

Promotor Prof. dr. Bernard De Clerck
Vakgroep Vertalen, tolken en communicatie
Copromotor Prof. dr. Véronique Hoste
Vakgroep Vertalen, tolken en communicatie

Decaan Prof. dr. Marc Boone
Rector Prof. dr. Rik Van De Walle

Dear Stakeholder

Exploring the language of sustainability reporting:
a closer look at readability, sentiment and
perception.

Nils Smeuninx

Proefschrift voorgelegd tot het behalen van de graad van Doctor in de Taalkunde

Research funded by an SBO grant from VLAIO (Agency for Innovation and
Entrepreneurship)

2018

Abstract

Amidst financial crises, an increasingly threatened environment, and countless worker rights conflicts more visible than ever due to growing globalisation, it can come as no surprise that ‘sustainability’ has grown from an attractive buzzword to a core component of almost all areas of public life. Governments and business are making changes – at times even reinventing their identities – to maximise not just what they have in the present, but what will be left of it in the future. Environmental, social and organisational sustainability have joined profit or loss as core elements of a company’s scorecard – or, if not that, then at least of their rhetoric. In many respects, a company that seeks only to compete financially is no longer competitive.

Along with this change has come a shift in (what many perceive to be) companies’ fiduciary duties. Companies can no longer prioritise their shareholders above all others – they are also accountable to their stakeholders: employees, consumers, communities local to their operations, etc. The primary vector for that accountability – the annual report – has evolved along with that change. A considerable majority of companies now publish a sustainability report in addition to or together with their (financial) annual report to disclose the key aspects of their non-financial performance, just as the annual report does for the financial. While the genre originally invited greenwashing – i.e. a focus on favourable presentation rather than substantive action – it continues to evolve towards greater (self-)regulation and mandatory disclosure.

A key difference between the genres, however, lies in its audience. While the financial report is a specialised genre chiefly aimed at experts (analysts and investors), the sustainability report, as a primary vector for non-financial accountability, addresses a potentially much wider group of heterogeneous stakeholders. That makes the sustainability report a specialised genre addressing a far less specialised group than its financially-oriented sibling. The financial report’s reputation for difficult language and impression management may not deter its expert audience, but that same language translating to sustainability reporting is likely to impede many of the readers not part of the core shareholder audience. This study investigates to what extent the sustainability report adapts to a wider audience, how that reflects in its language, and builds on those

outcomes to address how sustainability reports can better accommodate the linguistic requirements of their wider audience. After a theoretical exploration of the concepts of sustainability (reporting) and readability, we pursue four primary avenues of research into the linguistic genre traits of sustainability reporting.

A first, broad-scope exploration of a 470-document, approximately 2.75-million token corpus of year-2012 sustainability reports and their accompanying letters – as well as letters from annual reports – finds that sustainability content is, if anything, less readable than financial content. It makes this observation not just based on the conventional formula-based approach to readability measurement, but also integrates Natural Language Processing as a more nuanced estimate of a text’s characteristics. Notably, this enquiry finds little, if any evidence of obfuscation – companies concealing unfavourable outcomes in complex language – in spite of considerable evidence of the phenomenon in financial reports according to previous research. It does, however, find a notable association between a report’s language variety and its readability and syntax. For instance, UK reports are significantly more passive than US ones, which might impact the efficacy of cross-varietal reporting.

Second, we investigated to what extent the readability of a report’s accompanying letter influenced readers’ perceptions of the company by creating a lightly and heavily simplified version of such a letter and submitting them to a panel. Those amongst the panellists unfamiliar with the genre reacted more positively to the original, most complex version of the letter than those with previous experience did, while neither of the simplified versions showed a difference between the groups. These results are likely indicative of the efficacy of sustainability reports’ language, in spite of its difficulty: it may well have a positive influence on how laypersons – which many potential readers amongst its stakeholders may be – perceive the company. However, neither group’s opinion of the text or company declined as readability went up, signalling that it may be safer for companies to attempt to simplify their disclosures than they might expect.

A third inquiry investigated the extent to which the ‘Pollyanna Effect’ – a bias towards positivity in (reporting) language - occurs in sustainability reporting. This inquiry relied on annotators following an annotation scheme designed to account for the genre’s potential tension between various areas of performance, e.g. environmental stewardship coming at the cost of short-term profit. Analogous with financial reports’ reputation of excessive positivity, we found that reports likely contained more positive information than an aim of balance would suggest. More significantly, this enquiry, contrary to the first, did show an association between the positivity of the outcomes reported on and use of company-oriented agency framing; in other words, companies appear to attribute positive outcomes to themselves more than they do negative ones.

A final component attempted to discern what influences perception of readability for those unfamiliar with the genre. In an experiment based on human perception of readability, we attempted to distil the chief (perceived) contributors to readability or a

lack thereof from scorers' comments. This informed the training of a genre-adapted machine learning system meant to improve upon readability prediction for sustainability reporting. Traditional formulae tend to consistently rate corporate reporting as very difficult, but machine learning allows much greater nuance than 'generic' readability measures do, both in the features that inform it and the resolution of scores it can assign. This proof of concept demonstrates the merit of Natural Language Processing in helping authors write more accessible reports that better meet their widening audience's needs.

In that respect, this dissertation attempts to contribute to corpus linguistics research through genre adaptation of a diverse range of linguistic techniques, including NLP techniques, with machine learning as the most notable. It contributes to business communication research by analysing the linguistically underexamined genre of sustainability reporting.

In summary, we explore a number of ways in which the linguistic properties of sustainability reports keep the genre from achieving its potential as a vector for accountability and engagement towards a broad, heterogeneous group of stakeholders. The final part of the study attempts to formulate avenues for improvement. Based on the aforementioned results, we see the greatest potential for improvement in companies using an active, narrative style that creates engagement through personal pronouns, and trust that simpler language will not damage their perceived credibility or professionalism. On a process level, we find that although readability measures can be a valuable aid or 'second opinion', an author's own judgment of readability is inevitably more complete than any heuristic might be. However, a genre-adapted learner can still vastly improve on traditional readability formulae. Readability initiatives that set formula-based targets are likely to only hamper the genre by encouraging authors to 'write to formulae'. The author and editor retain access to the best yardstick for readability – the human ability to process language holistically. Nevertheless, easy reading remains 'damned hard writing'.

Samenvatting

Toenemende globalisering maakt recente financiële crisissen, arbeidsrechtswaardes en de druk waaronder het milieu staat steeds zichtbaarder. Het hoeft dan ook niet te verbazen dat het concept 'duurzaamheid' is uitgegroeid van commercieel jargon naar een hoeksteen van vrijwel alle aspecten van het openbare leven. Regeringen en bedrijven voeren veranderingen door – en vinden zichzelf zelfs opnieuw uit – om ervoor te zorgen dat ze in de toekomst zullen blijven behouden wat ze nu hebben. Milieu- en sociaal bewustzijn zijn structurele elementen van een organisatie geworden en staan naast winst of verlies op de eindbalans – of vormen minstens een belangrijk deel van het bedrijfsimago. Een bedrijf dat enkel op winst uit is, is niet langer concurrentieel.

Hand in hand daarmee zijn ook de verantwoordelijkheden van bedrijven verschoven. Bedrijven moeten niet enkel de belangen van hun aandeelhouders behartigen, maar ook die van alle belanghebbenden, onder andere werknemers, consumenten en de gemeenschappen rond bedrijfssites. De belangrijkste manier waarop ze die verantwoordelijkheid kenbaar kunnen maken is via het bedrijfsrapport. Dat genre is samen geëvolueerd met de hierboven beschreven verschuivingen. Ofwel omvat het jaarrapport nu ook niet-financiële informatie, zoals milieu- of sociale aspecten, ofwel geeft het bedrijf die vrij in een apart duurzaamheidsrapport. Hoewel vroege duurzaamheidsrapporten vaak meer om presentatie en retoriek gingen dan om inhoud, worden ook resultaten qua duurzaamheidsbeleid steeds belangrijker. Er is namelijk een evolutie naar meer (zelf-) regelgeving, en het wordt steeds vaker verplicht dergelijke informatie te publiceren.

Er is echter een cruciaal verschil tussen financiële rapportering en duurzaamheidsrapportering, namelijk het publiek. War het financiële rapport vooral experts (zijnde aandeelhouders en analisten) aanspreekt, heeft het duurzaamheidsrapport een potentieel veel breder publiek, namelijk alle belanghebbenden bij wat het bedrijf doet. Het blijft echter een gespecialiseerd genre dat hen toeschrijft.

Bedrijfsrapporten hebben de reputatie moeilijke taal te gebruiken, en soms zelfs die taal te manipuleren om de best mogelijke indruk te maken. De impact daarvan op de

experten die financiële rapporten lezen is gering, maar de impact op dit bredere publiek zou erg groot kunnen zijn; die moeilijkheid is een potentieel obstakel voor die belanghebbenden die minder ervaring hebben met het genre. Dit onderzoek peilt dan ook naar de manieren waarop en de mate waarin duurzaamheidsrapportering aangepast is aan haar bredere publiek en hoe dat het taalgebruik beïnvloedt. Op basis van de resultaten bespreken we ook hoe het dat publiek efficiënter kan aanspreken. We verkent de concepten duurzaamheid(srapportering) en leesbaarheid eerst op een theoretisch niveau, en onderzoeken het taalgebruik van duurzaamheid daarna aan de hand van vier methodes.

De eerste aanpak is een overzichtsstudie van de 470 teksten, onderverdeeld in 2.75 miljoen *tokens* (i.e. woorden en leestekens) die we voor dit corpus hebben verzameld. Dat corpus bestaat uit Engelstalige duurzaamheidsrapporten van beursgenoteerde bedrijven en de overeenkomstige brieven aan aandeelhouders en belanghebbenden, zowel voor de duurzaamheidsrapporten en financiële rapporten. In de grootst mogelijke mate omvatten de teksten boekjaar 2012.

Deze piste leert ons dat duurzaamheidsrapportering geenszins leesbaarder is dan financiële rapportering, en mogelijk zelfs minder leesbaar. Dat ontdekken we niet enkel via de ‘traditionele’ leesbaarheidsformules, maar ook door fijnmaziger de lexicosyntactische aspecten van de tekst te meten via computationele linguïstiek. Ondanks vorige studies die dat wel vaststelden voor zowel financiële als duurzaamheidsrapporten vinden we erg weinig indicaties op tekstniveau dat bedrijven zwakkere resultaten proberen te verhullen in moeilijker taalgebruik. We stellen echter wel vast dat de variëteit van het Engels waarin het rapport is geschreven een aanzienlijke impact heeft op leesbaarheid en taalgebruik. Zo bevatten Britse rapporten bijvoorbeeld minder passiefstructuren dan Amerikaanse, wat het goede begrip tussen de twee regio’s zou kunnen bemoeilijken.

Als tweede piste hebben we onderzocht in welke mate de leesbaarheid van de begeleidende brief invloed heeft op de perceptie van de lezer rond het rapport en het bedrijf. We hebben een originele brief en zowel een licht als zwaar vereenvoudigde versie voorgelegd aan respondenten die wel of geen ervaring hadden met bedrijfsrapportering. Elke respondent kreeg slechts een enkele versie te zien. We stelden vast dat respondenten die het genre niet kenden positiever reageerden op de moeilijkste tekst (het origineel), maar er geen verschil in perceptie was tussen beide groepen wanneer die de vereenvoudigde versies lazen. Met andere woorden, ondanks de negatieve reputatie van bedrijfsrapportering qua moeilijk taalgebruik schijnt die moeilijkheid wel een wenselijk effect te hebben op de reacties van leken. Aangezien de vereenvoudiging de mening van respondenten met meer ervaring niet aantastte moeten bedrijven echter ook niet vrezen dat eenvoudiger taalgebruik hen geloofwaardigheid zou kosten.

De derde piste heeft onderzocht in welke mate het ‘*Pollyanna Effect*’ – een neiging naar positief taalgebruik die sterk aanwezig is in financiële rapporten – terug te vinden is in

duurzaamheidsrapporten. Voor deze deelstudie hebben we een nieuw annotatieschema voor sentiment opgesteld en het voorgelegd aan taalexperten. Dit schema had het doel om de spanning tussen verschillende thema's in duurzaamheidsrapporten, zoals financiële en milieugerelateerde resultaten, op te vangen. We stelden vast dat ook duurzaamheidsrapporten meer positieve informatie bevatten dan een 'gebalanceerde' rapporteringsstijl zou insinueren. Opvallender echter was de link tussen positieve informatie en hoe het bedrijf zich als agens positioneert. Hoe beder de resultaten beschreven in de zin, zo blijkt, hoe sneller bedrijven geneigd zijn de verantwoordelijkheid aan zichzelf toe te eigenen (bijvoorbeeld door gebruik van de eerste persoon).

Een laatste onderzoekspiste probeerde vast te stellen welke elementen in de taal een invloed hebben op hoe leesbaar een tekst is volgens menselijk oordeel, en in welke mate we dat oordeel kunnen benaderen via een lerend computersysteem. Dat levert een waardevolle toevoeging aan leesbaarheidsonderzoek op, omdat de traditionele aanpakken, zoals leesbaarheidsformules, relatief beperkt zijn in zowel de aspecten die ze in rekening kunnen nemen als de nuance waarmee ze een oordeel kunnen uitdrukken. Dit toont de haalbaarheid – en het nut – aan van computationele linguïstiek als hulpmiddel voor zakelijke auteurs die beter tegemoet willen komen aan de leesbaarheidsvereisten van hun steeds bredere publiek.

Samenvattend kunnen we stellen dat deze studie bijdraagt aan corpuslinguïstiek door genreadaptatie van een breed scala aan technieken, waaronder een aantal uit de computationele linguïstiek, en bij uitstek lerende analysesystemen. Ze draagt bij aan onderzoek naar zakelijke communicatie door een linguïstisch onderbelicht genre te onderzoeken.

Concreter heeft dit onderzoek een aantal manieren verkend waarop het taalgebruik in duurzaamheidsrapportering het genre ervan weerhoudt een optimaal geschikt communicatiemiddel naar een breed scala aan heterogene belanghebbenden te zijn. Het laatste deel van dit proefschrift stelt dan ook een aantal manieren voor om dat taalgebruik te verbeteren. Op basis van de resultaten schijnen bedrijven er potentieel erg bij te baten om een actieve, verhalende stijl te gebruiken; persoonlijke voornaamwoorden blijken ook geschikt om de lezer geboeid te houden. Uit het corpus blijkt ook dat makkelijker taalgebruik de geloofwaardigheid van het bedrijf, noch de perceptie van meer ervaren lezers schijnt te schenden.

Qua schrijfproces stellen we vast dat hoewel leesbaarheidstools een belangrijk tweede perspectief kunnen bieden, de auteur nog steeds in eerste instantie moet afgaan op eigen inzicht in de taal. Qua taalgevoel blijft het menselijke inschattingsvermogen veel completer dan dat van software. Leesbaarheidsdoelwitten op basis van formules komen dat inschattingsvermogen waarschijnlijk niet ten goede, hoewel een lerend systeem al een hele verbetering kan bieden ten opzichte van die formules. Ongeacht de beschikbare hulpmiddelen blijft makkelijk leesbare tekst echter uitdagend om te schrijven.

Acknowledgements

With very few exceptions, I would like to thank everyone.

Genre conventions, however, dictate slightly more specificity than that. Without so much as the pretense of exhaustivity, the following people merit specific mention.

I can say with almost metaphysical certainty that you would not be reading these words were it not for the assistance, support and encouragement of my supervisor, Bernard De Clerck. To borrow an observation from others who have worked with him, he has been, throughout this process, essentially impossible to fault. I am, of course, also grateful to the co-supervisors for this project, Véronique Hoste and Walter Aerts, for their assistance in concept, execution and logistics throughout, especially when things proceeded less smoothly than they might have. Thank you for your patience.

I would also like to express my gratitude for the time and insights of this dissertation's jury, which includes Lies Bouten, Andrew Kehoe, Geert Jacobs, and Orphée De Clercq. I hope they will appreciate the irony of a dissertation on readability fully embracing the academic register, as I fear there is more than one instance where brevity won out over reading ease.

On a similar note, I would like to thank all the colleagues, students and annotators (which sometimes blur the line a little) that I've had the pleasure of working with at VTC. That crucially includes Koen and Ludovic for statistical support, and Frederic and Kara for their heroic annotation efforts. A specific word of thanks is, of course, also due to the past and current members of LT3, including (in alphabetical order) Arda, Ayla, Bart, Bram, Camiel, Claudia, Cynthia, Els, Gilles, Joke, Julie, Klaar, Laura, Lieve, Luna, Mariya, Marjan, Orphée, Peter, Sarah, Stef and Véronique (it turns out there's quite a lot of us).

Out of the above, a first special mention has to go to Peter for helping us cope with the psychological effects of Fridays. A second goes to Cynthia for lending – over the course of these five years – more ears than seems anatomically plausible. I hope I've been able to return the favour a little. Finally, an astronomical thank-you to Orphée is certainly in order – for helping me get to grips with readability as a concept, extracting all the necessarily metrics from the corpus, and for retraining her readability assessment system

to accommodate everything we'd learned about the genre. I quite literally could not have done this without her.

The same goes for Oveis Madadian, who extracted the performance data informing many aspects of this study from Datastream.

Finally, a word of thanks to VLAIO (the Flemish Agency for Innovation and Entrepreneurship) for funding the lion's share of this research (as well as UGent for a seven-month extension). Without them, I would have had to go hungry. The responsibility for managing to put on quite a bit of weight, however, is entirely my own.

In light of all of the help I've received throughout the project from those mentioned above, I hope you, the reader (and thank you for being just that!) will forgive me for using the academic 'we' throughout the dissertation. It certainly felt more intellectually honest.

Shifting now to the entirely personal, a word of thanks ought to go out to my parents, Annita and Herman, and grandparents Susanna, Gustave, Johanna and Freddy. Although this is neither a dissertation in genetics nor causality, I am fairly confident they played a key role in these words making it onto paper.

For keeping me sane, or at least closer to a baseline level of insanity, I am grateful to everyone who invited or joined me for a game (cardboard, digital or imaginary) throughout the course of the writing process, with a special mention to Sofie for... well, the better part of everything.

On a similar note I owe an acknowledgement to Jorien for everything we shared throughout the better part of this process, for oodles of patience and, as ever, for keeping the couch occupied.

I would also like to thank cats – all of them in general, and my own in particular.

Last but certainly also most, I am incomprehensibly grateful to Tine for the love and support throughout the last stretch, perhaps best exemplified by her willingness to proofread all 100.000-odd words that make up this dissertation within a span of 48 hours before submission. You are insane, and I love you (for it).

And, because it bears repeating: you, the reader. Thank you for reading this.

I hope that at least begins to cover everyone. It's been a hell of a ride. Thanks for bearing with me.

List of Tables

Table 1	Examples of difficult sentences in <i>The Cat in the Hat</i> and <i>Ulysses</i>	28
Table 2	Examples of easy and difficult sentences according to formulae.	32
Table 3	‘Bands’ of readability as distinguished by the Flesch Reading Ease Index.	34
Table 4	Overview of reading grade levels.	36
Table 5	Corpus composition by genre, region and industry.	81
Table 6	Corpus composition by unique company and genre.	82
Table 7	Mean, SD, minimum and maximum for different genres expressed in tokens before and after cleaning, and page numbers before cleaning.	83
Table 8	Illustration of text extraction issues for non-running text.	89
Table 9	Per-genre summary of minima, maxima, means and standard deviations (SD) for dependent variables.	97
Table 10	Summary of significance and effect size of genre as predictor, per dependent variable.	102
Table 11	Per-genre summary of predictors’ significance and effect size for general linear models predicting readability metrics.	107
Table 12	Means of dependent variables (FRE, Lexical Density, Subordination and Passives) per region, and significance of difference between those means for sustainability reports.	112
Table 13	Means and standard deviations for Flesch-Kincaid Grade Level and Gunning Fog Score for financial LtSs, divided by region and industry. ...	114
Table 14	Means of dependent variables (Lexical Density, Parse Tree Depth, Subordination and Passives) per region, and significance of difference between those means for financial LtSs.	115
Table 15	Means of dependent variables (FRE, FKGL, GF, Lexical Density and Passives) per region, and significance of difference between those means for sustainability-related LtSs.	116
Table 16	Summary of significance of differences within respondents’ answers, split along text and/or expertise.	141
Table 17	Means of significant differences between responses.	143
Table 18	Inter-annotator agreement scores (Cohen’s Kappa) for text-level sentiment annotations.	161
Table 19	Inter-annotator agreement scores (Cohen’s Kappa) for text-level sentiment annotations.	162

Table 20	Sum total of all negative and positive sentiment in all sentences in the subcorpus.	165
Table 21	Tallies of sentiment scores assigned to all sentences in the subcorpus.	165
Table 22	Means and standard deviations for performance scores for companies present in the subcorpus.	167
Table 23	Tallies for rhetorical move types for all annotated sentences.	168
Table 24	Tallies for agency types for all annotated sentences.	170
Table 25	Tallies for time frame types for all annotated sentences.	172
Table 26	Tallies for subjectivity types for all annotated sentences.	174
Table 27	Summary of ordinal regression predicting agency patterning.	180
Table 28	Categories, criteria and examples of scorer comments.	190
Table 29	Tallies of comments about improved, lowered or ambivalently affected reading ease based on aspects of language and content.	192
Table 30	Summary of significances and effect sizes for type of comment on a given category in a multiple regression model predicting assigned score.	215
Table 31	Mean differences between comment types per category.	216
Table 32	RMSE scores for different machine learning training scenarios.	224
Table 33	Correlations between readability predictors and normalised readability score.	227
Table 34	Correlations between genre-specific readability predictors introduced based on comments and previously used readability measures.	228

List of Figures

Figure 1	Visualisation of a sentence as parsed by the Stanford CoreNLP group (2016) online parser.....	46
Figure 2	Example of the WebAnno annotation interface.....	156
Figure 3	Stacked column chart representing positive, negative and neutral outcomes per performance aspect.....	165
Figure 4	Proportions of rhetorical move types.....	168
Figure 5	Proportions of agency framing types.	170
Figure 6	Proportions of time framing types.....	172
Figure 7	Proportions of subjectivity types.	174
Figure 8	Screen capture of the scoring interface.	187
Figure 9	Stacked column chart representing comments indicating greater reading ease, lower reading ease and ambivalent effects.	193
Figure 10	Relationship between number of positive comments minus negative comments and assigned score.	196
Figure 11	Boxplot of human-assigned scores before normalisation	214
Figure 12	Scatterplot and regression line for normalised averaged score and Flesch score.....	218

Table of Contents

Abstract	v
Samenvatting	ix
<i>Acknowledgements</i>	<i>xiii</i>
<i>List of Tables</i>	<i>xv</i>
<i>List of Figures.....</i>	<i>xvii</i>
<i>Table of Contents.....</i>	<i>xix</i>
Introduction: On Sustainability Reporting and its Stakeholders	1
<i>Part 1: Theoretical Framework.....</i>	<i>5</i>
Chapter 1 Corporate Sustainability Reporting	7
1.1 A Shift towards Sustainability	7
1.2 Legitimacy and Licence to Operate	9
Legitimacy as a Social Resource: the Marikana Miners' Strike.....	11
1.3 Four Pillars of Sustainability	12
1.3.1 Financial Sustainability	12
1.3.2 Environmental Sustainability.....	13
1.3.3 Social Sustainability	14
1.3.4 Sustainable Governance	15
1.3.5 Multiperspective Performance.....	16
1.4 Corporate (Sustainability) Reporting	16
1.4.1 Why Report?.....	16
1.4.2 The Letter to Shareholders	17
1.4.3 The Financial (Annual) Report	18
1.4.4 The Sustainability Report.....	19
1.4.5 Standardisation Initiatives.....	19
1.4.6 The Stakeholder.....	20

