
PC1_S_wins:English1		0.172**		
		(0.077)		
PC1_D_wins:English1			0.100**	
			(0.044)	
lexical.density_wins:English1				−0.053
				(0.139)
Constant	15.425***	14.554***	15.319***	14.645***
	(1.722)	(1.694)	(1.729)	(1.721)
Observations	1,202	1,202	1,202	1,202
R^2	0.152	0.166	0.147	0.146
Adjusted R^2	0.135	0.149	0.130	0.129
Residual Std. Error (df = 1177)	1.677	1.664	1.682	1.683
F Statistic (df = 24; 1177)	8.811***	9.773***	8.454***	8.402***

** This table provides the results of hypothesis 3, which assumes that the curvilinear relationship between the level of lexical complexity found in the white paper and the amount raised during the coin offering period is stronger for ICOs from English-speaking countries. Model 9 takes lexical complexity as independent variable, while model 10, 11 and 12 take lexical sophistication, lexical diversity and lexical density as respective independent variables. The 4 models include interaction terms between English and the corresponding independent variable. ***, **, * represent the significance level of the resulting coefficient, at a 99%, 95% and 90% confidence level respectively. The significance test applied in this table is the two-sided t-test. This table has been produced with RStudio.*

UNIVERSITÉ CATHOLIQUE DE LOUVAIN
Louvain School of Management

Place des Doyens, 1 bte L2.01.01, 1348 Louvain-la-Neuve
Boulevard Emile Devreux 6, 6000 Charleroi, Belgique
Chaussée de Binche 151, 7000 Mons, Belgique

www.uclouvain.be/lsm