1.4.7	Linguistic Inquiries into Corporate Reporting	22
Chapter 2	Readability	27
2.1	What is Readability?	27
2.2	Why Readability Matters	30
2.3	Readability Formulae	31
2.3.1	The Flesch Reading Ease Index.....	32
2.3.2	Flesch-Kincaid Grade Level.....	34
2.3.3	Gunning Fog Index	36
2.3.4	A Caution	37
2.4	Aspects of Readability	38
2.4.1	Readability versus Understandability.....	40
2.4.2	Text-internal Criteria	41
2.4.2.1	Word length, Sentence Length, and Word Rarity	41
2.4.2.2	Aspects of lexicosyntactic complexity	42
2.4.3	The Audience	48
2.4.3.1	Measuring Understandability: the Cloze Procedure	49
2.4.3.2	Subject Familiarity	50
2.4.3.3	Language Proficiency.....	50
2.4.3.4	Motivation	51
2.4.3.5	Difficulties in Quantifying.....	51
2.4.4	The Author(s).....	52
2.4.4.1	Language proficiency.....	53
2.4.4.2	Authorial (corporate) voice and language variety	54
2.4.5	Paratext.....	55
2.5	How machines process text and what that means for readability studies	56
2.6	Readability beyond formulae.....	57
2.6.1	Human Review and Readability Heuristics	58
2.6.2	Natural Language Processing	60
2.6.2.1	An Example: Word-Sense Disambiguation	61
2.6.2.2	NLP as a Tool	62
2.6.3	Scoring and Machine Learning	62
2.7	Moving Forward.....	65
Part 2: The Language of Corporate Reporting		67
Chapter 3	The Readability of Corporate Reporting	69
3.1	Motivation	69
3.2	Context and hypotheses	71
3.2.1	Genre and Audience	71
3.2.2	Language Variety	72
3.2.3	Company and Legitimacy	73
3.2.4	Impression Management: Obfuscation and Defensive Attribution	74
3.3	Methodology	77
3.3.1	Corpus Collection	78

3.3.2	PDF and Text Processing	83
3.3.3	Running Text Extraction	87
3.3.4	Corpus Processing	90
3.3.5	Statistical Analysis	91
3.3.6	Limitations.....	92
3.4	Analysis and Discussion	95
3.4.1	Exploring the Sample.....	96
3.4.1.1	Readability Formulae	98
3.4.1.2	Lexicosyntactic Features	99
3.4.1.3	Comparison between Genres	101
3.4.2	Fine-grained Analysis of Company Characteristics.....	105
3.4.2.1	Language Variety.....	111
3.4.2.2	Corporate Performance	118
3.5	Conclusions.....	120
Chapter 4	Readability Manipulation.....	123
4.1	Motivation	123
4.2	Hypotheses	125
4.3	Editing Process	126
4.3.1	More Readable – ‘Difficult’ (30-50 FRE score).....	128
4.3.2	Most Readable – Towards ‘Plain English’ (60-70 FRE score)	130
4.3.3	Risks of Writing to Formulae	132
4.4	Assembling the Questionnaire.....	133
4.4.1	Introduction and Informed Consent	134
4.4.2	Reading and Comprehension Test	135
4.4.3	Company and Composition	135
4.4.4	Biographical Data	138
4.4.5	Survey Process	138
4.4.6	Description	139
4.5	Analysis	140
4.5.1	Differences between Texts	143
4.5.2	Differences between Levels of Expertise	144
4.6	Limitations.....	145
4.7	Discussion and Conclusion	147
Chapter 5	Sentiment Analysis	149
5.1	Motivation	149
5.2	Hypotheses	153
5.3	Automatic Lexicon-based Sentiment Analysis	155
5.4	Annotating Sentiment in a Sustainability Context.....	156
5.5	Inter-annotator Agreement	161
5.5.1	Text-level Annotations	161
5.5.2	Sentence-level Annotations.....	162
5.6	Processing.....	162
5.7	Frequencies and Description.....	164
5.7.1	Sentiment scores	165

5.7.2	Rhetorical Moves	168
5.7.3	Agency.....	170
5.7.4	Time Frame	172
5.7.5	Subjectivity.....	173
5.8	Analysis	177
5.8.1	Sentiment and Performance.....	177
5.8.2	Industry and Region.....	178
5.8.3	Agency.....	179
5.9	Discussion & Conclusions	180
Chapter 6	Human Assessment and Machine Learning	183
6.1	Motivation	183
6.2	Scoring.....	184
6.2.1	Implementation.....	184
6.2.2	Participant Selection.....	185
6.2.3	Subcorpus selection	185
6.2.4	Scoring and Annotation Process	188
6.3	Outcomes	191
6.3.1	Initial Exploration	194
6.3.1.1	Lower reading ease	197
6.3.1.2	Increased reading ease	204
6.3.2	Implications.....	212
6.4	Exploration	214
6.5	Processing.....	216
	Comparison to Flesch Score	217
6.6	Machine Learning.....	219
6.6.1	Measures Used	219
6.6.2	Language Modelling.....	222
6.6.3	Prediction	223
6.6.4	Correlations & Discussion	225
Part 3: Towards Better Reporting: Discussion and Conclusions		233
Chapter 7	Discussion.....	235
Chapter 8	Implications	241
Chapter 9	Future Research.....	249
Conclusion	On Writing Corporate Reports	253
Bibliography	257
Appendix	269
	Appendix 1: Corpus Description & Additional Analyses.....	271
	Companies in Corpus	272

CSR Reporting Requirements per Country (year 2012)	278
Post Hoc Analyses for Full Corpus.....	280
Sustainability Reports – Readability Formulae	280
Sustainability Reports – Lexicosyntactic Features	287
Financial (Annual Report) Letters to Shareholders.....	295
Financial (Annual Report) Letters to Shareholders – Lexicosyntactic Features.....	301
Sustainability Report Letters to Stakeholders - Readability Formulae	310
Sustainability Report Letters to Stakeholders - Lexicosyntactic Features	317
Appendix 2: Manipulation	327
Readability Survey (adapted for offline display)	328
LtS - Original Version: FRE 13.8	334
LtS - ‘More Readable’ Version: FRE 36.6	335
LtS - ‘Most Readable’ Version: FRE 47.1.....	336
Appendix 3: Sentiment Annotation	337
Sentiment Annotation Guidelines	339
Sentiment Annotation Guide – How to Identify Different Aspects of Performance	349
Sentiment Annotation Guide – Reference Flowchart	350
Appendix 4: Human Assessment & Machine Learning	353
PVH – Excerpt from Sustainability Report	355
Freeport-McMoRan – Excerpt from Sustainability Report	356
Chevron – Excerpt from Sustainability Report	358

Introduction:

On Sustainability Reporting and its Stakeholders

Throughout the last decades of the 20th and first decades of the 21st century, corporate reporting – especially annual reporting – has diversified enormously in the scope of topics it covers (KPMG 2013, 2017; Lodhia & Hess 2014). From a focus on the financial bottom line (De Villiers, Rinaldi & Unerman 2014), the corporate annual report has evolved through a ‘Triple Bottom Line’ approach (Elkington 1998, 2004) that includes greater attention to the company’s environmental and social impacts. This, in turn, has transitioned towards a multi-faceted form of ‘Sustainability Reporting’ (e.g. Unerman, Bebbington & O’Dwyer 2007) and/or ‘Integrated Reporting’ (International Integrated Reporting Council 201; De Villiers, Rinaldi & Unerman 2014). The latter attempt to holistically capture all aspects, financial or otherwise, that are material not just to the company’s performance over the preceding and following fiscal years, but to its longer-term sustainability.

With the broadening of scope out from corporate (annual) reporting that delivers financial information to shareholders came the inception of a sister genre informing the company’s stakeholders in the wider sense – its employees, partners, local communities, etc. – of non-financial performance information pertinent to that wider group’s interest. As this genre increased in prominence and “often [...] in volume and complexity” such disclosures became known as ‘environmental reports’ or ‘social reports’ (De Villiers, Rinaldi & Unerman 2014) as they emphasise the company’s impact on their local and wider natural environment or social aspects such as employee health and safety as well as communication with local communities. A third prominent pillar of non-financial sustainability, in addition to social and environmental sustainability, is sustainable governance. The latter concerns itself with, for instance, management conflicts of interest. While the documents that disclose these non-financial areas of performance have had and continue to have many names and forms, one very common umbrella term is ‘(corporate) sustainability reporting’.

Financial and non-financial (reporting) content often exhibit a remarkable dynamic of similarities and differences that their similarly intersecting but divergent audiences accentuate. For one, financial disclosures almost invariably face more and stricter regulation than non-financial disclosures do, to the point of the former frequently being mandatory and the latter voluntary¹ (Aerts, Cormier, Magnan 2008, KPMG 2013). In terms of audiences, the company's shareholders, to whom the financial disclosures are most relevant, are only a fraction of the company's stakeholders, i.e. those groups whom the company's activities tend to affect directly or indirectly. As illustrated above, these can range from employees to local communities, but can also include governments, NGOs, concerned citizens, etc. This much wider group makes up significant portions of the audience for corporate disclosures that go beyond the strictly financial, as Townsend, Bartels & Renaud (2010) indicate.

The most frequent commonality between financial and non-financial corporate disclosures, however, is perhaps their similarity in presentation. Financial disclosures have a reputation as impenetrable documents – on average, deservedly so (see e.g. Curtis 1995 & 1998; Stanton & Stanton 2002; Li 2008) – and sustainability reporting tends to be very similar in form. However, financial disclosures generally address an expert audience whereas sustainability disclosures' wider stakeholder audience have a far more heterogeneous audience (Townsend, Bartels & Renaut 2010; De Villiers, Rinaldi & Unerman 2014).

While the linguistic properties of financial reporting have received considerable scientific attention, and sustainability reporting is frequently the subject of scholarly inquiries, fairly few of the latter discuss its linguistic properties. This study aims to address that gap by examining, *inter alia*:

- To what extent financial reporting's often complex language transfers to sustainability reporting, and how such a transfer might affect the genre's utility to a wider stakeholder audience (Chapter 3);
- How the readability of corporate reports and the performance underlying them interact and, as a corollary, to what extent obfuscation, as attested for financial reporting, occurs in sustainability reporting (Chapter 3);
- How changes towards optimised general-audience readability affect audience perception of non-financial disclosures, e.g. in terms of perceived credibility, professionalism and competence of the reporting company (Chapter 4);
- How positively or negatively charged non-financial disclosures can be, and to what extent sentiment surrounding a certain aspect correlates with performance relevant to that aspect (Chapter 5);

¹ This was the case for financial 2012, which this study's corpus drew its text from; sustainability reporting has continued to evolve towards more mandatory regulation.

- To what extent we can automate the readability estimation of this genre in order to enable authors to write better (i.e. more accessible) reports and enable audiences to be better (i.e. more informed) readers (Chapter 6); and
- How authors of sustainability reporting might optimise sustainability disclosures in order to best cater to its wider audience (Chapter 6 & Chapter 8).

The study will not only attempt to innovate by addressing a knowledge gap in current research, but also aims to advance the best practices in measuring the linguistic properties of corporate reporting through a number of pilot studies, i.e. into manipulation, sentiment and automating readability prediction. Whereas most studies measure these reports' readability using the so-called 'readability formulae' such as the Flesch Reading Ease Score (Flesch 1948), we start from these relatively shallow (see section 2.3) formula-based measures but expand on them through the following additions:

- Deeper-level linguistic features measured through Natural Language Processing (NLP) (see sections 2.4.2.2 and 2.6.2);
- Panel-based manual readability scoring (see section 6.2.4);
- Prototype machine-learning-based automatic readability scoring (see section 6.6);
- Questionnaire-based perception study into readability manipulation (see section 4.4);
- Manually annotated sentiment information (see section 5.4).

This thesis will first proceed by describing the corporate annual report and sustainability disclosures' place within the wider genre of corporate reporting (Chapter 1). It then contextualises the concept of readability and sets out a number of practice-oriented approaches to quantifying it (Chapter 2), formulating a number of hypotheses on the readability and other linguistic properties of annual (sustainability) reporting (Chapter 3). We then describe the data collection process, with an emphasis on the linguistic information and the challenges of compiling an NLP-suitable corpus from visually rich material. Next, we describe and characterise the data through a broad-scope full-text analysis of the whole corpus, which attempts to capture not just its readability in terms of formulae, but also its syntactic and lexical complexity (Chapter 3).

We proceed by describing our efforts to measure the effect of shifts in readability on readers' perception of company performance, competence and professionalism, as well as their trust in the company (Chapter 4). We subsequently describe efforts to develop a sentence-level sentiment annotation scheme capable of capturing multi-faceted company performance that incorporates both financial and non-financial pillars, and present the outcomes of a pilot study deploying this annotation scheme (Chapter 5). As a last aspect of the study, we set out the methods and analysis of an assessment-driven manual scoring experiment aimed at optimising a learning readability prediction system for the genre of corporate reporting, but also offer a qualitative overview of non-experts'

assessment of these reports' readability (Chapter 6). We then proceed from a discussion of results (Chapter 7) to an overview of practical implications for the authors and users of sustainability reports (Chapter 8). Finally, we give an overview of potential avenues for future research (Chapter 9).

Part 1: Theoretical Framework

Chapter 1

Corporate Sustainability Reporting

1.1 A Shift towards Sustainability

Over the past decades, sustainability has grown from an ancillary component of corporate operations to one of its cornerstones. It is now a requirement that must be satisfied for a corporation to remain viable. To illustrate: KPMG's 2013 Corporate Responsibility Reporting Survey found that 93% of the world's largest 250 companies issued sustainability-related disclosures, and over half of companies worldwide that issued one included such information in their annual reports; they assert that the debate whether or not to report is over. This shift – both in attitudes and practice – was by no means limited to the corporate sphere. It has occurred, and continues to occur, throughout a larger context of increasing societal and governmental attention to sustainable development (KPMG 2013, Costa & Menichini 2013). Among other factors, recent financial crises and scarcity concerns, global warming and biodiversity issues, as well as social equality and worker rights issues are highly likely to have contributed to an increased awareness of (and demand for) sustainability in virtually all aspects of life. Large businesses especially have become a prime target for ever-growing scrutiny regarding their operations' sustainability (KPMG 2013). One example of this shift towards greater scrutiny is the Dodd-Frank Wall Street Reform and Consumer Protection Act, which aims to “improv[e] accountability and transparency in the financial system” (2010) in the wake of 2008's financial crisis; the various carbon reduction initiatives in the wake of the 1997 Kyoto Protocol exemplify the same phenomenon.

The preceding examples illustrate how the corporate sphere plays a crucial role in meeting sustainability targets and implementing initiatives; that is why regulators often mandate that cooperation through incentives or legal action (i.e. the metaphorical carrot and stick). Because companies are inevitably – directly or indirectly – involved in the sustainable development of the societies in which they operate, many also present it as

an integral part of their day-to-day operations. For instance, nine different companies present in this study's corpus (which Chapter 3 describes in greater detail) draw on a 'corporate DNA' metaphor within one or more sustainability-related documents in order to describe how crucial sustainable development and related concepts (such as Corporate Social Responsibility) are to their business strategy and identity. However, as section 1.2 will illustrate, a company's assertion or perception of sustainable business being core to their operations can be a fundamentally one-sided thing; it by no means obliges parties external to the company to agree.

Nevertheless, throughout their sustainability-related communications, many companies rhetorically position sustainability and non-financial performance as equally important to the company as its profitability. While a minority of companies may explicitly draw that equivalence, many imply it through aphorisms such as 'sustainable business is good business' (e.g. PVH 2013's "CSR is not only the right thing to do, but also means good business"). Regardless of the motivations behind them, many companies also engage in voluntary sustainability initiatives not mandated by any law or regulatory requirement (Berliner & Prakash 2015). Issuing non-financial disclosures such as a sustainability report has become the prime example of such voluntary but necessary practices to the point of the company placing itself at a competitive disadvantage by not issuing one (KPMG 2013). As the assertion that 'sustainable business is good business' and variations thereupon imply, engaging in initiatives to boost non-financial performance can also benefit financial performance; the two are not zero-sum.

In terms of balancing a company's financial and non-financial performance, Elkington (1998) captured the imagination with the conception of a 'triple bottom line', which sees companies evaluate themselves on not only the profit or loss they made, but also their environmental and social impact. We can recognise the relevance of these non-financial performance aspects in, for instance, Thomson Reuters' ASSET4 database (see e.g. Thomson Reuters 2012). ASSET4 measures a company's performance based on four primary aggregate measures: the aforementioned financial, environmental and social performance, and the company's performance regarding governance, i.e. the extent to which management prioritises the share- and stakeholders' interest, rather than their own; this measure concerns itself with, for example, corporate transparency or corruption issues.

Perhaps one of the most foundational documents for the governmental, societal and corporate conception of sustainability was *Our Common Future*. The United Nations World Commission on Environment and Development's 1987 publication on sustainable development, also called 'The Brundtland Report', defines sustainable development as:

Development that meets the needs of the present without compromising the ability of future generations to meet their own needs.

This carried forward into many organisations' thinking about sustainable operations and development, be it corporate or otherwise.

As the approach set forth by, *inter alia*, Thomson Reuters' ASSET4 (see section 1.3) encompasses the three 'triple bottom line' performance aspects (financial, social and environmental) and supplements them with a fourth (governance), we will explore all four, focusing on their relevance to sustainability and sustainability reporting. We will preface that exploration with an overview of two concepts that connect these performance measures – more specifically 'legitimacy' and 'licence to operate'.

1.2 Legitimacy and Licence to Operate

As the previous section anticipated, while the common perception of environmental or social sustainability may be that they run counter to the company's financial interests, at least from the corporate perspective, they often – necessarily – align in the long term. An overwhelming majority of companies function poorly or not at all without the assent of stakeholders – ranging from local communities to governments – to their activities. Companies require licence, be it legal, moral, or both, from parties with an interest in the figurative and literal ecosystems in which they operate, and generally owe the parties that grant them this licence a certain measure of transparency and accountability in return.

In the case of the extractive industries, for instance, these forces can be especially powerful. Mining, oil production and other activities can have a substantial impact on the aforementioned ecosystems across all four pillars of sustainability, especially the environmental and social. Extracting resources almost invariably has an impact on the local environment. While this impact is not always a lasting or permanent one, it often can be. As local communities, various civil groups, and local or national governments generally have an interest in maintaining the state of the environment, they must ensure that the benefits of these operations outweigh the costs and risks. Transparency on the company's behalf then entails making both positive and negative outcomes of their operations visible and comprehensible to those various stakeholders (see Bouten 2011), and accountability entails – voluntarily or otherwise – minimising their operations' negative impact where possible and compensating through added value where minimising that impact is not possible. Such compensation need not be financial; it can occur within many of the various spheres and ecosystems in which the company interacts with these stakeholders.

For instance, the social ecosystem around the company is another area in which companies' operations can have a significant impact, both positive and negative, that the

company may wish to control to the greatest possible extent. As a significant portion of the extractive industries' workforce requirements entails manual labour, and many such companies operate in developing nations, the employee's position can be an especially vulnerable one. While an increase in available jobs can benefit local communities socio-economically, the type of work often poses a greater-than-usual risk of injuries or health issues. When compounded by potentially less established worker rights and safety standards, this same employment can also pose a risk to local communities that companies, to be able to continue operating, will likely wish to mitigate or compensate through other means. These can include community schooling, healthcare or other support programs.

Ensuring a positive balance between costs and benefits to stakeholders in its operations, for instance through such initiatives, is one of the primary means through which a company can secure and maintain what (*inter alia*) Deegan, Rankin & Tobin (2002) call 'licence to operate', i.e. local and wider-scale communities' continued assent to the company's activities. Many companies in extractive sectors indicate the importance of securing and maintaining this 'licence to operate'. If a company persistently flouts worker safety guidelines as a cost-cutting measure, they may reap short-term financial benefits, but are likely to threaten their licence to operate and, thereby, their operations' financial – as well as social – sustainability. As the above examples may illustrate, companies receive licence to operate when stakeholders deem their actions legitimate, i.e. “desirable, proper or appropriate within some socially constructed system of norms, values, beliefs and definitions” (Suchman 1995, p. 574).

This definition highlights how legitimacy depends on the system assessing it, and is thus subjective: what one group of stakeholders might consider legitimate corporate behaviour when assessing the costs and benefits of a company's operations, others might consider wholly illegitimate.

For instance, the difference in incidence rates of *karoshi*, or death by overwork, between Japan and Western labour markets (Nishiyama & Johnson 1997) may well indicate that working conditions and demands on employees conducive to this phenomenon are less acceptable in Western labour markets. That is, Western labour markets may be more likely to recognise workplace environments that would lead to *karoshi* as illegitimate, and thus prevent the root causes from occurring. Nishiyama & Johnson (1997, p. 630) indicate an “unpaid ‘voluntary’ work culture” and long working hours as potential causes, and see an increase of working hours in Japan relative to a decrease in Western labour markers since World War II. Increasing awareness of and initiatives, to counteract the issue (Yamaguchi 2016) also illustrate how perceptions of legitimacy can shift.

Legitimacy as a Social Resource: the Marikana Miners' Strike

Its necessity to corporate operations marks legitimacy as an “operational resource” (Suchman 1995, p. 576), which Tilling (2004, p. 4) cautions may have “particularly dire consequences for an organisation” when exhausted, “which could ultimately lead to the forfeiture of their right to operate”. One such dire consequence of strained legitimacy within the mining industry was the 2012 Marikana Miners' Strike (Flak 2012), which saw 49 deaths resulting from a clash between miners and South African police at a Lonmin-operated mine. The Bench Marks Foundation indicates a contributing factor in the perceived lack of legitimacy of Lonmin's operations. It partially attributes that perception to remuneration issues, (Van Wyk 2012), but also indicates that earlier violence in 2011 stemmed from “unacceptable” (p. 72) worker health and safety alongside “appalling” (p. 73) living conditions, “undermining and devaluation of property”, and a general lack of “the very important and priceless ‘Social Licence to Operate’” (p. 81). Van Wyk (p. 81) further asserts that while the company signals improvements in its CSR reporting, actual conditions have remained unchanged.

Viewed in light of the fatal shootings at Marikana, these comments illustrate a negative cycle of legitimacy loss. According to the Bench Mark Foundation, the company claimed to strive for transparency without managing to achieve it, nor did it respect its economic, environmental and social accountability. Because they perceived Lonmin's operations as illegitimate, local workers removed its licence to operate through the means most immediately available to them, i.e. a strike. As this strike escalated, the legitimacy of Lonmin's operations suffered further damage due to the ensuing casualties. While Lonmin continued to operate, the extensive damage to its licence to do so also contributed to the subsequent 2014 South African platinum strike (Grootes 2014). Lonmin likely regained some legitimacy in eventually reaching a compromise with the workers on strike, together with other platinum companies. However, the company and its employees suffered a considerable financial setback from the five-month strike (Parker 2014).

The case of Marikana is an unfortunate example of how a decision with short-term positive impacts on a company's financial performance – saving money compared to investing in worker pay and quality of life – can have considerably larger negative consequences on its financial and social performance in the longer term, as well as grave costs to society overall. Here, as is often the case, investing more in non-financial areas of sustainability would, while initially costly, have contributed to preventing historically large losses. As the next section will continue to explore, from a sustainability perspective, financial and non-financial performance are inextricably linked. As De Villiers, Rinaldi & Unerman (2014, p. 1044) phrase it, “[financial] measures account for past performance while non-financial measures have the potential to drive future performance.”

1.3 Four Pillars of Sustainability

1.3.1 Financial Sustainability

The concept of financial sustainability builds on that of financial performance through the explicit addition of a long-term vision. Financial sustainability entails not necessarily operating to optimise the next balance sheet, but all those in the (foreseeable) future; analogous to *Our Common Future's* definition of sustainability, we might characterise financial sustainability as results that meet the targets of the present fiscal year without compromising the ability of future fiscal years to meet targets. Such an approach can drive decisions such as how to obtain and expend available resources (which can, in turn, include legitimacy) and how to approach the market in which a company operates.

As the 'triple bottom line' concept implies, and because legitimacy can be construed as an operational resource, financial sustainability often – if not inevitably – goes hand in hand with environmental and social sustainability; “actions or impacts in one area will often lead to other impacts in other areas’ (De Villiers, Rinaldi & Unerman 2014, p. 1045). Although they are seldom fully borne by the company alone, the costs ensuing from unsustainable business practices can often make ‘doing the right thing’ for a company’s stakeholders the right decision from a financial perspective as well. It bears repeating here that companies themselves often highlight this dynamic: within the sustainability-oriented genres of the corpus, just under 10% of companies express an equivalence or metonymic relationship between sustainability or CSR practices – environmental, social, or otherwise – and “good business”, “good business sense” or “good business practice”. In other words, many companies assert an interdependence between financial and other areas of sustainability.

Within the ASSET4 database that this study relies on in order to quantify economic performance, three primary factors determine the aggregate score for economic performance (at time of data collection). In addition to conventional (financial) corporate performance measures based on Key Performance Indicators (KPIs) such as margins and sales per employee, ASSET4 includes client loyalty and shareholder measures (Thomson Reuters 2012). In that respect, ASSET4 goes beyond considering the traditional KPIs, but also includes a longer-term perspective in considering that although a company might have an exceptional year, that need not reflect in sustainable financials if they are unable to retain clients and shareholders. As they themselves describe it,

[t]he economic pillar measures a company's capacity to generate sustainable growth and a high return on investment through the efficient use of all its resources. It is reflection of a company's overall financial health and its ability to generate long term shareholder value through its use of best management practices (Thomson Reuters 2013).

1.3.2 Environmental Sustainability

Virtually any company's operations have some impact – positive, negative or both – on local or global environments. In many cases, this negative impact entails waste and emissions, but some industries, such as the extractive or agricultural ones, can have an even more direct and potentially permanent impact on the lay of the land or biodiversity around the area of operations. Environmentally sustainable practices are aimed at reducing or eliminating the company's negative impact on the environment or, ideally, substituting it with a positive impact.

In the case of the extractive industries, such initiatives can include site rehabilitation or reclamation initiatives alongside safety protocols to prevent environmental incidents, which can considerably impact a company's environmental footprint and, as previously mentioned, compromise its license to operate. Even in industries with a less immediately visible environmental sensitivity, environmental footprint remains an important facet, with many companies striving to reduce (greenhouse gas) emissions, (packaging) waste, energy usage, etc. Perhaps one of the most iconic examples of such a 'greening' scheme is the ever-increasing number of airlines (e.g. Brussels Airlines, Scandinavian Airlines System, Qantas, etc.) offering 'CO2 offsetting' facilities in which the passenger is able to donate a sum commensurate with their trip's CO2 output towards carbon neutralisation initiatives in order to offset the environmental cost of their travel. Notably, however, this particular scheme relies on the passenger to choose to travel sustainably; indeed, the companies above position themselves as more environmentally sustainable because they are enabling their passengers to make this choice (Brussels Airlines 2018, Scandinavian Airlines Systems 2018, Qantas 2018).

ASSET4 describes the environmental performance pillar as

[measuring] a company's impact on living and non-living natural systems, including the air, land and water, as well as complete ecosystems. It reflects how well a company uses best management practices to avoid environmental risks and capitalize on environmental opportunities in order to generate long term shareholder value. (Thomson Reuters 2013)

As with financial performance, ASSET4 derives a company's environmental performance from three KPI-based performance categories: resource reduction, emission reduction, and product innovation (Thomson Reuters 2012). In other words, a company that uses fewer resources and creates fewer undesirable by-products in delivering their product will perform better from an environmental point of view. ASSET4 considers each KPI that contributes to these three broader categories of environmental performance on an industry-by-industry basis (Thomson Reuters 2013). That is, identical performance on a given KPI may reflect poorly on a company operating in an industry with high standards

for that KPI, while it might reflect favourably on a company part of an industry that has a lower standard for that KPI.

1.3.3 Social Sustainability

From a social perspective, companies typically engage primarily with two stakeholder groups: employees and local communities. As is the case for environmental sustainability, a socially sustainable company will minimise and ideally eliminate its negative impact on these groups, and maximise their positive impact. For instance, companies generally strive to ensure continued employee health and safety (which many companies fold together with environmental sustainability under a 'health, safety and environment' or HSE label), which includes injury prevention and promotion of good health practices.

Another particularly visible aspect of social sustainability throughout the industries that this study examines is the previously cited issue of worker remuneration. The apparel industry faces frequent scrutiny for the percentage of its income that goes to the labourers producing garments and accessories, and in recent years the mining sector has faced several such crises, most notably in the form of the aforementioned strikes. This again exemplifies how pursuing non-financial sustainability, for instance by raising employee wages, might have a negative impact on short-term profitability, but also supports financial sustainability targets.

According to the ASSET4 database,

[t]he social pillar measures a company's capacity to generate trust and loyalty with its workforce, customers and society, through its use of best management practices. It is a reflection of the company's reputation and the health of its license to operate, which are key factors in determining its ability to generate long term shareholder value. (Thomson Reuters 2013b)

The aggregate Social Performance may be the most complex within ASSET4 as it contains the greatest number of KPI-based performance (sub-)categories at seven (Thomson Reuters 2012):

- Employment Quality
- Health & Safety
- Training & Development
- Diversity
- Human Rights
- Community
- Product Responsibility

Social performance is likely also the most complex measure with respect to ASSET4's peer group division. Out of a total 88 KPIs that make up the (sub-)categories of social

performance, the database considers 38 relative to the industry the company operates in, 41 relative to the region, and 9 relative to all companies present in the database (Thomson Reuters 2013). These 9 ‘universal’ KPIs apply to the human rights category, for instance whether the company has a policy to guarantee freedom of association and exclude child, forced, and compulsory labour (Thomson Reuters 2013a, 2013b). Considering non-human rights issues from a regional or industry perspective captures, for instance, how other benchmarks for labour standards (and rights) can differ across various regional labour markets: for instance, what might be progressive in one region may be the bare minimum in another.

1.3.4 Sustainable Governance

The concept of sustainable governance is in all likelihood the least intuitively accessible to non-expert users of corporate reporting. We might argue, analogous to financial sustainability, that sustainable governance as a set of systems and mechanism that ensure that management carries out – and continues to carry out – their fiduciary duties in serving shareholders’ and stakeholders’ interests rather than their own. Such mechanisms may include having an independent board of directors, a whistle-blower policy, or regular independent audits, while infringements on these mechanisms include bribery, corruption, fraud, and other forms of mismanagement. Fraud is perhaps the most iconic case of a practice that can serve the financial bottom line in the short term, but can prove highly destructive in the long term.

As described by ASSET4,

The corporate governance pillar measures a company's systems and processes, which ensure that its board members and executives act in the best interests of its long term shareholders. It reflects a company's capacity, through its use of best management practices, to direct and control its rights and responsibilities through the creation of incentives, as well as checks and balances in order to generate long term shareholder value. (Thomson Reuters 2013b)

ASSET4 quantifies its aggregate governance score based on five KPI-based performance categories (Thomson Reuters 2012):

- Board Structure
- Compensation Policy
- Board Functions
- Shareholders’ Rights
- Vision and Strategy

The KPIs that make up these categories are numerous. For instance, they include:

- whether the company has policies to maintain various independent committees (e.g. audit, nomination or CSR committees);
- to what extent these committees comply with regulations;
- number of board meetings and attendance, as well as representation and diversity in the board;
- (absence of) conflicts of interest, company compliance with various CSR (reporting) schemes; and
- controversies and organisational features (Thomson Reuters 2013b).

1.3.5 Multiperspective Performance

In summary, a sustainability perspective on performance causes a massive shift away from considering how well a company performed chiefly by its profits or losses. While interpretations vary of what sustainable performance is and what is relevant to which aspect, we find that ASSET4's approach aligns well with e.g. Loh, Thoman & Wang's (2017), and in turn aligns with e.g. the Global Reporting Initiative (GRI) and International Finance Corporation's (2010) interpretation thereof. As such, we see sufficient reason to continue with ASSET4-based performance data as a measure of how well companies performed.

This approach does, however, come with the caveat that it becomes increasingly difficult to answer whether a company performed well in anything resembling a binary fashion. Without wishing to imply that this is an entirely straightforward matter for a strictly financial performance perspective, a sustainability perspective is entirely too complex to gauge whether a company is 'in the black' or 'in the red'. For instance, long-term financial, social or environmental performance can all come at the expense of short-term performance. Alternatively, a company with an inevitable impact can still be a leader in minimising that impact. Chapter 5 explores in greater detail how companies rhetorically position themselves around these multiple perspectives that are sometimes at odds.

1.4 Corporate (Sustainability) Reporting

1.4.1 Why Report?

Considering all the different perspectives on corporate sustainability the previous sections have explored, it is safe to assert that any company's impact inevitably extends beyond the company itself. Most companies' operations will have a considerable number

of stakeholders at various levels of society, including but not limited to clients, investors, local communities and civil society in general. As a company's impact on these various stakeholders can be substantial, it follows that most jurisdictions require listed companies to issue regular reports, typically at least on an annual basis, detailing their operations. A primary aim of such reporting is to make the company's operations more transparent to stakeholders: Lodhia & Hess (2014), for instance, claim of the mining industry that "stakeholder pressures are paramount [and] companies [...] need to provide evidence of their social and environmental responsibility to their shareholders."

Additionally, we can note that scholars distinguish between two types of Corporate Social Responsibility (CSR) based on the impetus behind them: explicit CSR occurs when companies signal their own behaviour as engaging in CSR, while implicit CSR occurs when pre-existing structures organically impel corporations towards CSR, without their explicitly indicating that commitment. Matten & Moon (2008) and Jackson & Apostolakou (2010) present US and European corporate culture as prototypes of explicit and implicit CSR, respectively; they expect explicit CSR culture to incentivise more active CSR-related stakeholder communication.

This study's main concern is with three (sub-) genres relevant to corporate reporting: the (financial) annual report, the sustainability report, and the letter to shareholders. This section will first explore the nature of, and differences between, these three text types, then discuss why their readability and use of sentiment words merit study, and finally examine what previous linguistic inquiries into these genres have revealed.

1.4.2 The Letter to Shareholders

Before we continue exploring the tension between financial and sustainability (or simply non-financial) reporting, it is worth briefly establishing why we separate out the introductory letters that almost invariably precede the reports' body text. Various called 'CEO letters', 'chairman/woman/person's letters/addresses,' 'letters to shareholders,' and so on, this is a subsection of a larger corporate report in which the CEO, chairperson, or sometimes both (when they are different individuals), address(es) the readership directly. Typically, the more conversational tone is meant to facilitate at-a-glance insight into the company's past performance and future prospects, and such letters are frequently present in both (financial) annual reports and sustainability reports. As letters from the CEO and chairperson, when those are not the same individual, can co-occur in the same report and otherwise serve an interchangeable function, we will use the common denominator of 'Letters to Stakeholders' (LTS), which in the case of financial reporting will often subcategorise to 'Letters to Shareholders' in terms of intended audience.

We treat letters to stakeholders as a separate sub-genre because they are by far the most-read section of a report (Courtis 1998, Clatworthy & Jones 2003, etc.) and offer a substantially different rhetorical situation from the rest of the document. They often represent both a synthesis of and introduction to the report, and usually offer a discursively different frame from the rest of the report. This is because they almost invariably highlight the CEO or chair speaking to the reader directly through such means as the address (e.g. 'Dear shareholder' or 'Dear reader') and direct attempts to act discursively upon the reader (e.g. 'We/I hope that you...') that would be very unusual in other parts of a report. In addition to the above, as several studies (such as Smith & Taffler 1992, Courtis 1998, Clatworthy & Jones 2003) have investigated the LtS separately from other management commentary, for instance Management Discussion and Analysis (MD&A) sections or footnotes, we also separate LtSs from the rest of this corpus to better compare with previous research.

1.4.3 The Financial (Annual) Report

The first and likely most prominent of the corporate reporting genres is the annual financial report, often just known as the 'annual report.' It conventionally focuses on the shareholder component of the larger stakeholder population. As listed companies have a fiduciary responsibility towards their shareholders, it serves as a means of enabling those shareholders to make informed investment decisions, typically containing an overview of the company's earnings and losses for the financial year, as well as an explanation behind those numbers, expectations for the future, etc.. Over time, however, it has become more common for companies to include both financial and non-financial information in their annual report, as KPMG's (2013) aforementioned indication that 51% of listed companies included at least some non-financial information in their annual reports already suggested.

The preceding sections have also begun to explore how, given the conventionally fairly homogeneous shareholder and analyst audience, the annual financial report has an almost stereotypical reputation of containing fairly to extremely difficult language. This reputation exists both in the general opinion and according to scholarly inquiry (Courtis 1995 & 1998; Stanton & Stanton 2002; Li 2008, etc.). However, given the considerable scholarly attention financial reporting has already enjoyed, this study opts to cast the genre in a more liminal role. Financial reporting is only present in this study in the form of letters to shareholders extracted from annual (financial) reports. Instead, this study focuses most of its attention on a newer type of reporting: the sustainability report.

1.4.4 The Sustainability Report

Alongside the exponential surge in adoption of ‘sustainability’ thinking and practice, standalone environmental or social reports gradually evolved into more holistic ‘sustainability reports’ that disclose material information on multiple aspects of sustainable performance (environmental, social and governance aspects, sometimes in addition to financial aspects) and, in a number of reports, how they interrelate (De Villiers, Rinaldi & Unerman 2014). Such reporting typically happens on a voluntary basis (see e.g. Nazari, Hrazdil & Mahmoudian 2017), although exceptions to that voluntariness exist, with South Africa’s King Code one of the most notable (see below).¹ In spite of this voluntariness, there are a number of standardisation initiatives for non-financial and combined forms of reporting. Examples include the Carbon Disclosure Project (CDP), which seeks to facilitate and aggregate carbon emissions disclosures (CDP 2018), the Global Reporting Initiative (GRI), which seeks to standardise “multi-stakeholder”-oriented sustainability reporting (GRI 2018), and the International Integrated Reporting Council, which aims to “establish integrated reporting and thinking within mainstream business practice as the norm in the public and private sectors” (IIRC 2018). Given their relevance to this study’s corpus, we will briefly explore the latter two in greater detail.

1.4.5 Standardisation Initiatives

The dominant (but generally still optional) framework for sustainability reporting is the Global Reporting Initiative (GRI) guidelines, which offer scalability in the form of different levels of self-declared compliance and are the de facto standard for sustainability reports’ form and content (Temouri & Jones 2014). The GRI offers companies reporting non-financial performance (they themselves primarily use the term “sustainability reports”; GRI 2018) various levels of principles and requirements those companies may choose to subscribe to when issuing said reports. These requirements emphasise, *inter alia*, multistakeholder engagement, transparency, and materiality (i.e., identifying those issues that are key to a company’s sustainability). These requirements’ stringency increases as companies choose to report at higher levels of compliance, a choice which typically depends on the company’s available means and sophistication in terms of CSR processes and infrastructure. The GRI also facilitates independent assurance for that

¹ Appendix 1 contains a list of requirements that may impact CSR disclosures for regions represented in the corpus. As many are subject to interpretation (chiefly what is ‘necessary’ for a complete understanding of the company), there are only two companies in the entire corpus (Maurel & Prom and Eramet, i.e. the two French companies) whose reporting process is certain to have faced considerable influence from reporting requirements.

reporting and compliance process, which they describe as similar to but distinct from external auditing for financial reporting (GRI 2013).

Although the GRI guidelines distinguish sustainability reporting from financial reporting, financial and non-financial content are not always exclusive to different documents; a fair few companies combine the two into a single report. In the majority of cases at the time of data collection, this entailed including some non-financial performance information in annual reports that still emphasise financial performance. Other companies offer a summary of non-financial performance in their annual report, while publishing a separate report or website detailing non-financial performance. Others still offer more evenly weighted 'integrated' reports that attempt to devote equitable attention to every relevant performance aspect, be it financial or non-financial. Integrated reporting has received an impetus towards standardisation from the Integrated Reporting (<IR>) framework (IIRC 2013), although this framework was insufficiently in place at time of data collection to meaningfully account for its presence or absence in the corpus (see e.g. Wee et al. 2016). Notably, its adoption appears to be rapidly accelerating, with, South Africa mandating an integrated report – not just some form of sustainability reporting – as a listing requirement (Intitute of Directors in Southern Africa 2009; see also Nazari, Hrazdil & Mahmoudian 2017).

While an integrated approach to reporting might also integrate financial and non-financial audiences into a joint audience for a single document, it would not necessarily improve non-experts' ability to deal with reading difficulties. That is, it might still tailor its reading level to the experts rather than the wider audience.

1.4.6 The Stakeholder

For the purposes of this study, perhaps the most crucial difference between financial and non-financial disclosures remains that while the former almost invariably addresses the company's shareholders, the latter's content is far more likely to enjoy the interest of a wider group of corporate stakeholders. Shareholders have a direct financial stake in the company, and the company and shareholder's interests are typically aligned: when the company benefits, so does, generally speaking, the shareholder. While all shareholders are stakeholders, the inverse is not necessarily true.

Stakeholders, due to their greater diversity, are a more difficult group to describe. Sacconi (2004) distinguishes between direct stakeholders and indirect stakeholders. Direct stakeholders are those affected by the company's operations because they have made some form of investment in it, although this need not be a financial investment; it can also be one of time or trust. Even in these situations, companies can still count on somewhat aligned interest with direct stakeholders, many of whom are employees or partners that will typically want to see the company succeed. This same assumption does

not hold for indirect stakeholders, who are impacted by the company's operations, be it positively or negatively, without necessarily choosing to be. Communities local to the area in which the company operates are the prototypical example of indirect stakeholders. This potential lack of direct involvement sets indirect stakeholders apart, as companies cannot assume that these indirect stakeholders' interests align with their own; in fact, indirect stakeholders' interests - typically environmental or social interests rather than financial - can often be at odds with a company's.

These indirect stakeholders represent one of the key differences between financial reporting and sustainability reporting. They may have relatively little interest in the company's financial results, and thereby financial reporting, but they are an important part of the audience for sustainability reporting; some reporting companies explicitly address their customers, local communities, or concerned citizens and civil society in general. For instance, the Adidas Group's (2013) sustainability reports welcomes their reader to the report with the acknowledgement that "you and many of our consumers and stakeholders have high expectations [...] when it comes to [our] sustainability efforts." Total (2013), in turn, highlights its efforts to submit its CSR report to wider-circle stakeholders such as NGOs or governments local to their operations and gather their feedback. Third-party surveys, such as Townsend, Bartels & Renaut (2010), indicate that these reports' readership is indeed fairly diverse; more than one-third of the readers they polled were non-investors external to the company.

This same diversity can also translate into distance. These indirect stakeholders are, by their nature, furthest removed from the company's operations, while still being affected by them. Townsend, Bartels & Renaut (2010) found that the average reader read approximately three reports annually and only 5% of the audience read ten reports or more on a yearly basis; in other words, many readers in the sample are likely non-experts that lack the experience of veteran analysts or investors. Consequently, companies can make fewer assumptions about the interests or expectations of this wider stakeholder audience, or their expertise: these members of their reports' broader audience can be experts, laypersons, or anything in between. As such, although financial and sustainability reports share a very similar presentation, their content and delivery should exhibit substantial differences given the different audiences they address, especially in terms of linguistic and visual accessibility. However, "[f]rom an accountability perspective, *all* stakeholders who desire to have [the information contained in these reports] should be able to retrieve it, no matter how economically powerless they are" (Bouten 2011, p. 2). The question of these reports' accessibility is, from an accountability perspective, an absolutely crucial one.

In summary, this divergence in potential audiences between financial and non-financial reporting is one of the driving factors behind this study: as the following section will explore, the corporate annual report is a genre with linguistic peculiarities that may

be commonplace for its expert audience, but most likely translate poorly to a wider, more general audience.

1.4.7 Linguistic Inquiries into Corporate Reporting

As stereotypes surrounding the genre would have it, two of the more prominent features of corporate reports' language are complexity on the one hand (Curtis 1995 & 1998; Stanton & Stanton 2002; Li 2008, etc.) and (excessive) positivity on the other. The so-called 'Pollyanna Effect' (Hildebrandt & Snyder 1981, Rutherford 2005) posits that companies may exhibit extreme and potentially undue optimism, even when reporting poor performance. Consequently, it is a frequent perception that corporate reporting generally makes for difficult, cumbersome reading and contains company-serving bias rather than a balanced assessment of performance. This reputation of impenetrability and imbalance – and the previous studies confirming it – in large part motivate this study's paths of inquiry, which are readability and sentiment, respectively. The chief question becomes whether sustainability reporting exhibits the same linguistic tendencies because (or in spite) of its greater voluntariness. The following sections summarise the main issues facing the language of corporate reporting, and attempt to offer a succinct overview of previous studies' insights into those issues.

As the preceding sections explored, the crux of these reports' accessibility is that their extended stakeholder audience is likely not as well equipped to deal with linguistic complexity as financial reports' core audience of investors and analysts, nor with the complexity of its contents (De Villiers, Rinaldi & Unerman 2014). Even if we only consider the form rather than the content, text remains less effective if it is only accessible to the linguistically better-equipped section of its potential audience. As Curtis (1998, p. 459) phrased it, "effective communication of narratives will be improved if those responsible for writing prose passages are responsive to the reading and comprehension abilities of their audiences."

A company's awareness of its indirect stakeholders might impel it to write more accessibly when addressing those groups in addition to its direct stakeholders. However, we can discern at least two closely intertwined factors that may work against this shift towards more accessible writing for a broader audience. The first is sustainability reporting's similarity to financial reporting; Cho, Michelon & Patten (2012a, b), for instance, implicitly equate financial and sustainability reporting in terms of presentation through applying graphical content analysis methodologies previously supplied to annual reporting to sustainability reports. The second incentive for companies to maintain a comparatively difficult writing style is a possible bias towards equating complex writing with credible writing; Chartprasert (1993), for instance, found that

readers of informative prose attribute greater expertise to a more bureaucratic writing style.

In line with Cho, Michelon & Patten (2012a, b), we can note that throughout its rise to prominence, the sustainability report has largely adopted the style and presentation of its older sibling, the (financial) annual report. Many companies, for instance, preface their sustainability reports with a Letter to Stakeholders - a CEO letter or chairman's address - just as they do their financial counterparts, and, like the financial report, most sustainability reports offer a synthesis of graphs and figures with management commentary creating a narrative around performance and expectations for the various aspects. As the sustainability report spawned from the financial report, it is more likely to also emulate the latter's linguistic structure; to do otherwise would require a deliberate, conscious departure from that tradition on the author's part. It seems plausible for authors to assume that linguistic complexity might enhance credibility based on the same logic of the non-financial disclosure resembling the genre of financial disclosures in presentation. This might further discourage reports' authors from making sustainability reports more accessible. That is, authors' choice of linguistic structure might deliberately evoke the financial report's voice and linguistic patterns in order to simultaneously evoke its reputation, and legitimate the disclosure. We might construe this imitation as a process of mimetic isomorphism (DiMaggio & Powell 1983; see sections 4.1). Creating an alternate, multi-stakeholder-oriented voice might then also risk a perception, especially amongst the core stakeholders that financial reporting targets, of lessened credibility or authority.

We must also acknowledge the potential disconnect between those stakeholders that a company wants to engage with through its reporting, and those with which it *claims* to want to engage; Lu & Abeysekera (2014), for instance, find evidence of Chinese companies adopting disclosure strategies to "gain or maintain the support of particular powerful stakeholders" (p. 38). Prado-Lorenzo, Gallego-Alvarez & Garcia-Sanchez (2009) conceptualise stakeholder power, drawing on Ullmann (1985), as the extent to which a stakeholder "controls resources critical to the organization" (p. 96). Those stakeholders that are least equipped to deal with textual complexity may also be those with the least leverage over a company's operations, precisely because they are indirect stakeholders; local communities, for instance, may not be fluent in the language the report is written in. In short, sustainability reporting's textual accessibility may also suffer in situations where there is greater incentive for a company to *signal* that it is engaging with various stakeholders through channels like the report than there is incentive to *actually* engage with those stakeholders through these reports.

In terms of the sentiment (typically good or bad news) present in these reports, their reputation of favouring positive language can, again, be explained by the incentives in place: as companies will generally want their audience's impression to be as favourable as possible, they will typically benefit from emphasising good news over bad. In this

respect, the difference in legislation between financial and sustainability reporting may also influence the reports: financial reports are generally more heavily regulated and will thus offer fewer opportunities for impression management. As many companies issue sustainability reports on a voluntary (and thereby often unregulated) basis, most of their incentive to comment on less desirable outcomes will stem from reputation management concerns and the choice to address an issue rather than remain silent on it, rather than an ethically motivated (if potentially costly) desire for transparency. We find evidence of that tendency in Cho, Michelon & Pattern's (2012a, b) observations of graphical impression management (such as graphs with non-zero bases) that favours and distorts outcomes to present the companies in their sample in the best possible light; they also indicate the reports' typically voluntary nature as one of the likeliest contributors to that phenomenon.

Iivonen & Moisander (2014) interpret reputation-driven communication on negative (CSR) outcomes as a process of sense-making, in which "a disruption in the state of the world [breaches] the expectation of continuity" (p. 650). They indicate that companies may choose to engage in "narcissistic CSR" (p. 650) when "the very core of their business strategy is called into question and the interest of the organization appear to be at odds with the interest of their stakeholders and the public good." (p. 650) Lonmin's efforts to address the Marikana Miners' Strike and ensuing casualties is one such example in that remaining silent on the issue would have likely caused further adverse effects from an impression management perspective. In less severe cases, companies may choose to devote as little attention as possible (potentially none) to unfavourable CSR-related news and attempt to optimise the reputation their CSR efforts might yield them. Displaying such behaviour, especially when *reporting* on CSR benefits the company more than *pursuing* them, is often termed 'greenwashing,' which Nazari, Hrazdil & Mahmoudian define as "attempt[ing] to convey an image of responsible corporate citizenship [that] is inconsistent with actual social and environmental performance" (p. 167).

In summary, depending on the environment in which companies operate there may be considerable incentive for them to practice reputation and impression management. If their reports' readers can more easily decode the favourable news than the unfavourable, that may positively influence their impression of the company. This 'obfuscation hypothesis'² (Courtis 1998) is one of the cornerstones of research into financial reports' readability, but remains under-examined for sustainability reporting.

The more practically oriented chapters of the study will focus on four key questions, addressing them with a tailored dataset and methodological approach:

² The obfuscation hypothesis posits that companies will attempt to conceal unfavourable outcomes by making them more difficult to discern, e.g. phrasing them with greater complexity than favourable outcomes.

- Chapter 3: how does sustainability reporting compare to financial reporting in terms of readability, and how and why does that readability vary? We address this question with a large, diversified corpus in terms of genre, industry and language variety using quantitative methods.
- Chapter 4: how do sustainability reporting's wider and core audiences report to changes in readability? We present a group of laypersons and readers more experienced with corporate reporting with three versions of a Letter to Shareholders: an original, difficult one, a moderately simplified version, and a highly simplified version. We then query how these changes affect readers' perceptions of the text and company.
- Chapter 5: how positive is the genre and how do good and bad news impact its use of language? We had the Letters to Stakeholders from the corpus' sustainability reports annotated for good and bad news in terms of the different performance pillars as well certain linguistic choices (such as use of passive voice) and explore their incidence and patterns.
- Chapter 6: what determines the genre's perceived readability and how can authors optimise it? We analyse scores manually assigned to excerpts from sustainability reports as well as the reasoning behind those scores to discover what influences readers' perception of difficulty, and train a readability prediction machine learner based on those scores.

As a core component of three out of these four main research questions, readability constitutes this study's primary thrust of inquiry. To further explore the linguistic dynamics of the genre, the following chapter will detail how we can conceptualise and measure readability, and how inquiries into corporate reporting have approached in the past. Subsequent sections will then explore other means of quantifying and describing linguistic aspects of the genre of corporate reporting (and CSR disclosures specifically). These include machine-learning based complexity measurement and attempts to describe sustainability reporting from a per-performance-aspect polar sentiment perspective (see Chapter 5), as well as attempts to measure and manipulate audience perception. As these further approaches heavily tie into themes of readability and obfuscation, however, we begin by exploring these seminal vectors for inquiry into corporate reporting.

Chapter 2

Readability

2.1 What is Readability?

Although the concept of ‘readability’ plays a central part in this study of the linguistics of corporate reporting, it suffers from a lack of a single straightforward definition of or consensus on what falls within its purview. Is the readability of a text an intrinsic quality of the text, or does it vary as the readers do? Is it simply a matter of its words and their sequence, or do visual factors such as font and text colour also affect readability? How can or should we quantify readability? Because readability is so difficult to delineate, there are as many approaches to it as there are scholars examining it. This first section explores conceptions and definitions of readability. From those different conceptions, we will distil a working definition of readability suitable to this practice-oriented study. Finally, we describe the different ways in which this study attempts to quantify and measure readability.

Given the many ways in which one text – in the broadest sense of the word – can differ from another, not every text will be equally accessible to every potential reader. Capacity for dealing with different texts – literacy - can vary between individuals based on their command of graphemes, lexicon, syntax, world knowledge, awareness of intertextual connections between the text they are reading and others, and even the capacity of their memories (Jacobson et al. 2011). This means a text’s accessibility will also vary as its demands on potential readers do.

Prototypically – and stereotypically – we might illustrate this difference with the difference in linguistic complexity between Dr Seuss’ *The Cat in the Hat* (1957) and James Joyce’s *Ulysses* (1922). For instance, the following table compares some of the more difficult sentences in either:

Table 1 Examples of difficult sentences in *The Cat in the Hat* and *Ulysses*. Flesch Reading Ease Index (FRE) and Flesch-Kincaid Grade Level (FKGL) scores (see section 2.3) appended in bold.

<i>The Cat in The Hat</i>	<i>Ulysses</i>
<p>“Now look what you did!” said the fish to the cat. “Now look at this house! Look at this! Look at that!” You sank our toy ship, sank it deep in the cake. You shook up our house, and you bent our new rake. (p. 24)</p> <p>FRE: 115.6 FKGL: -1.2</p>	<p>Fatherhood, in the sense of conscious begetting, is unknown to man. It is a mystical estate, an apostolic succession, from only begetter to only begotten. On that mystery and not on the madonna which the cunning Italian intellect flung to the mob of Europe the church is founded and founded irremovably because founded, like the world, macro- and microcosm, upon the void. (p. 266)</p> <p>FRE: 45.3 FKGL: 12.1</p>
<p>Then we saw him pick up all the things that were down. He picked up the cake, and the rake, and the gown, and the milk and the strings, and the books, and the dish, and the fan, and the cup, and the ship, and the fish. (p. 57)</p> <p>FRE: 98.4 FKGL: 5.4</p>	<p>No question but her name is puissant who aventried the dear corse of our Agenbuyer, Healer and Herd, our mighty mother and mother most venerable and Bernardus saith aptly that she hath an <i>omnipotentiam deiparea supplicem</i>, that is to wit, an almightiness of petition because she is the second Eve and she won us, saith Augustine too, whereas that other, our granddam, which we are linked up with by successive nastomosis of navelcords sold us all, seed, breed and generation, for a penny pippin. (p. 511)</p> <p>FRE: -12.4 FKGL: 35.9</p>

Few would contest that these are polar opposites in terms of linguistic complexity.¹ This chapter aims to explore why such a contrast exists and how it expresses itself, while the

¹ Although the difference in readability is clear, these results already indicate a few issues with the formula-based method of quantifying readability that section 2.3 will explore in greater detail. For one, according to the Flesch-Kincaid Grade Level score, the second example from *The Cat in the Hat* requires six years of additional reading experience compared to the first. This is due to the much longer sentence, but the actual difference in difficulty that humans would perceive is much smaller than that. Conversely, the first example from *Ulysses* benefits from the two shorter sentences preceding it in order to appear, to the formulae, more readable than it likely is. The second example from *Ulysses*, in turn, registers amongst the most difficult sentences one can plausibly encounter in English due to its extreme length. Strictly speaking, the FKGL formula estimates this sentence as requiring 24 years of education past secondary to fully decode it. Without wishing to speak to that

majority of the study (chiefly Chapter 3, Chapter 4 and Chapter 6) explores linguistic complexity in corporate (sustainability) reporting.

As to how linguistic complexity manifests in written language, we can observe, for instance, that *The Cat in the Hat* employs shorter and more simple (as opposed to complex) sentences, as well as frequently occurring and monosyllabic words. It also assumes practically no world knowledge or familiarity with other works on the reader's part. By contrast, *Ulysses'* stream-of-consciousness writing style at times shirks the very concepts of syntax or (shared) vocabulary, and the title already announces itself as an intertextual reference. We might argue that *Ulysses* would be unable to achieve what it sets out to achieve using simpler language; its relative impenetrability is part of its reputation, appeal, and even status as a work of art. However, that also limits the audience with which it can meaningfully interact.

The intended audience of these texts, as the previous sections have explored, plays a crucial role and is perhaps *the* defining difference (out of many) between the two works. While *Ulysses* demands profound mastery of not only the language, but its cultural background in the widest sense, Dr Seuss' writings are tailored to novice – and often younger – users of English. These include children in the process of (native) language acquisition that do not yet have the linguistic competence or experience to interact meaningfully with texts that are more complex. We do not mean this to imply that the readability of a text – as an artefact – changes depending on the audience. Dale & Chall (1948) or McLaughlin (1969), however, do consider the audience a part of readability.

We would rather argue that based on how a text's readability manifests, it will be more or less appropriate for certain audiences to engage with and extract from it what they want – the text remains equally readable regardless of the audience, but different audiences and texts are better suited to one another. An audience's ability and willingness to engage with texts less readable than they are comfortable with will, in turn, vary depending on individual factors such as interest and motivation.²

Having established how texts can differ in readability – as we move towards a working definition of that concept – we must also wonder how we can meaningfully measure it. Offering language learners, as well as students in general, reading material that is sufficiently challenging to enable learning, but not so challenging as to incite frustration, can be crucial (Shanahan, Fisher & Frey 2012) to their learning progress. In order to facilitate this process, scholars such as Rudolph Flesch (1946) and J. Peter Kincaid (et al.,

assessment's accuracy, we can assert that none of these analyses account for the background knowledge (or lack thereof) required to understand the texts.

² We note that in an educational context, challenging reading material can often be the optimal choice to help build reading proficiency, with the learner's motivation, intrinsic or extrinsic, helping them over the hurdle of a more difficult text. This increases their ability to deal with other texts of similar difficulty (Peterson et al. 2000, Shanahan, Fisher & Frey 2012).

1975) have attempted to create numerical expressions of a text's readability. Such formulae express difficulty through a scale with predetermined 'bands' of difficulty, such as 'easy' or 'very difficult' (Flesch 1946 p. 205) or a grade level that estimates which level of education a given text might require to decode. Such formulae typically employ weighted sets of easily computable textual characteristics such as average word and sentence length, rather than 'deeper-level' characteristics such as syntactic complexity or required background knowledge. Although they only offer a surface-level estimate of a text's readability, these formulae often correlate with more advanced complexity measures (e.g. Pearson 1974). Nonetheless, an important caveat of readability formulae is that they are a resource-efficient means of estimating – but only estimating – a text's readability; the popular readability formulae are not a universal, authoritative expression of how accessible a text is. By design, they cannot be.

The subsequent sections explore text difficulty and the different ways it manifests, then looks at some of the theory and numbers behind the formulae. Sections 2.6 and 2.7 then explore what alternatives to readability formulae can contribute to the question of text difficulty.

2.2 Why Readability Matters

DuBay's (2004) *The Principles of Readability* offers a comprehensive overview of the formula-based approaches to readability in addition to elucidating why readability research matters: many important texts in everyday life are too difficult to be generally readable. DuBay cites the example of child-safety seat installation instructions, the readability of which Wegner & Girasek (2003) examined and found written at a 10th grade level, three levels more difficult than the 7th grade textual complexity that the average U.S. adult is equipped to deal with (Snyder & Hoffman 1993).³ These results prompted a surge of attention to these instructions' language, and justifiably so: overly difficult writing is less effective writing, because it is less likely to communicate to its entire audience what it means to communicate. In the case of safety instructions, making sure the largest possible audience can understand the message is certainly paramount. We can apply the same reasoning to most forms of professional writing: more readable writing has a better chance to be fully understood by more of its audience, and is thus more effective.

³ The National Assessment of Adult Literacy (2016) shows little to no evolution between 1993 and 2016 in terms of reading proficiency in the U.S.

We deliberately specify ‘most’ forms of professional writing. In some cases, the more ‘professional’ or authoritative voice that more complex language can evoke may at times be more desirable than ease of understanding. Furthermore, some writers may not wish for the audience to wholly understand the message in specific cases, such as potential attempts by corporate report writers to obfuscate less favourable information that they must nevertheless report on. We elaborate on this previously anticipated ‘obfuscation hypothesis’ (Courtis 1998, Rutherford 2003, Bayerlein 2010, etc.) in later sections; we will first explore how we can measure this linguistic complexity.

2.3 Readability Formulae

In professional contexts, just like in educational contexts, authors and users of texts generally want to ensure that a text is appropriate reading material for a particular audience. Composing a text that better (more clearly) addresses its audience is almost always more resource-effective. If authors want to ensure (cost-)effective communication, they need some means of measuring, or at least estimating, how readable that text is. The conventional solution to that problem, amongst professionals and academics alike, continues to be the aforementioned use of readability formulae. Readability formulae rose to prominence in the early-to-mid 20th century, and heavily influenced thinking about readability (DuBay 2004). For instance, the most popular readability formulae which still see use today, such as the Flesch Reading Ease Index (Flesch 1946), the Flesch-Kincaid Grade Level (Kincaid et al. 1975) or Gunning Fog Score (Gunning 1952),⁴ compute a text’s readability only through average word and sentence length. The table below offers a few examples from the corpus of relatively easy and difficult sentences according to these formulae, and their scores on the respective scales:

⁴ These three formulae, though by no means the only ones, are the main readability formulae that continue to see use in modern readability research, with for instance Courtis (1995, 1998) using the Flesch Reading Ease Index (with the former also using the Fog score), Abu Bakar & Ameer (2011) using the Flesch-Kincaid Grade Level, and Farewell, Fisher & Daily (2014) using all three measures. Including all three thus enables greater comparability with other studies and corpora.

Table 2 Examples of easy and difficult sentences according to formulae.

Formula	Easier	More Difficult	Source ⁵
Ease Index Flesch Reading	The process is summarised in the following five steps. (66.1)	The Asia business unit has responsibility for operations in Laos and for supporting the implementation of business development strategies within Laos and the region. (2.7)	PanAust
Grade Kincaid Flesch-	Long term, however, we are aiming for an annual reduction of 10%. (7.8)	In 2012 we completed the first phase of our work with external consultants DuPont who helped us to develop and implement core Health and Safety standards and procedures. (13.9)	Kazakhmys
Index Gunning Fog	Treatment success rate is in excess of 90 per cent, which is among the highest in the country. (7.2)	The nature of occupational illnesses is changing. Health conditions such as stress, fatigue and the normal results of ageing, such as reduced physical capacity, present different challenges from the traditional mining health issues. (16.3)⁶	Rio Tinto

Readability formulae see widespread use in corporate communication, but also serve as a legislative standard. For instance, many U.S. States mandate a maximum level of complexity for documents that require a stakeholder’s understanding for informed consent, such as insurance policies; for example, the US National Association of Insurance Commissioners (1995) requires a minimum Flesch Reading Ease score of 40 for life insurance policies. This reveals a first caveat to formula-based readability: a solely formula-based line of thinking might tempt writers to minimise word and sentence length at the expense of other factors that might influence readability more, such as how common those words are.

2.3.1 The Flesch Reading Ease Index

The pioneers of formula-based readability estimation attempted to create an objective means of estimating textual complexity that an author could calculate by hand. The 1940s saw the development and introduction of Rudolf Flesch’s Reading Ease Index (FRE), perhaps the best-known readability formula to this day. Flesch’s (1979) *How to Write Plain English* explores some of the thought processes behind the Reading Ease formula, which requires six steps to calculate (by hand, if desired) for a given piece of writing:

⁵ Each of these examples originates from the company’s sustainability report covering 2012.

⁶ As with the examples from *The Cat in the Hat* and *Ulysses*, we can note that although the Gunning Fog Index here indicates the lowest readability (i.e. highest number of years required for comprehension) out of these examples, the score could be further inflated by removing the shorter first sentence.

1. Count words

(Contractions, hyphenated words, abbreviations, figures, symbols and combinations thereof count as single words.)

2. Count syllables

(Count as pronounced, favouring pronunciation with fewer syllables. Abbreviations, figures, symbols and combinations thereof count as single syllables.)

3. Count sentences

(Count full units of speech divided by period, colon, semicolon, dash or question mark or exclamation point, but disregard those if within sentence.)

4. Average syllables per word

5. Average words per sentence

6. Determine score

- Multiply average sentence length by 1.015.
- Multiply average word length by 84.6.
- Add them together.
- Subtract from 206.835.

(Condensed from Flesch 1979)

Ease of use is important in these early formulae: *How to Write Plain English* even includes a readability chart that lets users determine readability by drawing a line between the average number of words per sentence to the average number of syllables per word. That line then intersects a line of readability scores; the point where they intersect gives the user the score. Flesch tries hard to keep the formula easily computable for anyone, although a calculator will probably be useful. A computer can calculate the FRE score for a piece of text almost trivially; results may be slightly less precise, however. A computer may not always count syllables or sentences entirely correctly. Pronunciation matters when counting syllables, so a computer needs some system to separate a sequence of letters (i.e. a word) into syllables. ‘Leicester’ and ‘Worcester’ are perhaps two of the best known examples of words with fewer syllables than the word picture might suggest, but words that violate conventional pronunciation rules, like ‘colonel,’ similarly need a more sophisticated approach than counting non-sequential vowels as syllables. In other words, these systems can make mistakes. Similarly, the question whether punctuation marks two sentences or exists within one can also cause errors. Nevertheless, modern computer systems can calculate these formulae extremely quickly. Due to the above issues, implementing readability calculation is not trivial, even though the calculation itself is.

The FRE offers a weakly bounded 0-100 interval, with higher scores representing higher readability. It distinguishes at least six tiers of readability (Flesch 1979), which the following table captures:

Table 3 ‘Bands’ of readability as distinguished by the Flesch Reading Ease Index.

FRE Score Band	Difficulty
90-100	Very Easy
80-90	Easy
70-80	Fairly Easy
60-70	Plain English
30-60	Difficult
0-30	Very Difficult

The most important tiers for this study are 70-100 and 0-30. Texts with an FRE above 70 should be universally readable, while the FRE considers those texts under 30 ‘very difficult’. As the FRE is weakly bounded, it is possible, if rare, for texts’ readability to exceed 100 or go below 0; while the FRE does not define these areas, we can safely assume they are particularly easy or difficult to read, respectively: with 0 meaning “practically unreadable” and 100 meaning “easy for any literate person” (Flesch 1962, p. 216). The highest score for running text is 120, assuming only sentences of two words, each a syllable long. Theoretically, single-word, monosyllabic sentences would yield a score of 121. As English allows for arbitrarily long sentences, there is no theoretical minimum to how low the Flesch Reading Ease score can go.

2.3.2 Flesch-Kincaid Grade Level

Kincaid et al. (1975) attempted to address a weakness they perceived in using the FRE to determine the appropriateness of reading material for navy personnel. While the FRE offered a useful general estimation of how readable a text would be to the general public, it required conversion in order to determine whether a text was suitable to a given reading level.

The Flesch-Kincaid Grade Level is an adaptation of the Flesch Reading Ease score that expresses difficulty as a grade level, without requiring that conversion. That makes it inversely proportional to the FRE: whereas a higher FRE score indicates greater readability, a higher Flesch-Kincaid grade level implies a less readable text. As with the FRE score, the formula allows for arbitrarily low levels of readability (or high levels of reading difficulty), and scales with average number of syllables and words per sentence. It offers a more precise standard variant, and a simplified, but slightly less accurate variant (by one-tenth of a grade level) that is easier to calculate manually.

The authors indicate the following steps, which are very similar to those for the FRE, to calculate the grade level for a text:

1. Count words

(Contractions, hyphenated words, abbreviations, figures, symbols, and any groups of letters surrounded by whitespace count as single words.)

2. Count sentences

(Count grammatically independent units, including sentence fragments, separated by period, colon, semicolon, dash or question mark or exclamation point. When such interpunction occurs within a single sentence, only count it as one sentence.)

3. Count syllables

(Count as pronounced. Count symbols and numbers as they are normally pronounced.)

4. Average words per sentence

5. Average syllables per word

6. Determine score:

Standard: $0.39 * (\text{words/sentence}) + 11.8 (\text{syllables/word}) - 15.59$

Simplified: $0.4 * (\text{words/sentence}) + 12 (\text{syllables/word}) - 16$

(Condensed from Kincaid et al. 1975, p. 38-39)

The authors' original study did not label this formula the 'Flesch-Kincaid Grade Level' score; later users did. Kincaid et al. simply conceived of it as a different way to calculate and express readability, focused on Navy use, that would align with the core goals (and variables) of the FRE. They composed the formula based on cloze testing (see section 2.4.3.1), with grade levels for specific texts determined by 50% of readers at a given reading grade scoring at least a strictly corrected 35% on a text with every fifth word blanked (Kincaid et al. 1975). The FKGL score rapidly exceeded the relatively narrow purpose of determining readability of Navy material, however: it is suitable for, and has seen use in, virtually any type of readability estimation, likely because of its more intuitive to interpret output.

The US Grade level scale not only indicates level of education required for comprehension, but in many cases also allows a user to estimate the appropriate age (in terms of text composition, not necessarily content) for a given text.⁷ The table below summarises Flesch-Kincaid Grade Level scores, educational levels, and typical ages ranges associated with them (lower is more readable). We can expect the scale to be less accurate towards the extremes.

⁷ The US grade level consequently requires conversion when comparing it to other regions' educational standards. As a rule of thumb, subtracting six from the score gives an estimate of years of primary education and above required for understanding (assuming normal progress).

Table 4 Overview of reading grade levels.

FK Grade Level	Educational level	Age range
-3.4-1	Pre-elementary (e.g. nursery schools)	<6
1-6	Elementary or primary education	6-12
6-12	Secondary education (middle and high schools)	12-18
12-16	Undergraduate	18-22
>16	Postgraduate (e.g. doctoral)	>22

The Flesch-Kincaid Grade level score offers less of an intuitive cut-off point for universal readability than the Flesch Reading Ease Score's '70 or above' for Plain English. Other studies, however, do suggest thresholds for grade level readability: Doak, Doak & Root (1996) approach grade levels from the perspective of medical document readability. They point to 5th-grade reading level as a threshold of functional literacy; they expect readers that read at a 5th grade level to struggle, although they can technically read, as the vast majority of texts in daily life are above the 5th grade reading level. When they conducted their study, they found the average reading level of adult Americans to lie between 8th and 9th grade reading level.

Based on these findings, we would argue that the mark for (near-) universally understandable text on this scale would lie around the 5th-grade mark, and the mark for generally understandable text would be an 8th-grade reading difficulty at the highest. We do not expect these targets to have changed meaningfully since Doak, Doak & Root's study: the National Assessment of Adult Literacy (2016) shows no significant evolution in average reading ability among US adults between 1992 and 2003, for instance, except for a modest improvement in quantitative literacy. Prose and document literacy did not improve.

2.3.3 Gunning Fog Index

The closely related Gunning Fog Index (1952) predates the Flesch-Kincaid Grade Level by over two decades. Similar to the FKGL, it attempts to translate the text's surface features into a grade-level expression of reading ease or difficulty. Approaching the readability issue from a textbook publishing background, Robert Gunning focused on the use of readability techniques for communication professionals (DuBay 2004).

This focus on readability for professionals likely contributed to the Fog score's emphasis on ease of calculation and use. For instance, where the Flesch score and its grade-level derivation deal with syllables by averaging the number per word, the Fog

Index makes whether a word is a 'hard word' a binary question – either it is polysyllabic, i.e. a three-or-more syllable word, or not (in a later revision the question becomes whether it is monosyllabic).

As the Fog Index has seen a number of revisions, we will synthesise a means of calculating its various forms from DuBay (2004):

1. Select a **100-word passage**, but respect sentence boundaries.
2. Count the **number of sentences**.
3. Count the **number of words**.
 - a. Count the number of **polysyllabic words** (three or more syllables) **OR**
 - i. Label these 'hard words'.
 - ii. Ignore proper nouns, familiar jargon, compound words, and frequent suffixes.
 - b. Count the number of **monosyllabic words** (optional).

These are only relevant for the Sumner and Kearl revision.
4. Average **words per sentence**.
5. **Calculate percentage** of monosyllabic words out of total words (optional).
6. **Calculate:**

Original: $0.4 * (\text{average sentence length} + \text{polysyllabic words})$.
Sumner and Kearl (1958) revision: $3.068 + (0.0877 * \text{average sentence length}) + (0.0984 * \text{percentage of monosyllabic words})$.
Kincaid (1975) revision: $((\text{easy words} + (3 * \text{hard words}))/\text{sentences}) - 3) / 2$

As the revised formulae make apparent, precision and refinement come at the expense of rapid manual calculation.

We can assume the same Plain English and universal understandability readability targets apply for the FKGL and GF, although results will differ due to the emphases they place. Specifically, the GF is more sensitive to word length, i.e. weighted to attach greater difficulty to polysyllabic (as opposed to mono- or disyllabic) words. Both can exceed the highest US grade level of 12, and then simply estimate the number of years beyond secondary education, e.g. higher education.

2.3.4 A Caution

While we would argue that the FRE, FKGL and GF scores are fairly shallow in their approach to text analytics, as they only consider surface-level, easily-computable variables, they are a useful tool – as Flesch (1946) himself puts it, a yardstick. He argues that while it seems like “a very crude way of dealing with writing [...] it is based on some very complicated facts of human psychology” (Flesch 1979, p. 21). He presents sentence length as an indicator of complexity as longer sentences mean more information for the mind to process when it

reaches a stop. He links this with, for instance, more subordinate clauses, and asserts that longer sentences mean more mental work for the reader. He argues the same of words: longer words contain more affixes. For instances, ‘unmistakably’ is more difficult to process than simply ‘take’. For all the reasoning behind it, however, Flesch (1946, p. xii) also cautions not to “wallow in the little rules and computations but lose sight of the principles of Plain English. What [he hopes] for are readers that [...] won’t expect more from it than a rough estimate.” The very worst-case scenario in that respect is when authors begin “writing to the formula[e]” (Klare & Buck 1954, p. 139) – that is, optimising their language to exhibit the highest possible readability according to formulae, rather than accommodating the readability requirements of their audience as best they can.

2.4 Aspects of Readability

It would be reductive to claim that the above formula-based approaches *define* readability as a function of word and sentence length; they are chiefly an effective, if not always accurate, means of approximating a text’s readability. The definitions of, and approaches to, readability vary immensely between practice- and theory-oriented approaches, but neither approach denies that readability is, at heart, a very complicated interplay of countless factors that are difficult to fully capture. DuBay (2004, p. 3) nevertheless offers an aphoristic attempt that is nearly impossible to argue against: “Readability is what makes some texts easier to read than others.” DuBay (2004, p. 3) reiterates a few definitions that scholars on readability have offered in the past:

The degree to which a given class of people find certain reading matter compelling and comprehensible. (McLaughlin 1969)

The ease of understanding or comprehension due to the style of writing. (Klare 1963)

DuBay describes Dale & Chall’s (1949, p. 5) as perhaps “the most comprehensive” definition:

The sum total (including all the interactions) of all those elements within a given piece of printed material that affect the success a group of readers have with it. The success is the extent to which they understand it, read it at an optimal speed, and find it interesting.

As becomes apparent from the above definitions, specificity, completeness, and computability are almost impossible to ensure without compromising the other(s). For

instance, DuBay's definition is succinct and virtually all-encompassing, while the Flesch, Flesch-Kincaid or Gunning Fog score's approaches to readability omit many potentially relevant factors such as the audience, the subject matter, or even word frequency or rarity. Conversely, DuBay's definition cannot reasonably compute or quantify a text's readability where the formulae can. Similarly, Klare offers a number of important variables (comprehensibility and the position of the author) while placing less emphasis on the position of the audience. The following non-exhaustive list contains some of the more prominent aspects of a text that can influence reading ease:

- Text-internal elements
 - Sentence length (Flesch 1946, Gunning 1952, Kincaid et al. 1975, etc.)
 - Word length (Flesch 1946, Gunning 1952, Kincaid et al. 1975, etc.)
 - Word difficulty or rarity (does the text use (mostly) everyday language?) (Dale & Chall 1948, 1995)
 - Syntactic complexity (Tierney, Anders & Mitchell 1987, Collins-Thompson 2014)
 - Number of passive structures (SEC 1998, Ownby 2005, Plain English Campaign 2013, Wink 2016)
 - Syntactic depth (how layered are syntactic relationships?) (Pearson 1974, Beaman 1984, Dell'Orletta et al. 2014)
 - Extent of subordination (how many embedded phrases does the text contain?) (Beaman 1984, Pitler & Nenkova 2008, Dell'Orletta et al. 2014)
 - Lexical density (what is the balance between content words and function words?) (Halliday 1989, Harrison & Bakker 1998, Castello 2008)
 - Cohesion & coherence (De Clercq 2015, Todirascu et al. 2016)
 - Semantics (underlying patterns of meaning, such as coherence and cohesion; e.g. vor der Brück & Hartrumpf 2007)
- Audience (Taylor 1953, Bean 2011, Wray & Janan 2013)
 - Subject familiarity (to what extent do readers have the necessary background knowledge?) (Bean & Weimer 2011, Wray & Janan 2013)
 - Linguistic proficiency (how good is the reader's command of the language?) (Davison 1985)
 - Motivation (why is the reader reading the text?) (Bean 2011, Wray & Janan 2013)
- Author
 - Writing style (what are the particularities of how the author writes?) (Davison 1985, Plain English Campaign 2013)
 - Linguistic proficiency (how good is the author's command of the language?) (Tierney, Anders & Nichols Mitchell 2013)

- Linguistic precision (how well does what the author conveys match what they want to convey?) (Flesch 1962)
- Language variety (which dialect or sociolect, in the broadest sense possible, does the author typically employ, which are they trying to employ, and how does that align with the reader's? See sections 2.4.4.2 and 3.2.2.)

Before exploring the above elements in greater detail, we will attempt to distil a working definition of readability based on what we have discussed so far. Our definition will focus primarily on how readability *varies*, as that is what this study examines, and makes no pretensions at being comprehensive, or fit for other purposes:

A text becomes more readable when its language better helps readers obtain information they want from it.

We limit our definition to the text itself, and consider the readers only in the abstract sense, as the same changes can help different kinds of readers. A more readable text might mean the difference between a very difficult and a functionally unreadable text for a novice reader, or between a fairly and entirely accessible text for a very experienced reader. We assume that most ways in which a text's readability can differ will impact a novice and experienced reader in the same direction, if not to the same extent.

This differs from some approaches to readability, such as Gilliland's (1968) or Wray and Janan's (2013), that approach readability as the interaction between text and reader, rather than just the language of the text itself. While this interaction certainly lies at the very core of the reading experience, the following section explores why we differentiate text-internal readability from reader-text interaction.

2.4.1 Readability versus Understandability

From a terminological perspective, we do not adhere to the notion that readability varies along with the reader; rather, we consider it an intrinsic quality of the text that derives from countless textual features. The experienced reader may still experience less difficulty reading a less readable text than an inexperienced reader would. The factors that contribute to or detract from the experience, however, will largely be the same. To draw back on the initial example, even a reader equipped to tackle *Ulysses* will find *The Cat in the Hat* an easier, if not necessarily more stimulating, read: for better or worse, it challenges the reader less than *Ulysses* does, but does not allow for the same depth or complexity of thought. We draw the boundary of what constitutes 'readability' around the text as an artefact. The more sophisticated reader might find *Ulysses* a far more compelling read than *The Cat in the Hat*. Rather than consider engagement an aspect of the text's readability itself, however, we consider it part of the reading experience, which a text's readability will contribute to along with a variety of other factors. In short, we

consider readability text-internal, and the reading experience – though certainly crucial – to derive from the text’s interaction with its audience. Consistent with Smith & Taffler’s (1992) text-internal approach to readability and differentiation from audience-dependent reading difficulty, we will refer to the ease or difficulty a given audience has understanding a text as that text’s ‘understandability’.

The next section explores text-internal criteria, and why we separate them from the reading experience. Sections 2.4.3 and 2.4.4 explore text-external aspects of the reading experience, and in doing so provide the other half of what, taken together with readability, we consider ‘understandability’.

2.4.2 Text-internal Criteria

In brief, we separate text-internal elements of readability from other parts of the reading experience because text-internal elements are far more quantifiable than text-external elements, and, ideally, measurable using Natural Language Processing (NLP) techniques. In virtually all cases, we only have access to the text as an artefact when trying to estimate how easy or difficult it would be to read. That means that the only useable measures of difficulty we have available will be those we can objectively count. For the more subjective, qualitative aspects that author and audience traits entail, we will at best be able to use countable elements of the text (such as use of personal pronouns) as proxies.

2.4.2.1 Word length, Sentence Length, and Word Rarity

We will discuss word length, sentence length and word rarity, which are the base components of the more prominent readability formulae, the calculation of which section 2.3 explored in greater detail. Chall (1996, p.24) sees a common element in these formulae being a function of “some [measure] of word difficulty” and “some measure of sentence complexity,” but also immediately cautions that merely using shorter words and sentences – writing to the formula, so to speak - will not necessarily make a text easier to understand. Neither are complex words or sentences the root cause of reading difficulty; rather, these are easily measurable features that correlate with reading difficulty. For instance, Pearson (1974) indicates that sentence length and transformational (i.e. syntactic) complexity will tend to co-vary.

Word length and word rarity can both serve as difficulty measures for vocabulary. The intuition is that shorter or more common words are simpler, and more complex words will be longer. Most formulae that use it (such as the Flesch Reading Ease Index, Flesch-Kincaid Grade Level or Gunning Fog Score) express word length in syllables. Formulae that use word rarity (such as the Dale-Chall Formula) determine it by considering words that occur on a predetermined list common words, and those that do not uncommon

words. In this approach to measuring word rarity, the composition of the lexicon of ‘common’ words of course plays a crucial part in the outcome.

Advances in computational technology (see section 2.6.2) have also enabled other word rarity metrics that would be all but impossible to calculate by hand. One notable example termhood, i.e. the extent to which “a word/phrase [...] carries a special meaning” (Vu, Aw & Zhang 2008). Although there are many ways to calculate termhood for a given word in a given (set of) documents, one fairly intuitive approach is Term Frequency – Inverse Document Frequency (TF-IDF). The Term Frequency is the number of times we can find a term in a document divided by its total number of terms. We then multiply it by the Inverse Document Frequency – the natural logarithm of the number obtained by dividing the total number of documents in a corpus by the number of documents with the term in it (Salton 1989). While it is entirely intuitive that a text with a greater number of highly specific words will, *ceteris paribus*, likely be more difficult to understand, especially to non-experts, than one with relatively fewer. It is important to note that the TF-IDF approach considers a set of documents rather than a single one. For instance, in the scope of a general, diversified corpus, words like ‘sustainability report’ or ‘turnover’ might show a high termhood; in a corpus of only corporate reporting, they are less likely to, but company-specific technical terms still would. While our initial full-corpus analysis does not integrate termhood as a difficulty measure, Chapter 6 explores termhood in greater detail.

2.4.2.2 Aspects of lexicosyntactic complexity

While sentence length may serve as an indirect proxy for syntactic complexity, a frequent criticism echoed by studies into corporate reporting is that while the formulae it is often a core component of are useful, but ultimately also limited in their explanatory power. Curtis (1998), for instance, refers to the inevitable reductiveness of attempting to capture the complexity of language into a single variable. Li (2008), in turn, explores how readability formulae are unable to measure comprehension difficulty, and their coarseness makes them relatively ill-suited to absolute readability judgments. At the same time, Li considers formulae appropriate for relative assessment. While this study accommodates for both objections – the first in defining and approaching readability separately from comprehensibility and the second by limiting absolute readability judgments to wide ‘bands’ rather than fine-grained individual scores – such criticism does underscore the need for finer-grained features, as well as the difficulty of computing them. To expand on these analyses, the full-corpus analysis chiefly focuses on syntactic aspects of readability conducive to intuitive interpretation, while further analyses such as a machine learning-based readability assessment system integrate a larger number of deeper-level features, such as TF-IDF (see above).

Motivation

As the preceding sections explored, performing the full-corpus analysis with a tractable number of computable, interpretable features meant selecting potential candidates. Of the features above, the full-corpus analysis draws on the text's number of passives, extent of subordination, parse tree depth (i.e. syntactic depth; see below) and lexical density (with the latter, as the name implies, a lexico-syntactic feature rather than strictly a syntactic one). While incorporating semantic information as von der Brück & Hartrumpf (2007) did, for instance, would likely greatly enhance the richness of the analysis, implementing such techniques is highly demanding – enough so that the addition thereof would come at the expense of breadth in other areas.

While providing an overview of all the potential syntactic contributors to or detractors from readability would merit a study of its own (see e.g. Bailin & Grafstein 2016), we make a deliberate decision to limit the initial broad-scope analysis to easily tractable ones (in this case lexicosyntactic features). By contrast, an iconic problem of readability prediction can illustrate the difficulty of integrating potential predictors that are far more difficult to quantify once we move beyond the lexicosyntactic.

The problem of quantifying textual cohesion (Todorascu et al. 2016), alternatively framed as the task of coreference resolution (De Clercq 2015) continues to prove a difficult task, in spite of the potential added value of being able to track and quantify semantic patterns throughout the entire length of the document. An ideal coreference resolution system might be able to automatically trace how meanings and ideas exist and evolve throughout a document. For instance, such a system might be able to recognise that this paragraph (partially) equated quantifying textual cohesion with coreference resolution in the preceding sentence; it would be able to discern how that idea re-occurs later in the text, even under the guise of other synonyms. Clearly, such a task requires far more advanced NLP techniques than recognising passive structures or counting lexical and function words would. As the aforementioned authors illustrate, this is a problem that continues to require focused studies' worth of effort in order to advance the state of the art and, in spite of those efforts, continues to present difficulties; as Todorascu et al. (2016, p. 995) phrase it,

parametrization [of cohesive features] requires heavy NLP processing and is prone to errors [...] [Cohesive features] do not seem to contribute much to the prediction of text readability when compared with simple predictors such as word frequency and sentence length.

De Clercq (2015) similarly found mixed results for the efficacy of coreference resolution in enhancing readability predictions, concluding that “it is a hard task of itself” but can nevertheless be of use (p. 187). This illustrates how, based on the current state of NLP technology, there are syntactic features that may well contribute to readability

assessment and prediction tasks, but are too technically demanding to implement in a broad-scope study with a fairly large corpus. Accordingly, although Chapter 6 attempts to apply De Clercq's (2015) and De Clercq & Hoste's (2016) more advanced approach, the broader-scope analysis of 2.75m words across several hundred documents requires less demanding approaches to quantifying readability, and we therefore opt not to integrate such cutting-edge features. The subsequent sections will explore the less demanding lexicosyntactic features used in Chapter 3 (number of passives, syntactic depth, subordination and lexical density) in greater detail.

Number of passives

A considerable number of style guides, ranging from the SEC's (1998) Plain English handbook to the Plain English campaign's own handbook (2013) or Wink's (2016) guide to academic writing recommend against indiscriminate use of the passive voice. They emphasise that authors should prefer the active voice, except in a few specific cases. The two most frequent arguments against the passive voice both relate to cognitive load: a passive-voice construction is generally longer than an equally informative active-voice construction, and as the passive voice is not the default voice in English, speakers of English find the order of elements more difficult to process.

The first argument against, we can easily intuit: any given full passive sentence in English must be longer than its active equivalent given the addition of the 'to be'-form and the 'by' introducing the agent. While the short form of the passive can be shorter, it also omits information (the agent) compared to the active-voice form. For instance:

- (1) **Active:** The company made mistakes.
- (2) **Passive:** Mistakes were made by the company
- (3) **Short passive:** Mistakes were made.

The second argument entails that as the default valency pattern in English (SVO) starts with the agent (subject) and proceeds through the verb to the object(s), readers and listeners need to mentally reconstitute passive-voice forms (Pearson 1974) towards that SVO order. As the active voice does not require that step, it places fewer demands on the reader's mental faculties. Both of these elements contribute to justifying a text's number of passives as a (partial) measure of its readability. Ownby (2005), for instance, examines its impact on readability, while Abu Bakar & Ameer (2011) quantify it as a readability measure for CSR reporting.

The specific cases where style guides condone the use of the passive typically involve the short passive, and its ability to omit the agent: the passive is useful when the author wishes to emphasise the action or the object rather than the agent, or conceal who carries out the action altogether. As both academic and corporate writing tend to be action or result-oriented rather than people-oriented, the passive voice helps achieve such a style. Even appropriate use does not make the passive easier to process, however, as the SEC

guidelines (1998) indicate. Section 3.2.4 explores the short passive in corporate reporting as a defensive attribution strategy.

In terms of benchmarking, fairly few previous studies quantify the number of passives in general written English. Roland, Dick & Elman (2007) do offer some insight, placing the percentage of passive verbs out of total verb phrases at 9% for the (Written) British National Corpus, 3% in its spoken equivalent, 11% in the Brown corpus, 2% in the Switchboard corpus, and 9% in the Wall Street Journal Treebank. Consequently, written English with a percentage of passivisation significantly above 10% may read as considerably more passive than most text; for spoken language, this threshold is likely substantially lower.

Syntactic depth

Pearson (1974) examines how syntactic complexity can manifest, comparing the perspective that sentences with more embeddings are more complex ('deep structure') with the perspective that deeper embeddings mean stronger semantic wholes and less need for inference on the reader's part ('chunking'). Although the results of Pearson's experiments disfavour the 'deep structure' perspective, Beaman (1984, p. 45) posits that 'it has generally been accepted that syntactic complexity in language is related to the number, type and depth of embedding in a text.' Dell'Orletta et al. (2014) integrate syntactic depth (as parse tree depth) as a feature for readability prediction, and Collins-Thomson (2014) does the same, consistently finding it in the best-performing models. We will also operationalise syntactic depth as parse tree depth. Dell'Orletta et al. (2014, p. 167) present two similar approaches to quantifying it:

1. The depth of the whole parse tree, calculated in terms of the longest path from the root of the dependency tree to some leaf.
2. The average depth of embedded complement 'chains' governed by a nominal head and including either prepositional complements or nominal and adjectival modifiers.

We will favour the first approach, taking the whole length of the syntactic tree underlying the sentence. We define parse trees as the tree diagrams that the CoreNLP parser, which our analysis uses, generates when it analyses a sentence. These diagrams indicate which elements of the sentence depend on others, and how they interrelate. Figure 1 offers an example of such a parse tree for the sentence 'in fact, much of the water used in production is of such good quality that we have official approval from relevant authorities to discharge it directly into rivers' (Infineon 2013). The parser this example outputs trees

very similar the one present in the CoreNLP package (Manning et al. 2014), but may not be identical.⁸ At its deepest, as indicated by the brackets, the parse tree is 12 levels deep.

Figure 1 Visualisation of a sentence as parsed by the Stanford CoreNLP group (2016) online parser.

Parse

```
(ROOT
  (S
    (PP (IN In)
      (NP (NN fact))))
    (, ,)
    (NP
      (NP (RB much)
        (PP (IN of)
          (NP
            (NP (DT the) (NN water))
            (VP (VEN used)
              (PP (IN in)
                (NP (NN production)))))))
        (VP (VBZ is)
          (PP (IN of)
            (NP (JJ such) (JJ good) (NN quality)))
          (SBAR (IN that)
            (S
              (NP (PRP we))
              (VP (VBP have)
                (NP (JJ official) (NN approval))
                (PP (IN from)
                  (NP (JJ relevant) (NNS authorities))))
              (S
                (VP (TO to)
                  (VP (VB discharge)
                    (NP (PRP it))
                    (ADVP (RB directly))
                    (PP (IN into)
                      (NP (NNS rivers))))))))))))))
```

Subordination

The above figure can also illustrate the concept of subordination. Beaman defines it as ‘the asymmetrical relationship between an independent and dependent clause(s) [...] introduced by an over subordinating conjunction’. As such, subordination is a type of embedding but not every embedding is subordination. In the tree above, ‘SBAR’ (preceding ‘IN that’) denotes elements that introduce a dependent clause, e.g. subordinating conjunctions or relative pronouns. Beaman (1984, p. 45) argues that

⁸ As the parser inside of the CoreNLP package and its implementation on the website keep evolving, it is very difficult to ascertain whether these were identical versions.

‘syntactically complex authors [...] use longer sentences and more subordinate clauses that reveal more complex structural relationships.’ Beaman’s (1984, p. 80) own research also directly uses subordination as a syntactic complexity measure, but finds that “the evaluation of syntactic complexity is more complex than [subordination implying complexity]”.

As with syntactic depth, extent of subordination might correlate with difficult reading material more than it causes difficult reading material. Analogous with Pearson’s (1974) argument about syntactic depth, using less subordination may make a text’s form more straightforward to process, but shift difficulty to the reader needing to infer more semantic relations. These caveats in mind, we also acknowledge that recent NLP-based studies into readability such as Dell’Orletta et al. (2014, p. 167) included subordination-based features as an ‘index of structural complexity in language’. Given the dissenting perspectives on how subordination interrelates with textual complexity⁹, we include the feature in our own study to observe how it varies within the corpus, without forgetting the caution that the relationship between subordination and linguistic complexity is less than straightforward.

Lexical density

Another less than straightforward text-internal variable is lexical density. The issue is, again, how it relates to complexity, but that relationship is not so much controversial as just more difficult to describe. We will investigate that relationship, but will first explore Halliday’s (1987) definition of lexical density:

Lexical density is the proportion of lexical items (content words) to the total discourse. It can be measured in various ways: the ratio of lexical items either to the total running words or to some higher grammatical unit, most obviously the clause; with or without waiting for relative frequency (in the language) of the lexical items themselves¹⁰. (quoted in Castello 2008, p. 49)

In other words, language that devotes relatively more words to content words, rather than the relationships between that content (expressed by function words), is more lexically dense. What complicates the connection between lexical density and complexity is that lexical density manifests differently between different modes of communication:

⁹ Pitler & Nenkova (2008, p. 190), for instance, find a very weakly, though not significantly, positive association between human-assessed readability and use of subordination. They observe that ‘while for children or less educated adults these constructions might pose difficulties, they were favoured by [their educated adult] assessors’.

¹⁰ Section 2.4.2.1 explored the notion of considering the frequency of a particular lexical item relative to its larger linguistic concept when exploring ‘termhood’ as a more computationally intensive method of analysing a given text’s lexical composition.

Castello (2008) compiles various attestations that the lexical density of written language tends to be over 40%, while spoken language generally has a lexical density under 40%. Flowerdew (2012, p. 29) echoes Halliday's (1989, p. 63) notion that '[t]he complexity of written language is lexical, while that of spoken language is grammatical.' Johansson (2008) reinforces Castello's thesis, finding that adults produce narrative and expository written texts at a lexical density of 39%, while children – here representing less experienced language users – produce similar texts at a 32-33% lexical density. Based on these outcomes, it appears linguistic proficiency may correlate with ability to process lexical density, although we must also acknowledge that the set of lexical items a user of a given language can produce will almost inevitably be smaller than that they can (correctly) interpret. This is the difference between a language user's active and passive command of the language, respectively.

Consistent with the aforementioned findings, Gibson (1993, p. 357) finds that academic abstracts with lower lexical density appear to be more 'reader-friendly' than those with high levels of lexical density. While there is some logic to an initial intuition that texts with a higher lexical density are informationally richer and thus easier to understand as they may offer greater specificity and require less inference on the reader's part, Castello (2008) compiles further studies with results that align with Gibson's findings. As such, we see sufficient justification to use lexical density as a reading difficulty measure within the corpus, which consists entirely of written text. We do note that its relationship with reading difficulty may not be entirely linear. For instance, texts with a lexical density lower than 40% may violate expectations that readers that have of written language. Informational density may turn into informational sparsity, and the diminishing returns of increasingly low lexical density may into adverse effects by actually decreasing readability. Chapter 6 explores reader response to either extreme of lexical density.

2.4.3 The Audience

We do not intend to minimise the position of the audience in examining readability issues; a text's readability is only relevant insofar as there is an audience to be readable for. Similar to how many readability formulae conceive of it, we see readability primarily as a barrier to entry, which audiences must overcome in order to obtain the information they need from the text. High readability reduces the effort the audience must expend towards those goals; in some cases, such as a novice user of English trying to read *Ulysses*, the required effort would be so staggering as to make the task effectively impossible. When the cost of expending that effort would be so great as to no longer outweigh the benefits a reader can gain from engaging with the text, we might well expect them to give up altogether. Following this reasoning, the 'obfuscation hypothesis' (e.g. Courtis 1998; see section 3.2.4) would see writers complicating undesirable information that they must

nevertheless communicate to the point where not understanding what is written is less costly than (and therefore preferable to) expending the effort required to understand.

Peterson et al. (2000) find that subject familiarity, linguistic proficiency and motivation will all influence the reader's performance along with a host of other factors, some of which may be as variable as the reader's health, mental state or level of fatigue¹¹. These factors already illustrate why computationally estimating readability (adhering to a text-internal approach) is considerably more feasible than computing understandability along the cloze procedure's logic. Even if we could compile a model of all text-external factors and how they influence reading ease, these highly volatile factors could still affect how easily any individual processes the text.

2.4.3.1 Measuring Understandability: the Cloze Procedure

In spite of the above, we do briefly wish to acknowledge those approaches that consider the audience an inextricable part of a text's readability, such as the cloze procedure introduced by Wilson Taylor in 1953 (DuBay 2004). In a cloze test, the party administering the test deletes words from the test (often along a set pattern) in order to measure how accurately the reader can fill in the gaps. We would assert that, in the first place, this procedure measures the individual's understanding of the text (which, as described above, readability will influence) and is therefore well suited to measuring individual variation within a relatively homogeneous group such as language learners of a comparable level. As DuBay (2004) indicates, this approach better complements reading tests than it does readability formulae, although it does reinforce the notion of a 'threshold' of readability (section 2.4.3.4) in estimating the 'frustration level' in dealing with a text to be below 35% accurately predicted words in a free choice test and below 50% in a multiple-choice test (Bormuth 1969).

It certainly also merits mention that the cloze procedure has offered valuable insight into which qualities or traits enable greater understanding in a given text's audience. The same factors that enable a reader that performs better on a cloze test ensure that they can better understand the text, although the dynamics might differ slightly. For instance, in a cloze test, the tension between active and passive vocabulary persists even with non-domain or topic-specific words: it is not enough for the reader to derive the meaning of a missing word from the context; they must be able to produce the appropriate word. Very similarly, in terms of subject familiarity we can also note that especially in an 'open' testing scenario (without multiple choice), the cloze procedure requires more than a

¹¹ Precisely due to these factors' considerable variability, we might expect them to add a great deal of randomness to an individual reader's perceptions of, and ability to deal with linguistic complexity. This further incents a combination of deterministic, computer-aided techniques and human readability assessment with multiple assessors.

passive knowledge or understanding of the relevant topic: the reader must be able to actively produce the correct words rather than simply recognise them and react to them. This, too, illustrates the exponential increase in complexity that directly factoring the reader into readability assessment (or, more precisely, understandability assessment) can cause, and why we opt for an audience-agnostic definition of readability.

2.4.3.2 Subject Familiarity

A reader's degree of familiarity with the topic of a given text is paramount in determining how much of the information missing from a text they can infer from what is available to them. If the complete text contains no information – or, to anticipate the factor of linguistic proficiency, words – that would be new to them, they can rely on their own internal representation of what the text means to convey in order to fill in the gaps in understanding with a greater degree of certainty. The further removed the topic of the text is from their own experience, the fewer inferences they will be able to make between what they know and what information remains after eliminating words based on the cloze procedure. That is, a reader might be perfectly able to understand and produce every word in a text, but fail to understand how they combine and what that combination means in context. While many readability heuristics are concerned with the average length of words, longer words are not always more specialised or domain-specific: 'company', for instance, registers as a 'complex word' to the Gunning Fog Index, but seems unlikely to meaningfully impede most readers' progress.

2.4.3.3 Language Proficiency

A more linguistically proficient reader will also, regardless of how familiar they are with the subject matter, be better able to decode the various relationships that a text expresses. Readers can benefit from a wider general vocabulary (as texts will contain both general and topic-specific words) and a better understanding of the syntax that expresses the relationships between the concepts that a text contains. This ties back into text-internal readability: the simpler the language a text expresses itself in, the more readers across the strata of linguistic proficiency will be able to extract information from the text. Additionally, general linguistic proficiency does not replace familiarity with the topic: a highly proficient language user may well be able to understand a given word, but without topic familiarity does not necessarily know its specific meaning or connotation within the topic-specific context. For instance, a reader might know the word 'material' can refer both to matter or the quality of being relevant to a given topic, yet be unable to fully decode the nuances of 'materiality' in a corporate reporting context, for instance because they are unsure which of those two potential meanings 'materiality' draws on.

2.4.3.4 Motivation

A reader's motivation (Peterson et al. 2000), finally, can influence how much effort, attention or other resources they are willing to invest in achieving their goals with the text. An intrinsically motivated reader, such as a highly interested one, may go to greater lengths to fully understand the text than an extrinsically motivated one (such as one assigned to read a text). A more motivated reader might experience less frustration, while a less motivated reader may be quicker to decide that the benefits of reading the text do not outweigh the costs. The more motivated reader is more likely to engage with the text for longer, or more deeply, and be less impeded by frustrating elements. Consequently, such a reader will probably perform better. As with the preceding two criteria, motivation connects to text-internal variables of interest to computer-assisted readability prediction – albeit somewhat more tenuously. Style guides as diverse as the SEC's Plain English Guidelines (1998) and the less businesslike Plain Language Action and Information Network's (2017) Dash's Writing Tips indicate that the active voice generally makes for more compelling or interesting reading than the passive. A more interested reader will find the activity of reading itself more rewarding, and accordingly be more intrinsically motivated to continue reading.

2.4.3.5 Difficulties in Quantifying

The fairly tentative exploration above illustrates how deeper insight into a reader's motivation – like many other aspects of understandability – requires more empathy on the assessor's part than a formula or other automated heuristic is currently capable of. The same applies to topic familiarity and linguistic proficiency: formulae, amongst other techniques, attempt to express appropriateness for a given level of education (as a proxy for linguistic proficiency or likely familiarity with the topic), but do not integrate it into the calculation; their approach is audience-agnostic. Similarly, we consider the above factors crucial to understandability (which pertains to the reading experience; the interaction between text and reader) to a given reader or audience. We consider them less relevant, due to difficulty of implementation and the desire for an audience-agnostic approach, to our working definition of readability (which pertains to intrinsic qualities of a text's language that make it easier or harder for the reader to achieve their goals).

In summary, this study adheres to an audience or readership-agnostic approach typified by readability formulae. This approach seems better equipped to help improve a text's general readability by focusing on those factors likely to impact reading experience across all potential strata of readership (such as lexicosyntactic complexity). It does so at the expense of affording the audience fairly little attention. However, as we have previously explored, tailoring the text to a specific audience's reading experience can detract from others'. In minimising the role of audience-specific means of analysis, this study can take multiple audiences' perspectives (for example that of expert analysts vs.

corporate stakeholders) throughout, rather than restricting itself to a single perspective. However, the appropriateness of non-financial disclosures' linguistic difficulty remains an important question throughout.

2.4.4 The Author(s)

While it is possible to approach readability as a text-internal quality in order not to restrict the analysis to any particular audience group, isolating the author from the text is altogether less feasible. On the most basic level, two individuals expressing the same information will do so in different words arranged in a different fashion. As a text-internal approach to readability implies that readability will vary based on which words the author uses in which sequence, that author will have a far more inextricable, if still indirect, influence on how readable a text is. This is not a direct influence because, while the author creates the text, and different properties of and circumstances surrounding different potential authors will yield different potential texts, in most cases that creative process is no longer ongoing by the time the audience starts interacting with a text.

Changing a text's author after it is written, in as much as such a thing is possible, does not alter the text as an artefact, and the exact same sentence written by James Joyce or a sustainability committee would be just as readable if it were written by Dr. Seuss, or a novice language user. A reader's *belief* about who wrote the text might alter their reading experience, in that they might experience difficulty differently based on their expectations. Burgoon & Miller (1985, quoted in Dillard & Pfau 2002, p. 124), for instance, posit that an audience's expectations can alter the communicative experience and that

[h]ighly credible communicators have the freedom (wide bandwidth) to select varied language strategies and compliance-gaining techniques in developing persuasive messages, while low-credible communicators must conform to more limited language options if they wish to be effective.

Within the genre of corporate reporting, we might expect this to imply that the (perceived) credibility and status a report's author possesses enables them to take linguistic liberties (e.g. writing complex language) that another author might not be able to, which might cause a feedback loop when non-financial disclosure language resembles that of financial disclosures. In doing so, it may appeal to the same (perceived) credibility of a financial disclosure by using the same language. Furthermore, based on this proposition, we might interpret the prefacing of corporate reports with LtSs, often written by one of the highest-status members of the company, as an attempt to generate credibility and linguistic affordances for the rest of the report.

In spite of the above, a change in authorship does not directly alter the text, and consequently does not directly alter its readability; if two authors, for whichever reason,

produce the exact same text, they will be equally readable (within the scope of this study's definition), even though they might not be perceived as such based on the audience's expectations. As we proceed to explore the potential influence an author has on the reading process, we note that corporate reporting, as a genre, will likely be somewhat less sensitive to individual authors' influences as a report typically has more than one author.

2.4.4.1 Language proficiency

We have already explored how greater linguistic proficiency can help audiences better understand a text: as the cloze procedure exemplifies, a more experienced language user is less likely to find a given text frustrating to deal with. While a particular text may be difficult and frustrating to deal with, even for an experienced language user, all other things being equal, we can expect it to be even more so to a less experienced reader. This association is less straightforward to conceptualise concerning the author: while a less proficient author is less likely to use difficult or uncommon words and structures, as they will be less familiar with them, they will also be less capable of reflecting on which synonym or structures expresses a particular idea most efficiently or accessibly.

Crossley et al. (2011), for instance, illustrate these apparent contradictions by initially asserting that attention on text cohesion and syntactic structuring only comes during later stages of writing development, but find that while "cohesive devices are important indicators of writing development [...] fewer cohesive devices are the mark of more mature writers" (p. 304). They explain this phenomenon by indicating that more advanced writers use more implicit cohesive elements by linking various elements of the text through syntactic patterning rather than explicit cohesion markers, while less proficient writers introduce more cohesive markers. Similarly, higher-proficiency writers make less common but more specific lexical choices, which are likely to lower reading ease for less proficient readers but improve it for more proficient readers.

Based on these outcomes, it appears that a broader command of the language implies more options to choose from, for better or worse. As Tierney, Anders & Nichols Mitchell (2013, p. 39) phrase it, "there is a trade-off between the effort exerted by a writer and that demanded of a reader." A better command of the language does not remove that trade-off. Or, as the variously attributable (O'Toole 2014) saying goes, "easy reading is damned hard writing."

Less proficient users may also use less accurate language and commit more errors. If we conceive of linguistic errors as a form of noise in the communicative channel (Kernighan et al. 1990, Church and Gale 1991, and Mays, Damerau & Mercer 1991), we can straightforwardly associate more errors with higher reading difficulty, as this noise would be a text-internal element that impedes a reader's ability to extract what they want from the text. All other things being equal, a less proficient author is more likely to introduce errors that lower readability. We note, however, that linguistic errors will be

rare in many kinds of corporate communication, especially corporate reporting, given the (potentially multi-stage) review and editing processes many such documents go through. That makes linguistic precision a less important concern for reading difficulty, as far as this study is concerned. Linguistic errors do occur, especially when the reporting company does not natively operate in the language they report in (typically English), but are less relevant to this study than others that deal with reading difficulty.

2.4.4.2 Authorial (corporate) voice and language variety

We can extrapolate the effects of the corporate editing process to other textual aspects. While their influence on a text's readability is undeniable, many of the above factors will likely have a more modest impact on the language used in corporate reporting, and, consequently, this study. Many components of these reports are not the result of individual effort; rather, the composition process is collaborative throughout its various stages, mitigating the effects of traits particular to any one individual. If a document goes through several steps of drafting, editing and review involving various departments, any individual author's voice will tend to normalise towards a shared 'company voice'. This 'design by committee' is likely to reduce individual stylistic variation, and will (assuming a well-functioning drafting, editing and review process) improve accuracy and minimise errors. As this makes it simultaneously more difficult and less relevant to account for individual differences between those involved in the composition process, we focus on those text-internal criteria particular to the resulting 'corporate voice'. Those include frequency of passive structures (which SEC 1998 cautions against), as well as a component of the authoring process likely to be fairly uniform amongst the various authors that comprise the corporate voice, and thereby likely to survive the editing process: the language variety it uses.

While not every collaborating author will prefer the same variety of English (some might be expatriates, others might work in different parts of the world, etc.) or even employ it as a native language, most texts with a functional editing and review process will consistently strive towards a given variety of English that aligns with the 'company voice'. We examine language variety's influence on readability based on previous findings that both syntax and lexicon show meaningful differences between varieties of English: for instance, Precht (2003a, b) finds US English to be more active and less modal than British English. If text-internal features relevant to readability, such as the aforementioned use of passive structures, can vary between varieties of English, this gives us cause to address the gap in research on readability between language varieties, and investigate how the latter can influence or predict the former.

Chapter 3 offers a more detailed overview of readability inquiries into corporate reporting, as well as expanding on language variety and how it affects text.

2.4.5 Paratext

In part because our analysis techniques are far better suited to it, we primarily examine text-internal readability issues rather than text-external legibility issues. That is not to say the paratextual aspects of corporate communication do not merit examination, such as for example Cho et al. (2012a, b) illustrate examining visual distortion in graphics accompanying corporate sustainability reports. They find that some corporate reports also use paratext (especially graphs and figures) to obfuscate or distort (deliberately or otherwise) numerical or textual elements. Section 3.2.4 expands on this practice and the visual aspects of corporate reporting,

We also note that our working definition does not necessarily include paratextual aspects such as typography or supporting visuals; our means of analysis isolates textual content from aspects of presentation (c.f. section 3.3.6); and DuBay (2004) considers such presentational issues as the font a text is composed in to belong to the domain of ‘legibility’ rather than readability. Like DuBay, we will define legibility as the visual or presentational aspects that make a text easier to decode, and thus distinct from readability: the same sequence of words could be more or less legible if written in a different font, but its readability would not change. We must note, however, that paratext may improve understandability when visual aids or typography succeed in facilitating information processing; emphasising key words in bold, for instance, may help readers better decode the text. As section 2.5 will explore, the inability to capture such aspects is one of the primary drawbacks of this study’s approach to quantifying readability, but an inevitable one given the size of the corpus.

Notably, neither readability formulae, nor the more advanced Natural Language Processing (NLP) methods that subsequent sections will explore in greater detail, typically incorporate paratext. We can explain this at least in part as a technical restriction: computers deal with text differently than humans do; otherwise there would be no need for such heuristics as readability formulae or machine learning to approximate the reading experience for humans. On a very basic level, these heuristics treat text as a linear sequence of characters. Introducing elements that do not fit within this linear pattern – such as a text box next to a paragraph, or an explanatory graph or image – impedes the tools’ ability to accurately process the input. While more advanced techniques may be able to glean more information from the same textual material, and thus offer a more nuanced assessment, font, graphs, and other paratextual elements (must) remain invisible to these techniques. These additional elements modify, enhance, supplement, or in some cases even obfuscate or work against that linear stream of characters - hence ‘para’-text. From a readability, legibility, or understandability perspective, paratext is most meaningful in how it interact with the text proper. Decoding that interaction requires interpretation, judgment, and awareness of the context, and is thus considerably more challenging to automate from a technical perspective.

2.5 How machines process text and what that means for readability studies

The difficulties that computers face in capturing paratext illustrate how humans ‘read’ differently from machines. We have already touched upon how the font in which a text is written influences its legibility (for humans). Text printed in a cursive, heavily stylised or otherwise unusual font may slow down the reader’s ability to recognise each individual character compared to fonts that are more common. Even amongst those, preference for serifed and sans-serif fonts can vary based on a number of factors. These factors include document type (lengthy print vs. short copy), position in the document structure (e.g. headings vs. body text), audience (adult readers vs. younger readers), medium (print vs. electronic), etc., but no firm consensus exists on which use case is optimal for which type of font (Strizver 2017). Preference for one font over the other in terms of legibility¹² may also simply stem from greater familiarity enabling faster processing (Strizver 2017); variations in use of font within a text may similarly enhance or disrupt reading flow.

However, use of a particular font over any other does not impact the processing fluency of computer-based or -assisted readability analysis (although it may impact some of the preparatory steps such as Optical Character Recognition, commonly called OCR). Moreover, the plain text format typically required for Natural Language Processing (Unicode Transformation Format – 8 bit or simply UTF-8) does not support text formatting (hence ‘plain text’). In other words, most of these readability heuristics simply have no way of knowing in which font the text exists; they cannot ‘perceive’ formatted characters as human readers do.

Even if a readability heuristic could access this information, how could they meaningfully use it to more accurately assess a text’s readability or understandability? They would need to weigh and numerically incorporate the impact of the relevant font – which there exists little consensus on – considering the above factors such as which genre the text belongs to, who is likely to be reading the text and how, along with a host of other contextual information, to achieve a meaningful result. As this is currently very difficult, if not impossible, even for highly advanced methods, the logical choice is, again, for

¹² We acknowledge that in discussions of fonts, ‘legibility’ and ‘readability’ can have their own meanings, different from how we have defined the terms. For instance, Haley (2017) describes legibility as “a function of typeface design [...] an informal measure of how easy it is to distinguish one letter from another in a particular typeface.” Readability, conversely “is a gauge of how easily words, phrases and blocks of copy can be read.” In this context, legibility is a purely font-internal property, while readability also depends on layout decisions. The contrast with our own approach to readability as a concept reinforces how our approach to readability is simply a working definition, and illustrates how context-specific the situational use of such relatively generic terms can be.

computer-driven or -assisted readability research to separate text-internal readability (which it can estimate through processing the text itself) from the presentational issues that DuBay (2004) refers to with ‘legibility’.

The same need to understand the context as well as the text impedes these automated readability analysis heuristics’ ability to engage meaningfully with other paratext such as pictures and graphs. In an ideal scenario, such illustrations help understand or reinforce the text, and vice versa; if a reader is unsure of how to interpret the one, they can fall back on the other (Beattie & Jones 2000). We see this in illustrated children’s books that visually represent a story’s most important moments just as well as in corporate reporting, where a graphic or tabular representation of an earnings report can provide an at-a-glance summary of the narrative, while the narrative ties together what the reader is seeing.

As the above illustrates, automatic readability analysis, be it formula-based or more in-depth, is above all an amalgam of heuristics meant to approximate how difficult a human reader might find a text. As the preceding sections have already explored how readability formulae function and how we can expand on them by capturing syntax, the following section will offer a general description of more elaborate (often NLP-driven) techniques we will use to expand on formula-based readability assessment as well as how previous studies have used those techniques. The separate chapters pertaining to every technique will contain a ‘Methodology’ section detailing their technical implementation.

2.6 Readability beyond formulae

As Flesch warns, we cannot restrict ourselves to readability formulae alone, even if we are only aiming to explore text-internal readability. Even when we do not consider the audience as determining a text’s readability, but rather as only interacting with it, their relative shallowness and simplicity suggests that readability formulae cannot achieve everything we might want out of a readability metric. Their relatively high computability comes at the expense of specificity and completeness. The typical readability formula does not explain *why* a text is difficult to read, and only considers the most surface-level aspects of readability, based on the assumption that those surface-level aspects will correlate meaningfully with deeper-level ones. While there are studies (e.g. Pearson 1974) in support of that assumption, formulae are not the be-all and end-all of readability analysis. This section considers readability analysis approaches other than formulae, exploring their merits and weaknesses.

Throughout the advance of computer technology, the traditional formulae have remained the most easily computable means of estimating a text’s readability, having

evolved from a count-by-hand process to such a degree of automation that popular text processors such as Microsoft Word automatically compute it as a software feature. However, most lack the specificity to meaningfully explain their result beyond the text containing long words or sentences, and even the relative contribution of each of those factors is typically opaque. This requirement for computability impacts the writing process.

The inverse – the most complete and specific readability analysis process with, accordingly, the lowest computability – is editing or review by human intelligence. Modern computer technology can quantify, weigh and combine vastly more aspects of a text than might have been automatable at the inception of readability formulae. However, considering all the possible aspects of how fluidly a text will interact with an audience requires empathy, subjectivity and even a measure of creativity – some of the most difficult cognitive processes to automate or replicate through computation.

As DuBay (2004, p. 31) paraphrases the cognitive perspective on reading, “the reader constructs meaning by making inferences and interpretations [and] using metacognition, the ability to think about and control the learning process.” In other words, a reviewer or editor assessing a text’s readability must assess their own position and how they interact with the text. They must also imagine how other readers might react to aspects of the text, and even entertain what is not there, imagining how else a text might look and how accessible it might be to the audience when altering a phrase, word, structure, etc.

2.6.1 Human Review and Readability Heuristics

Based on the above, we can assert that readability formulae and human review are two extremes of a continuum, with a great number of readability heuristics between them. On the one extreme, an editor, proofreader or author can deal with a text in a truly holistic fashion. They can offer specificity in which aspects they perceive to influence a text’s readability, what their influence is and how important they are, and are not limited to any predetermined or pre-programmed set of features in doing so. They can be as complete as is necessary or appropriate in doing so.

The primary drawbacks of a holistic editing process performed by a human editor are its low computability and low speed. Using readability formulae, a modern personal computer can analyse several complete texts before an editor is finished with a first sentence. Its advantages make manual review a superior means for qualitative analysis of a single or a few texts, but these disadvantages make it all but unsuitable for quantitative analysis of a whole corpus.¹³

¹³ This is with the exception of distributed approaches such as crowdsourcing (which can generate their own issues, such as inter-rater variability).

Another drawback is quantifiability: most people would have little difficulty comparing one text's readability to another's (see e.g. Tanaka-Ishii, Tezuka & Terada 2010), and will likely be able to justify their decision (as De Clercq 2015 and section 6.3 can provide examples of). However, they are less likely to be able to assign them an arbitrary number or score, especially without a frame of reference. Conversely, most readability formulae produce exactly such a number, which is far less content-dependent by design. Alternatively, we might argue that a readability formula uses the entire language – or the hypothetical set of all possible texts – as a point of reference.

As we have addressed, human review and assessment of texts draws on facets of cognition that are currently very difficult for computers to replicate. Yet computer-assisted or – driven editing and proofing continues to become more commonplace, and more technically advanced, with the most recognisable example probably Microsoft's spelling and grammar checker, which is now a component of the Windows operating system. These tools, however, supplement rather than replace 'conventional' editing for any type of copy that needs thorough revision; their design is typically primarily rule-based, so they are less able to consider the text as a whole or adapt to unfamiliar problems in the way that an editor might (Microsoft 2018). Such a system typically evolves through the addition of more rules and features (such as Office 2007's inclusion of 'contextual spelling', which helps differentiate between 'no' and 'know', 'than' or 'then', 'complement' or 'compliment', etc. based on where they occur; MSDN Archive 2006). These tools can help address the most frequent or common errors (although human judgment still applies, for example in dealing with false positives) so the human intelligences involved in the process can devote more attention to resolving those issues the automatic systems cannot. Especially regarding syntax, such tools can often indicate problems with the text's readability or ease of processing (for example long, complex or passive sentences) but are less able to automatically suggest alternatives than they are on a word level.

While human assessment is difficult to replicate, spelling and grammar tools are built upon heuristics that simulate various aspects of human assessment. As computational power and techniques continue to advance more rapidly, ease of computation has become less of an obstacle for quantitative readability analysis. For readability analysis, it remains a pipe dream to have every single data point describing a large corpus be the result of human assessment. Especially given the ever-increasing size that a corpus can be, sufficient examples of human assessment can allow computers to approximate it through various Natural Language Processing (NLP) techniques, and even generalise towards a more holistic perspective through a process called machine learning. The next section offers an overview of NLP, chiefly as it pertains to readability assessment, while the end of the chapter anticipates machine learning (which requires both human assessment and NLP to enable it), as a further elaboration on readability assessment to explore deeper

into the study. Chapter 6 presents the outcomes of an exploratory implementation of machine learning based on this corpus.

2.6.2 Natural Language Processing

Without wishing to delve too deeply into semantics, we must first acknowledge that all natural language invariably requires processing to achieve any (linguistic) effect. Any receptive process (reading or listening) taking place within a natural language¹⁴ chiefly consists of the reader or listener processing that language. Language production (i.e. writing or speaking) arguably draws on that same ability. That is, while Natural Language Processing or NLP is a subdomain of computational linguistics that this study draws on in order to advance readability analysis within the field of business communication, natural language processing is not a skill exclusive to computers. When this study refers to NLP, it will invariably be in the computational sense of a broad set of heuristics and other techniques meant to make computers approximate human processing of language (Jurafsky & Martin 2014).

One of the crucial differences between Natural Language Processing and computers' abilities to process non-natural languages, such as programming languages, lies in the extent of ambiguity that they allow for. Generally speaking, a programming language requires that there is only one way to parse or interpret a piece of code in order to compile it (i.e. make it possible to run the code), and they satisfy that requirement by making it impossible for well-formed code to be ambiguous. Natural languages, belonging first and foremost to human communication rather being a set of instructions for computer systems, have no such restrictions. Resolving ambiguity becomes a core component of processing (Jurafsky & Martin 2014). Disambiguation frequently relies on the context of an utterance; 'I saw John with the pair of binoculars yesterday' can imply that the speaker saw John through a pair of binoculars the day before, or that they saw John, and John had a pair of binoculars. While the speaker might disambiguate the utterance by outright stating 'I saw John through the pair of binoculars yesterday', the utterance is equally viable and well-formed when they choose to use the more ambiguous 'with'.

Ultimately, the aim of many NLP techniques is to interpret language well enough to achieve a specific linguistic goal (Jurafsky & Martin 2014), which can range from gauging the difficulty of a piece of text, as will be the case for this study, to detecting errors and suboptimal language use (and potentially suggesting improvements). The latter might be needed for a word processor spell checker, but has also been used for quality estimation in (machine) translation tasks (see e.g. Tezcan 2018). Where humans tend to rely on

¹⁴ We can, for these purposes, define a natural language as any language with two or more native speakers.

context and world knowledge (Jurafsky & Martin 2014) in order to carry out such tasks, NLP approximates that awareness and knowledge through a series of scaffolding heuristics. Achieving goals through NLP almost invariably entails completing a number of sub-tasks, such as parsing the sentence into structural trees, which requires awareness of each word's part of speech, which dovetails with the NLP system's ability to disambiguate between different potential meanings of that word (Jurafsky & Martin 2014). There are typically numerous ways to resolve any of these (sub-)tasks, and many of them continue to see steadily improving performance as techniques advance.

As we have previously indicated, readability assessment typically sees trade-offs between completeness, specificity and computability, which means that the more in-depth analysis NLP techniques offer are significantly more computationally complex than those required to calculate the readability formulae.

2.6.2.1 An Example: Word-Sense Disambiguation

To illustrate more advanced NLP techniques' greater complexity compared to calculating readability formulae, the problem of word-sense disambiguation (WSD) offers an intuitive illustration of the challenges that NLP techniques often face. An iconic example is the task of differentiating between the various senses of the word 'bank': is it a financial institution, the side of a river, or does it refer to slanted movement? As a human might, a word-sense disambiguation system will often look at the context in which the word occurs.¹⁵ In the case of 'bank' for instance, both humans and machines will typically be able to decode the ambiguities in the following sentences based on their context:

- (4) I took the money to the bank.
- (5) Sitting on the bank, we watched the boats go by.

The conventional interpretations for these two senses of the word 'bank' would, respectively, be 'financial institution' and 'side of the river' as 'money' co-occurs with the 'financial institution' sense of the word and 'boats' can co-occur with 'river', which then leads to the 'side of the river' interpretation of the word. While in neither case the other interpretation would be ungrammatical or impossible, they would present far less common scenarios, such as a riverside drop-off from a crime story or a financial institution offering a rooftop view of a harbour.

¹⁵ Although virtually all WSD techniques will draw on a word's context, this is not the only option; Navigli (2009), for instance, describes a system that always chooses the most frequent sense of a given word as a fall-back option for WSD tasks.

2.6.2.2 NLP as a Tool

This dissertation does not attempt to advance or even thoroughly explore the current state of the art of NLP techniques from a technical level; as problems such as disambiguating ‘bank’ are subtasks of subtasks required for a sentence-level analysis that grow in complexity for every word in the sentence, NLP is vastly more technically complex than formula-based readability assessment is. What is more, while the temporal and computational requirements of readability formulae scale approximately linearly with the number of words in a text, more advanced NLP tasks such as coreference resolution (see section 2.4.2.2) can scale exponentially if every subsequent word requires examination for coreferential connectors with every preceding one. In other words, in terms of technical complexity, a detailed overview of NLP techniques lies both beyond the scope of this study and the technical expertise required to carry the study out (we refer, for those purposes, to Jurafsky & Martin 2014).

Instead, this study attempts to advance the state of the art of corporate readability in terms of technical approaches to the genre; for these purposes, we use the readymade CoreNLP toolkit, which its authors present as

“[a]n integrated NLP toolkit with a broad range of grammatical analysis tools [...] for arbitrary texts [...] with the overall highest quality text analytics [which aims] to make it very easy to apply a bunch of linguistic analysis tools to a text.” (Stanford NLP Group 2018)

CoreNLP integrates a part-of-speech tagger, named entity recognition system, parser, coreference resolution system, sentiment analysis, pattern learning, and information extraction. As, based on the above, this study treats Natural Language Processing as a (set of) tool(s) rather than an area of research, section 3.3.4 will discuss the most pertinent aspects of this toolkit to the rest of the study in somewhat greater detail when exploring its methodology. For technical detail, however, we refer chiefly to Manning et al.’s (2014) overview of CoreNLP’s various components as an introduction to the various problems NLP tends to encounter and attempt to resolve. Chapter 3 explores the approach to and results of the formula and NLP-based readability assessment of our full corpus. This represents the fully automated component of the study.

2.6.3 Scoring and Machine Learning

Building on fully automated or automatable formulae and Natural Language Processing techniques, we also aimed to develop a more fine-grained, genre-adapted readability assessment system specific to corporate (sustainability) reporting. In terms of genre, the readability formulae exhibit the relative weakness that their scope is the widest possible – all (potential) texts present within the language. Lexico-syntactic features extracted

through NLP, while more nuanced, similarly face the issue that, in as much as there are points of reference, they typically originate from very general corpora, rather than genre-specific ones.

We can expect corporate reporting to be on the less readable end of any spectrum these ‘yardsticks’ can capture. Consequently, we saw another means of measuring readability in first having a group of highly proficient second-language learners assess the relative difficulty of a number of excerpts from sustainability reports. We believed this group could stand in for the educated readers likely to be part of sustainability reports’ wider audience. This enabled us to use machine learning techniques to have a fully automated system detect which measurable qualities of the text best predicted those experienced language users’ scores, thus enabling a fully automated genre-adapted assessment process that would nevertheless be able to approximate the score a member of the audience might assign to it.

While specific implementations vary, on a conceptual level, readability analysis through machine learning begins with texts annotated for readability through human assessment (the ‘gold standard’). It then extracts as much information about these texts as is available and automatically computable (these are called ‘features’ and might include the ‘classic’ word and sentence length variables, but also syntactic analyses, word sequences, relational models of the information in the text, etc.). It then attempts to compute the best (most useful) relationship between that information and the assessors’ judgment. It might find certain elements more informative of that judgment, and will weigh them accordingly. Greater availability of information for every element to be predicted will enable (but not necessarily guarantee) greater accuracy. However, perfect accuracy is virtually impossible: some variation in the outcome variable or assessment will depend on factors not measured (such as how well the paratext supports the text), the assessor, or circumstance (this might include factors with more short-term variation, such as fatigue, as mentioned in section 2.4.3.1). Like readability formulae, this technique estimates rather than sets in stone how readable a given piece of writing is, but places less of an emphasis on computability and can therefore conduct a more complete analysis with greater specificity. For instance, it is possible – and often optimal – to train a machine learning system on a specific genre or corpus, which in the case of readability can enable far greater granularity than the formulae’s ‘one-size-fits-all’ approach.

As this more fine-grained approach works best with a tailored corpus and manually annotated training data, it is significantly more difficult to implement than a formula-based approach. One of the difficulties lies in computing the assessment: formula-based processing can occur in real time, for example in Microsoft’s Office suite, while advanced machine-learning-based language analysis systems such as the Stanford NLP Group’s Sentiment Analysis suite can take several seconds to process a sentence, and analysis typically requires the sequential application of several such tools. However, a much greater obstacle to this approach lies in enabling machines to perform the analysis in the

first place. Obtaining usable textual data is a first hurdle, annotating it is a second, and enabling the machine to process the relevant data a third. Section 3.3 explores this in further detail.

However, the outcomes demonstrate that the result may be worth the effort. In short, the machine learning approach can essentially compile a substantially more advanced (both technically and in what it can consider as variables) readability formula specific to a (sub-)category of texts. It is a heuristic counterpart to the truly holistic approach of human assessment, but offers enough of the benefits of human assessment to integrate it into a modern study into genre-specific readability analysis and investigate its merits as such.

Due to the same factors that complicate its implementation, the main strength of a machine learning approach is that it can (but often must) be tailor-made to a specific category of text. To achieve this, we implement De Clercq's (2015) and De Clercq & Hoste's (2016) infrastructure and methodology (which we expand on in section 6.6) with the aim of expressing a text or text fragment's difficulty within the genre as a number between 0 and 100, with the latter being easier to read. This measure invites comparison with the similarly bounded Flesch score, but that comparison is not necessarily appropriate. As the Flesch score attempts to describe the full range of easiest to most difficult reading without genre adaptation, we should (and will) compare Flesch score results with the output of a machine learner trained on a general corpus. We can expect texts within the single genre of corporate reporting to cluster around one or two strata of difficult on the Flesch scale, with a learning system trained on a general corpus likely showing a similar clustering.

The main aim and advantage of training the system to detect variations in readability within the single genre of corporate reporting is granularity over the other approaches: while we can reasonably expect corporate reports to occupy less than half of the general scales, the genre-specific approach enables a full 0-100 range of resolution in terms of readability variability. We note that while we cannot strictly adhere to the 'readability' extreme of the 'readability-understandability' continuum using these techniques, gathering data from multiple human assessors moves us considerably closer to quantifying readability rather than understandability than using a single assessor would. Using NLP techniques to measure text-internal features, even when they serve as proxies for more audience-dependent aspects of the text (because we attempt to predict the score an assessor would have given based on those text-internal features), further helps us emphasise general readability over audience-specific understandability. Chapter 6 describes the result of this first proof-of-concept foray into machine-learning based readability assessment tailored to corporate (sustainability) reporting.

2.7 Moving Forward

As we have discussed a sizeable host of concepts throughout this chapter, we offer a brief overview of how the following chapters and inquiries therein will deal with those aspects of readability.

This chapter has explored corporate communications, with an emphasis on corporate (sustainability) reporting, and the linguistic features and corporate voice that characterise it, as well as addressing the sometimes problematic ways the genre interacts with its audience. We have examined who the audiences are, and which implications they might have for report composition. What follows in Chapter 3 is based on that initial conceptual exploration. That next chapter will contain an overview of previous studies into the readability of corporate reporting, both financial and nonfinancial, and the hypotheses we will investigate based on both those findings and the concepts we explored throughout this chapter.

We will investigate these hypotheses through a descriptive inquiry into our corpus of corporate communications that will use readability formulae and deeper-level syntactic measures (passives, parse tree depth, subordination and lexical density) to estimate each document's readability with every data point representing a text. We also analyse how these aspects differ as language variety and corporate performance differ. Chapter 3 also elaborates on the relevance of both language variety and performance. An overview of the methodologies we used in compiling and analysing this study's corpus will precede the analysis proper.

Chapter 4 will then examine how variation in the language of corporate reporting impacts reader perception of the text and company. As the present chapter explored, the reader's perception of the text and audience can also greatly impact the reading and understanding process. Companies may wish to exert the greatest possible influence on that reading experience in a process called impression management, which this theoretical overview has begun to explore and Chapter 3 will expand on. This perception study will examine to what extent differences in the variables we have indicated as measures of readability also influence the audience's perception not only of readability, but also of the company that produced the text (e.g. its professionalism or credibility).

Subsequently, Chapter 5 presents the results of an exploratory study into assessing sentiment as it occurs within sustainability reporting, given the greater extent of tension between various areas of favourable or unfavourable news in that a given outcome might, for instance, be financially favourable but at odds with social sustainability. This also allows us to test to what extent the 'Pollyanna Effect' of corporate reporting being highly positive regardless of actual performance applies to sustainability reports.

Finally Chapter 6 will apply a two-pronged approach to the audience: we will gauge how (human) users of these reports assess their reading ease or difficulty, and which

factors they indicate as influential, and then see how we can translate these audience-specific observations into audience-agnostic, text-internal readability features. Based on those outcomes, we will attempt to build a genre-specific readability analysis system tailored to the language of corporate reporting, although that genre-specific quality will come at the expense of some audience-agnosticity.

A discussion of these components' findings and their implications follows this final aspect of the study in Chapter 7 and Chapter 8 respectively, accompanied by an overview of future research avenues in Chapter 9. The conclusion attempts to summarise what all of the above means for sustainability report writing.

To the greatest extent possible, these different areas of inquiry will stand alone as parts of the study. Each will give a synopsis of relevant previous research. Each will then delineate the methodology we employed to carry them out and the hypothesis we tested. Nevertheless they will to some extent build on one another. A notable instance thereof will be that every subsequent aspect draws on (parts of) the corpus collected for the initial fully automated formula and accompanying NLP-based investigation of the genre's readability.

The final aim is to have a thorough overview of the various ways in which we might quantify the language of corporate (sustainability) reporting, what that language says about the genre, and what the implications are for those writing and reading it.

Part 2: The Language of Corporate Reporting

Chapter 3

The Readability of Corporate Reporting

3.1 Motivation

This chapter describes the corpus in terms of its readability, which we will quantify through the readability measures that Chapter 2 describes. We also explore the impact of language variety and industry, in addition to examining how performance and readability measures interrelate.

As sustainability reporting has continued to mature, scholars have subjected it to much of the same scrutiny they have financial reports. They have queried quality and scope of content, rhetoric, and use of visuals, etc. (e.g. Barkemeyer, Comyns, Figge & Napolitano 2014; Cho, Michelon & Patten 2012a, 2012b; Hahn & Kühnen 2013; Parsons & McKenna 2005). Inquiries into the genre's readability, however, have been relatively minimal, with Abu Bakar & Ameer (2011) and Farewell, Fisher & Daily (2014) amongst the few to conduct them. Much like similar inquiries into the readability of financial reporting (e.g., Courtis 1995, 1998; Leheavy, Li & Merkley 2011; Li 2008; Rutherford 2003; Smith & Taffler 1992a), they find that non-financial disclosures tend to occupy the most difficult strata available to readability formulae. Courtis (1998, p. 460), for instance, found a sample of corporate annual reports "possibly beyond the fluent comprehension of the vast majority of readers". In other words, while their similarity in form belies potentially very different content, that similarity may extend to the text's readability characteristics.

Furthermore, relative to that of financial reporting, there are two complicating factors that we consider likely to influence the readability of sustainability reporting: its wider audience compared to financial reporting, and the tension between the act of publishing a sustainability report and the value of its contents in terms of impression management. Furthermore, as the relatively lower expertise of a sustainability reporting audience might render readers more susceptible to other variables in the text overall, we will also examine the impact of language variety and industry on text characteristics. In an

increasingly international reporting space, we might expect readers across different varieties of English to have different expectations about the language a report should use.

In addition to the multiple language varieties present in the corpus and the focus on Natural Language Processing techniques, this study also expands on the aforementioned previous inquiries into the language of sustainability reporting. Whereas Abu Bakar & Ameer (2011) as well as Farewell, Fisher & Daily (2014), use approximately 200-word excerpts from the report (Courtis 1995 similarly uses three 100-word passages per report), this study uses full texts wherever possible. That leads to a substantially larger corpus in terms of tokens than those aforementioned studies, with approximately 2.75 million tokens to, for instance, Abu Bakar & Ameer's (2011) 66000 tokens.

3.2 Context and hypotheses

3.2.1 Genre and Audience

To reiterate the differences in (implied) audiences, companies can safely make a number of assumptions about the readers of financial disclosures, prime amongst which some level of financial sophistication, in addition to shareholders' goodwill towards the company and aligned financial interests. This applies to a lesser extent for strict-sense stakeholders that make up another part of non-financial disclosures' wider audience. These have voluntarily made some investment in the company, such as time and effort or social engagement, as might be the case for an employee. In addition to these aligned interests, a company addressing this group can assume some elementary familiarity with its operations. Few assumptions, not even that of a shared language, need still hold true for the final audience component of broad-sense stakeholders, such as local communities, which imposes substantial consequences on how companies should address them (cf. e.g. Townsend et al. 2010, Jenkins & Yakovleva 2014).

Based on a study of 76 sustainability reports issued by well-performing companies (either based on consistently good performance or improvements during the fiscal year), Farewell, Fisher & Daily (2014) found that none of the reports' text could be labelled as conventionally readable according to the Flesch Reading Ease Index – that is, even the most readable did not score above 50. They indicate that although this level of difficulty is consistent with that of other types of corporate communications, that does not diminish companies' responsibility to use accessible language, especially given the potentially weaker command of the report's language a sustainability report's audience may have. One example of this may be that a British company operating mining sites abroad should not necessarily assume a native-level command of English amongst local communities.

As Li (2008), for example, finds that older firms generally have more readable reports, and Rutherford (2003) finds that greater organisational complexity can yield less readable reports, we might expect that especially those companies with the experience and sophistication to publish (high-quality) sustainability reports are potentially also most at risk of producing inaccessible ones when their size also implies complexity. Although the few studies that have investigated sustainability reports' readability found them to occupy the same readability strata, there is still some room for variation within those strata.

However, as Courtis (1998, p. 459) claims for the annual report we can also expect that sustainability reporting “will be more or less useful depending on the extent to which, *inter alia*, its content is readable and understandable.” Consequently, we expect companies to make some linguistic affordances for sustainability reports’ wider audience of stakeholders by making a deliberate effort to communicate in language the entire potential audience can understand. While achieving a ‘Plain English’ level of readability may be an unrealistic target given the complexity of the subject matter, a company with an interest in transparency should optimise the readability of sustainability reports to maximise their return on the considerable investment of resources that reporting demands. While section 3.2.3 offers alternatives to that transparency perspective, we will formulate a first (and perhaps somewhat naïve) hypothesis for this chapter, which large parts of this study will then proceed to nuance:

H1: Sustainability report content will be more readable than financial report content.

While an ideal scenario for sustainability reporting and its wider audience would likely see its readability around a FRE score of 70, suggesting fairly general readability, what little previous research there is into these documents’ readability already suggests that such an outcome will be highly unlikely. Nevertheless, companies’ often explicit awareness (see Chapter 1) that they are addressing a wider, less expert audience could lead to a significant difference in readability even within the same stratum. That is, there is still some room for variation within, for instance the ‘difficult’ or ‘very’ difficult bands of readability in which previous research places both financial and sustainability content.

3.2.2 Language Variety

Farewell, Fisher & Daily’s reference to readers’ command of the language a report is written in highlights another issue. As section 2.4.3 examined, a more proficient reader will be better able to achieve what they want to with a text, and a reader that does not have, for instance, English as a native language may be at a disadvantage as it is an extremely common reporting language. This is probably because of English’s ubiquitous position as a global business lingua franca. However, precisely because English is so geographically diverse, it is also possible for a report to be composed in an English that is not the reader’s native variety, such as might be the case for a British reader of an American report.

Most scholars have hitherto neglected the impact of language variety on corporate reporting, or – at most – indirectly integrated it as a variable. For instance, Leuz et al. (2003) distinguished between three clusters of declining legal enforcement: the US, the UK and Australia belong to the cluster with the highest enforcement, most European

countries to the second, and Greece, Portugal, Italy and Spain, along with India, to the last, which faces the least legal enforcement. As the study finds that clusters with greater legal enforcement exhibit less earnings management (i.e. optimising the numerical content of financial disclosures, albeit through legal means), we might similarly expect that the countries in the first cluster will exhibit less textual manipulation (which the following sections will expand on), and thus better readability, than those in the clusters with a lesser extent of enforcement. More recently, Cho et al. (2012b), also drawing on Leuz et al.'s framework, find a greater skew towards positive graphs in countries from less-regulated clusters, similarly suggesting manipulation. Language variety is present by proxy in this analysis as we mainly find those countries with English as a sole official language in the first cluster and countries that employ Business English as a Lingua Franca (BELF) in the second and third, the latter of which the more linguistically diverse India also occupies. Section 3.4.1 further details corpus composition.

However, scholars such as Precht (2003b) and Creese (1991) suggest variation between varieties in the same cluster in their application of such syntactic and semantic elements as passivisation, impersonalisation and directness. As corporate reports reach ever-increasingly international audiences (Townsend et al. 2010), we wish to examine how textual complexity, expressed both as a 'shallow' formula and a set of linguistic features, differs across the five varieties present in our corpus. For instance, a British report might contain more passive structures in order to express itself less directly and maintain an (expected) discursive distance from the British reader, but might, in doing so, strike an American reader as evasive. Section 3.4 will explore how, rather than the different clusters of institutional climates that might impact reporting, it is chiefly the different varieties of English that represent different linguistic attitudes and influence report readability. This awareness, too, can help companies write reports that better communicate what they want to communicate, and better enable readers to approach texts with an appropriately critical attitude. Based on the aforementioned studies, we formulate the following hypothesis:

H2: Corporate (sustainability) report readability will vary along language variety (regional) lines.

This study's corpus contains five varieties of English, represented by five regions: Australian English, British English, US English, Indian English, and Business English as a Lingua Franca as used in non-UK Europe.

3.2.3 Company and Legitimacy

While we have briefly touched on how organisational complexity might influence readability, another company-specific feature likely to impact the reporting process is

the industry in which the company operates. Farewell, Fisher & Daily (2014) cite this as an avenue for research, expecting different industries to address the different relevant issues through different structures, as reporting complexity can vary across industries. For instance, many extractive industries have inherently non-renewable production processes, which will typically require addressing in a sustainability report and render the company's environmental performance a sensitive issue. Similarly, companies with a history of worker rights issues, might face additional scrutiny or legislation in these areas, such as is the case with the US semiconductor industry which requires companies that file with the SEC to disclose and describe their use of conflict minerals (Higgins 2014).

Especially for these industries with a greater sensitivity to CSR issues, we might expect considerable tension between what a company can gain from the mere act of reporting on its non-financial performance compared to what they can gain from reporting transparently, if the latter also accounts for the potential cost of reporting unfavourable news. Given how different industries may need different degrees of circumspection when reporting on their CSR efforts, we formulate a third hypothesis:

H3: Corporate (sustainability) report readability will vary along industry lines.

However, we should also nuance this potential tension between what a company can gain from *claiming* to want to engage in CSR and what they can gain by actually doing so by noting a shifting attitude amongst readers. KPMG (2013), as well as Townsend et al.'s (2010) survey, indicate that a shrinking minority of readers sees sustainability reporting as 'greenwashing' (see also e.g. Hrasky 2012, Boiral 2013), i.e. insubstantive impression management, and respondents considered a desire for corporate accountability the prime motivation behind reporting.

Scholars, in turn, appear more sceptical than readers do: Parsons & McKenna (2005) and Boiral (2013), conversely, signal how language used in sustainability reporting can twist its narrative frames to the company's advantage, and Story & Neves (2015) indicate the risks of alienating readers when they perceive corporate social responsibility initiatives as purely strategic. '[I]f organizations do not engage in CSR they may jeopardize their brand and reputation, which, in turn, could decrease short- and long-term profitability,' potentially endangering the company's social and environmental licence to operate (Deegan et al. 2002).

3.2.4 Impression Management: Obfuscation and Defensive Attribution

Like Farewell, Fisher & Daily (2014) found Abu Bakar & Ameer (2011) attest consistently high reading difficulty across a sample of Malaysian CSR communications. Notably, the Malaysian stock exchange mandated sustainability reporting as a listing requirement at the time of data collection. Both studies also found a partial positive relationship between

the readability of a firm's report and its financial performance (expressed, among other variables, by its profitability). They compared these findings which Cho et al. (2010, p. 431), who report "significantly more 'optimism' and less 'certainty'" in environmental disclosures issued by worse performing companies. Abu Bakar & Ameer see this as evidence in support of the 'obfuscation hypothesis' for sustainability reporting.

Although obfuscation in sustainability report text remains under-examined, numerous studies (e.g. Aras & Crowther 2008; Bebbington & Larrinaga-González 2008; Laufer 2003; Neu, Warsame & Pedwell 1998) have focused on the genre's rhetoric. Differences in how composers of sustainability reports represent information might, intentionally or otherwise, cause the "gap [...] between corporate sustainability talk and practice" that Cho, Laine, Roberts & Rodrigue (2015, pp. 78) emphasise. Parsons & McKenna (2005) offer the example of a company making non-falsifiable promissory statements with no set timeframe (e.g. 'we will ensure shared value through community engagement') as a means of rhetorically manipulating sustainability reports.

Accordingly, this study also aims to contribute to greater textual characterisation of the genre by querying whether and how these reports exhibit textual obfuscation. The 'obfuscation hypothesis' (see e.g. Curtis 1998) posits that companies will make less favourable results more difficult to decode, typically out of impression management concerns. The underlying notion is that by making them easier to decode, companies can highlight more favourable results. Conversely, presenting unfavourable results in a more complex fashion relegates them to the background, incents glossing over them, and also impedes full understanding. For the purpose of this study, we will define obfuscation as the presence of a barrier to comprehension of unfavourable (performance) information where such a barrier is absent or less present for favourable information. The prototypical example of this would be a scenario in which a well-performing company explains how it is doing accessibly and clearly where a poorly-performing company uses ambiguous or impenetrably structured phrasing in an attempt to conceal the state of affairs from its readers. Even so, as we have previously explored the latter group of companies may still issue reports, hoping to derive legitimacy from the act of reporting itself.

The focal question in Curtis' (1998) study is whether financial report readability varies as corporate performance does. Within Curtis' sample, there was no significant difference in readability between 'good news' and 'bad news' segments, but companies with more press exposure did show lower average readability and more fluctuation in their texts' readability. Farewell, Fisher & Daily (2014) similarly attest readability variability between sections but does not relate it to performance. Rutherford (2003) and Bayerlein et al. (2010) find that companies do not significantly obfuscate unfavourable information, which contrasts with Curtis' (1995) observation of profitable companies producing more readable chairperson's addresses. Smith & Taffler (1992b) present similar results in favour of obfuscation. Dempsey et al. (2010, p. 19) suggest that companies will

manipulate readability to conceal present unfavourable outcomes, but report on past unfavourable outcomes with greater readability as they are “inclined to improve readability in order to convince capital providers they are worthy of receiving external funding”. In terms of sustainability reporting, Abu Bakar & Ameer (2011, p. 9) find that higher-growth companies publish more readable CSR reports, possibly because “[they] want their stakeholders to easily comprehend the messages in the CSR disclosures”. As a corollary, lower-performance companies that practice CSR reporting primarily in order to legitimise their operations may have less incentive to make the actual outcomes known to their audience.

As the above illustrates, research into obfuscation in financial reporting narrative proves inconclusive and, at times, contradictory. The state of similar inquiries into sustainability reporting is equally problematic, albeit not in its controversy, but in how limited it is. Accordingly, to better investigate this contested issue, we present the following hypothesis:

H4: Corporate (sustainability) report readability will correlate with corporate performance.

We must note, however, that although Abu Bakar & Ameer (2011) investigated obfuscation based on financial performance, sustainability reports integrate several pillars of corporate performance. Cho et al. (2010) show an alternative in examining obfuscation in environmental reports based on environmental performance. As the ASSET4 database that sections 1.3 and 3.3.1 expand on contains four aggregate performance measures we can subdivide the above hypothesis:

H4a: Corporate (sustainability) report readability will correlate with financial performance.

H4b: Corporate (sustainability) report readability will correlate with environmental performance.

H4c: Corporate (sustainability) report readability will correlate with social performance.

H4d: Corporate (sustainability) report readability will correlate with governance-related performance.

Another form of impression management similar to obfuscation occurs in defensive attribution behaviour - “a defensive framing tactic that shifts the blame for negative outcomes away from themselves” (Merkl-Davies and Brennan 2007, pp. 11), as for example Aerts (1994) and Aerts et al. (2011) also recognise. We might expect companies to attribute positive outcomes (such as decreased carbon emissions) to themselves, but negative outcomes (such as an increase in injury rates) to external factors. We might expect different agency patterns between these two attributive frames, with companies more frequently employing the active voice for self-attribution (e.g. ‘We decreased

carbon emissions through supply chain optimisation’) and the passive voice for external attribution (e.g. ‘Worker injury rates were exacerbated by unusually poor weather’).

In contrast to textual obfuscation, defensive attribution does not outright obscure the content of the message, although it can still impede processing; defensive attribution primarily obscures the responsibility behind the message’s content. Potential attribution strategies for impression management include, as Aerts & Cheng (2011) indicate, implicitly denying responsibility for negative outcomes and dissociating the actor from the outcome. As Thomas (1997, 51) finds that the companies use the passive voice to “distan[c]e the messenger from the message” in their chairperson’s addresses and do so more frequently when profits decline, we examine passivisation as an indicator of external attribution as well as a readability measure. Although Chapter 5 will explore agency patterns in greater detail, we do have a limited ability to gauge company use of such textual dissociation tactics through the same tools we use to measure obfuscation. As an addition to the previous hypotheses, we incorporate the following into our analysis:

H4e: Extent of passivisation in corporate (sustainability) reporting will correlate inversely with corporate performance.

Finally, what also merits mention is that if we shift our focus to the use of visuals, we find two studies (Cho et al. 2012a, 2012b) that provide evidence for companies practicing impression management by, for instance, presenting more positive visual information and using non-zero axes for graphical representations of numerical data to emphasise (arguably exaggerate) or conceal results. Companies operating primarily in regions with lesser legal enforcement (as described in Leuz, Nanda & Wysocki 2003) appear to do this more often. We consider such graphical manipulation practices a form of visual obfuscation as it presents an additional hurdle (i.e. the mental recalibration of the axes) that the reader needs to overcome in order to obtain information that the company might want to draw attention away from. While textual obfuscation in corporate (chiefly financial) reporting remains a contested phenomenon and this study is unable to examine visual obfuscation, Cho et al.’s evidence of visual obfuscation in sustainability report reinforces Abu Bakar & Ameer’s (2011) plea for further examination of textual obfuscation in this primarily voluntary, less-legislated genre.

3.3 Methodology

This section describes the various steps taken throughout the study. It discusses, in order, corpus collection, corpus processing, and statistical analysis.

3.3.1 Corpus Collection

Examining the above hypotheses imposes substantial restrictions on the corpus: it needs to be sufficiently large, contain as little noise as possible, represent several (sub-)genres of corporate reporting as well as several varieties of English and, perhaps most problematically, have both financial and non-financial company performance data available for each of the texts contained within.

We found no corpus that could meet these requirements, so the first practical step of this study was collecting one. As the performance information requirement proved most stringent, especially combined with requiring several reporting regions, we opted to first find a database that offered financial and non-financial company performance across multiple regions, and collect reports based on the companies contained within. The most suitable database proved to be Thomson Reuters' ASSET4, which offers both discrete data points for each of the performance aspects and an aggregate score that represents the company's performance for that aspect relative to others; depending on the variable, this is within the same region, the same industry, or worldwide.

Thomson Reuters (2018) claims of the Environmental, Social and Governance (ESG) data present in the ASSET4 database that

[it includes] information on over 7,000+ global companies and over 400 metrics, including all exclusion (ethical screening) criteria and all aspects of sustainability performance. The data is gathered from publicly available information sources and is manually collected to ensure that the information is standardized, comparable and reliable. All of the ESG data collected is quality controlled and verified in a rigorous process by [Thomson Reuters'] experienced analysts and robust automated checks.

From ASSET4's industry categories, which align with the Thomson Reuters Business Classification (TRBC) categories, we selected four industries. Oil and gas, as well as mining and metals, represent the more environmentally sensitive extractive industries, which typically have an inevitable destructive impact on the environment (see e.g. Jenkins & Yakovleva 2014 for mining). The apparel and semiconductors industries represent more socially sensitive industries. We assert these two industries' social sensitivity based on frequent and prominent workers' rights cases, for instance the Savar building collapse in Indonesia (BBC 2013), and conflict materials issues (Bafilemba et al. 2014).

For each of the five regions (US, UK, EU, AUS, IND¹) in the study, we collected up to three texts for each of the companies present in any of the ASSET4 industries, as available:

¹ Each of these represent a geographically distinct variety of English with distinct linguistic influences (e.g. Indian English having a rich diversity of L1 speakers and L2 speakers with disparate linguistic backgrounds,

- Letters to Shareholders
 - o From the (financial) annual report
 - o From the standalone sustainability report
- One sustainability report, prioritised as follows:
 - o A standalone sustainability (or, e.g. CSR) report
 - o Sustainability-related sections or chapters within a company's annual (financial) or integrated report
 - o The company's 'sustainability' section on the corporate website, provided it offered information for the specific financial year rather than a general overview of the company's CSR activities. These were typically still labelled 'sustainability report', but were not standalone publications (i.e. PDF files).

As many companies that had ASSET4 data within Thomson Reuters' Datastream interface still did not offer standalone sustainability publications (and far fewer still offered self-declared 'integrated reports'), this tiered approach was necessary in order to obtain sustainability disclosures for those companies. In the absence of a dedicated sustainability report, we considered sustainability-themed content in the company's most prominent publication to be its de facto sustainability report for that year.² Given the topic of the study, we considered only those documents written in English, even in cases where more or longer documents were available in another language. While performing these readability and other linguistic analyses across multiple languages would be fascinating, answering the question of how to compare texts across different languages, each with their own linguistic norms, would merit a study in its own right, as would the logistics of gathering a sufficiently large multi-language specialised corpus. An inquiry into the translation of corporate reports or variation between languages would likely be highly informative, but there are also severe technical impediments to a corpus-based NLP approach in languages other than English. These impediments include the greater availability of richer and more advanced background corpora and NLP tools in English, and potentially the comparability of tools' output between languages.

European Business English as a Lingua Franca only having the latter, etc.). While South African English would have been a valuable candidate in terms of linguistic diversity, reports would have suffered comparability issues with the rest of the corpus due to South Africa's early adoption of mandatory integrated reporting in 2009 (Institute of Directors in Southern Africa 2009).

² This clustering together of different types of sustainability report for 2012 aligns with Habek & Wolniak (2016). In two cases, there is a small but significant difference between reports originating from the annual report and those published as standalone documents. Standalone reports exhibit slightly higher readability (18.87 compared to 19.49; $p = 0.022$, Partial $\eta^2 = 0.033$) and lower use of passive structures (0.26595 vs. 0.31132; $p = 0.003$, Partial $\eta^2 = 0.56$). While this represents only two out of seven readability measures, neither effect is particularly large, and the difference seems unlikely to register to a casual reader or impact other analyses, these observations offer a minor (due to its inconsistency) amount of support to the hypothesis that report writers may increase readability for sustainability content's wider audience.

We extracted data for the companies' financial year 2012 and attempted to obtain the closest corresponding report. In some cases, such as when companies issued biennial sustainability reports, we selected those reports that contained the greatest proportion of the 2012 calendar year. This ensured the best temporal match possible with the performance data. 2012 was the most recent available (fiscal) year at the time when data collection started.

The majority of companies had at least a letter to shareholders available from the financial report, and given the above methodology, more companies had (de facto) sustainability reports available than letters to stakeholders from sustainability reports, as only the latter category required that they originate from a standalone publication. The tables below describe the final totals; these indicate that while the collection methodology (i.e. collecting, for the chosen parameters, all available texts for companies present in ASSET4) makes for a representative corpus, balance suffers, with some cross-sections containing no documents at all. While, ideally, we would have seen a more even distribution of companies and reports between the different regions and companies, we consider the present scenario the best possible. Pre-screening industries and regions and selecting them for the sake of corpus balance rather than sensitivity to sustainability performance would have made the sustainability report-based approach less relevant, and added another substantial logistical hurdle before data collection.

Table 5 Corpus composition (number of texts) by genre, region and industry.

Genre & Region \ Industry	Mining	Oil	Semiconductors	Apparel	Grand Total
Fin. oriented LtS	95	82	30	12	219
USA	11	35	22	4	72
UK	18	11	2	0	31
Europe	17	15	5	8	45
Australia	44	18	1	0	63
India	5	3	0	0	8
Sust. oriented LtS	38	35	12	3	88
USA	4	14	8	2	28
UK	14	7	1	0	22
Europe	9	8	3	1	21
Australia	8	5	0	0	13
India	3	1	0	0	4
Sustainability Report	78	59	16	10	163
USA	9	18	10	2	39
UK	18	11	2	0	31
Europe	17	16	4	8	45
Australia	29	11	0	0	40
India	5	3	0	0	8
Totals					
USA Count	24	67	40	8	139
UK Count	50	29	5	0	84
Europe Count	43	39	12	17	111
Australia Count	81	34	1	0	116
India Count	13	7	0	0	20
Grand Total	211	176	58	25	470

Table 6 Corpus composition (number of texts) by unique company and genre.

Region & Industry \ Genre	Financial LtS	Sustainability LtS	Sustainability Report	Unique Companies
Australia	63	13	40	64
Mining	44	8	29	45
Oil	18	5	11	18
Semiconductors	1			1
Europe	45	21	45	48
Apparel	8	1	8	9
Mining	17	9	17	17
Oil	15	8	16	16
Semiconductors	5	3	4	6
India	8	4	8	8
Mining	5	3	5	5
Oil	3	1	3	3
UK	31	22	31	31
Mining	18	14	18	18
Oil	11	7	11	11
Semiconductors	2	1	2	2
USA	72	28	39	78
Apparel	4	2	2	4
Mining	11	4	9	13
Oil	35	14	18	37
Semiconductors	22	8	10	24
Grand Total	219	88	163	229

Table 7 Mean, standard deviation (SD), minimum and maximum for different genres expressed in tokens before and after cleaning (described in 3.3.2), and page numbers before cleaning.

	Tokens before cleaning				Tokens after cleaning				Pages in PDF			
	Mean	SD	Min	Max	Mean	SD	Min	Max	Mean	SD	Min	Max
Financial LtS	1875.14	1399.21	274	11780	1793.35	1284.65	252	9766	3.33	2.299	1	17
Sustainability LtS	913.85	552.87	238	3391	885.36	541	225	3272	1.72	0.8	1	4
Sustainability Rep.	18974.48	18241.2	369	98925	14184.27	13650.29	345	78701	45.38	48.366	1	388

3.3.2 PDF and Text Processing

This section describes the process of converting reports from the format in which they are typically available online to the one required for NLP-based techniques. While it proved a highly technically involved process, ensuring the ‘cleanest’ corpus possible was absolutely crucial as every step of NLP-based analysis can introduce more noise to the data and, what is more, amplify existing noise. Minimising this error percolation placed considerable technical requirements on the corpus.

The vast majority of the LtSs and sustainability reports in the corpus came in the PDF format³, which *inter alia* Evans (2007) cautions can be both time-consuming and difficult to convert into the plain-text formats necessary for most computer-aided purposes that a corpus can serve, from searching through automatic processing. Automated readability estimation, both through formulae and more advanced NLP techniques, also necessitates well-formed sentence structures with clearly defined sentence ends, so the often non-linear layout of corporate reporting PDFs posed an additional problem: many of them use text boxes, columns, graphs, tables and other elements that impede conversion to the very linear plaintext TXT format. Fully automatic conversion software yielded issues that made it unsuitable to the level of precision we required (such as generating inconsistent casing), struggled to process these non-linear sequences, and at times simply halted due to report length, which frequently exceeded 100 pages per report.

We explored two options: automatic PDF conversion to Microsoft Word files (.doc or .docx) and automatic conversion to plain text. While automatic conversion to .doc(x) formats often yielded acceptable results, the issue of converting a non-linear visual format to a linear one persisted. Conversion from .doc(x) to .txt generally (re-)introduced the problems that initial conversion to .doc(x) avoided, but was inevitable as the NLP

³ A very small minority of documents was only available as a website. Depending on the complexity of that website’s design, we either saved those websites as .pdf files (for the complex websites) and then converted them to .txt using the above process, for parity, or in the case of very simple websites copied them straight into a .txt file.

suites we use require plaintext input. Even at the .doc(x) stage, however, we frequently encountered misplaced elements such as text boxes, images and text wraparound, as well as inconsistent fonts, including capitalisation errors, with some of the output containing erratic casing.

For purposes that can use .doc(x) formats without further conversion, automatic .pdf-to-.doc(x) may prove sufficient – and vastly more labour-efficient than manual conversion – but it did not resolve the issue of non-linear .pdfs translating poorly to linear .txt. Converting straight to .txt format simply compounds the .doc(x)-related issues and nonlinearity, sidestepping only the image-related ones, and results in a poorly sequenced text file with, for example, the content of text boxes in the original .pdf at times intersecting paragraphs mid-sentence.

To avert most of those issues, we used ABBYY FineReader OCR software, which allows for manual, per-page reordering as well as editing of textual elements, such as tables. While offering a considerably higher level of fidelity and utility compared to automatic PDF conversion, this process was substantially more time-intensive⁴, which constrained corpus size and potential for expansion. Nevertheless, the ability to apply human judgment to text order on every page yielded results that better corresponded to the design of the original text in .pdf format than otherwise possible. This was neither an automatic nor a deterministic process; as it relied on human intervention to step in where automatic processes fell short, it did introduce a measure of subjectivity.

We attempted to be as consistent as possible in focusing on four key interventions:

1. Tagging numeric and mixed-content tables;
2. Discarding metatext.
3. Normalising casing; and
4. Joining fragmented paragraphs and sentences.

In order to ensure the first intervention, we manually added a <table> tag to the start, and a </table> tag to the end of every table element's text output within FineReader. The presence of these tags enabled us to formulate subsequent regular expressions⁵ to not extract numerical or mixed-content table elements, while judging tables with running text on how closely they resemble conventional running text.

As FineReader supports autodetection and hiding of metatext, the second intervention was the easiest step out of the four. When FineReader encounters a recurring textual layout element (such as a page number), it marks that element as a 'header or footer':

⁴ While orders of magnitude faster than fully manual text conversion or copy-pasting, this approach still required human intervention. Notable cases include whenever a table or image occurred (to flag them as such), or a paragraph broke across page limits (to reattach the text to one of two sides); as a result, processing a long (e.g. 200+) document could still take upwards of several hours, up to a day.

⁵ We executed all regular expressions using PowerGrep version 4.

FineReader recognised the text, but did not insert that text when we saved the document as a .txt file. As it is possible to tag text boxes as headers or footers, we were able to manually tag or untag textual elements as metatext where required.

We achieved the third intervention through a two-part sequence of regular expressions, with '=' signifying replacement:

1. `\b([a-z]?[A-Z][a-zA-Z]*)\b => \L1`
*Within word boundaries, detect every sequence of letters (word) that starts with a lowercase letter but contains at least one uppercase character.
Replace this uppercase character with its lowercase equivalent.*
2. `\b([A-Z])([A-Z]*[a-z]?[A-Z][a-zA-Z]*)\b => \1\l2`
*Within word boundaries, find any sequence of letters (word) starting with an uppercase character that contains at least one uppercase character other than the first.
Preserve the first uppercase character. Replace the other uppercase characters with their lowercase equivalent.*

We normalised only towards lowercase. Working together, these regexes capture (1.) words that start with lowercase letters but contain uppercase letters, and (2.), words that start with an uppercase letter but contained at least one lowercase letter. We assumed that such erratic casing is the result of erroneous conversion, and thus normalise them towards conventional sentence case, only maintaining initial uppercase letters to respect potential capitalisation of proper nouns etc. In other words, this normalisation is fairly conservative; it does, however, disrupt unconventional capitalisation patterns in those rare cases the author intended them as such. One example of such an unconventional pattern would be Apple's 'i'-range of products, such as iPhone, which the above solution would convert to 'iphone'. We considered these fringe cases an acceptable cost of implementing the above algorithm.

Regarding the joining of fragmented sentences and paragraphs (the fourth intervention), one of the main weaknesses of the FineReader software suite for our purposes was that it treats individual pages as separate sub-documents. That resulted in the software splitting paragraphs or sentences that ran across multiple pages in two. We remedied this by manually joining them together. We took the start (or end) of the fragmented paragraph or sentence and added it to the last (or first) fragment of the paragraph on the next (or previous) page. This simulated the paragraph starting and ending on a single page. We see no downsides to this approach, except for how time-consuming it is and potential oversights due to the manual annotation. To minimise such oversights, we also automatically detected and repaired sentences containing erroneous linebreaks (as would result from a page transition in FineReader) in the output using regular expressions. We applied the following regular expression replacement sequence (with '=' indicating replacement), and went back to correct the files in FineReader where necessary:

1. $^((?:[a-z,"()]+){2,}.*?[.?!]) \Rightarrow \text{<LAMatch>\1}$
 Tag every potentially erroneous start of line with '<LAMatch>' (signifying 'line anchor match') tags.
 Detect potentially erroneous starts-of-line: start of line anchor followed by at least two sequences lowercase letters or mid-sentence punctuation such as quotation marks or brackets followed by a space, followed by sentence boundary punctuation (full stop, question mark, exclamation mark).
2. $(?<=[.?!] ?)((?:[^\r\n][^\r\n]{4,}))$ \Rightarrow \text{<LBMATCH>}$
 Tag a number of potentially erroneous ends-of-line with '<LBMATCH>' (signifying 'line break match') tags.
 Verify that at least one sentence boundary punctuation mark followed by a space exists before the potential match.
 Match any string of non-linebreak, non-sentence boundary characters followed by at least four words (i.e. sequences of non-linebreak, non-punctuation characters followed by one or more non-capitalised, non-linebreak, non-punctuation character, separated by spaces) followed by the end-of-line anchor.
 This matches at least four-word-long sentence fragments that do not terminate as we would expect sentences to, nor appear to start a new sentence, that strictly follow normal sentence case.
3. $(?<=[.?!] ?)((?:[^\r\n][^\r\n]{3,})(?:[^\r\n][^\r\n]{3,}))$ \Rightarrow \text{<LBMATCH>}$
 Tag a number of potentially erroneous ends-of-line with '<LBMATCH>' (signifying 'line break match') tags.
 Verify that at least one sentence boundary punctuation mark followed by a space exists before the potential match.
 Match any string of non-linebreak, non-sentence boundary, non-capitalised characters, followed by sequences of more than three non-linebreak, non-sentence-boundary, non-lowercase characters, followed by a non-linebreak, non-sentence boundary, non-capitalised sequence of characters, followed by the end-of-line anchor.
 This allows a sequence of capitalised words to exist within a paragraph, as was sometimes the result of font-related capitalisation issues. The capitalisation normalisation regex will have already converted any instance of non-sentence case to full capitals.
4. $(?<=[.?!] ?)([^\r\n][^\r\n]{3,})$ \Rightarrow \text{<LBMATCH>}$
 Tag a number of potentially erroneous ends-of-line with '<LBMATCH>' (signifying 'line break match') tags.
 Verify that at least one sentence boundary punctuation mark followed by a space exists before the potential match.
 Match strings of three or more sequential words without initial capitals (no linebreaks or sentence boundary punctuation) followed by a space and at least three non-

lowercase, non-sentence boundary, non-linebreak characters before the end-of-line anchor.

This pattern attempts to capture partial sentences where the very last textual element does not match either sentence case or full capitalisation (for instance a number), but looks at capitalisation in the preceding sentence more stringently in exchange.

5. `<LBMatch>\r?\n<LAMatch> =>`

Remove any linebreaks, optionally preceded by carriage returns surrounded by erroneous linebreak or line anchor tags on both sides, as well as the tags themselves.

6. `<LBMatch> =>`

Remove erroneous linebreak tags that step 5 did not remove (i.e. tags without matches).

7. `<LAMatch> =>`

Remove erroneous line anchor tags that step 5 did not remove (i.e. tags without matches).

As with the above regexes, this sequence was the result of trial-and-error iterative improvement; it provided good results for this corpus, which favoured precision over recall, but may perform significantly differently on other (types of) text. As the PowerGREG regular expression tool also enabled us to manually examine cases where either the start or end of a string contained a tag, we were further able to increase accuracy by manually examining those instances of potentially problematic linebreaks before steps 6 and 7 eliminated the non-matching tags.

3.3.3 Running Text Extraction

Finally, as our various means of analysis benefited substantially from receiving full-sentence input, we stripped away the largest possible amount of non-full-sentence content (e.g. bullet points and tables) through a regular expression. This process removed over a million tokens relative to output of the PDF-to-text-conversion process, but yielded a (sub-)sample of relatively reliably full-sentence text. The regular expression was the result of extensive trial and error combined with iterative improvement; it is dense, but functional in how closely tailored it is to this corpus:

```
(?<=^\d{0,3}\W*\d{0,3}\W*)(?!<\W*\t\W*\t\W*)((["'"] ?)?\w[^\t]+?(?<=^[^\t]+?[^\t]+? [^\t]+?)[:;!]( and| or)?,( ?["'"] ,))?(?=[0-9]\W)? ?$)
```

Verify that the following exists before the match: a start-of-line anchor, optionally followed by up to two sequences of three digits separated by whitespace. The regex does not capture them, but does allow for their existence, to capture text behind enumerations. Verify that the match is not preceded by two or more tabs separated by non-word characters (as could be the case for tables, for instance, which we want to exclude). Match optional quotation marks optionally followed by a space, followed by a word character (letter, digit or connector punctuation) followed by any sequence (as short as possible) of non-tab characters.

Verify that the preceding characters can be defined as three sequences of arbitrarily long (but as short as possible) non-tab characters separated by spaces (this ensures paragraphs are at least three words long).

Match end-of-sentence punctuation: semicolon, colon, full stop, question mark or exclamation mark.

Match an optional 'and' or 'or' preceded by a space.

Match an optional comma.

Optionally match an (optional) space followed by closing quotation marks.

Verify that the match is followed by an optional space, optionally followed by a digit or non-word characters, followed by an optional space, and finally the end-of-line anchor. (This accommodates numbers or other characters that refer to foot- or endnotes; unless these are present, the next character must be the end-of-line anchor.

This regex extracted 2.75 million words out of the initial 3.95. While it may seem wasteful to discard more than a million tokens of potentially usable text – over one fourth of the initial corpus – with a very stringent regular expression. We made this choice because we needed to prioritise precision over recall – that is, making sure that every instance of full text we extracted was in fact full text was far more crucial than extracting every bit of running text. If we pass along a false positive to the NLP tools we use, we make it process data is not designed to process. That would introduce (avoidable) noise into the data, making them less reliable. On the other hand, generating more false negatives simply means generating less data to pass along to the next step. While that might make the data we pass along to the next steps less representative – by discarding usable elements of the corpus – it does not make the data less reliable. While neither situation is optimal, we choose – by lack of a flawless means to distinguish usable running text from non-running text – the lesser evil of favouring fewer false negatives. In that respect, the corpus prioritises quality over quantity; section 3.3.6 explores that decision in greater detail.

In terms of representativeness, this regular expression extracted more than 60% of the total tokens in over 96% of texts (453/470) and more than 40% of the total tokens in over 99% of texts (467/470). Of the three remaining texts, only one of them (Repsol's 2012 sustainability report) proved truly problematic at 1.6% tokens extracted; the next lowest percentage of tokens extracted was 31%. Closer investigation revealed why the regex was able to extract so little running text in Repsol's sustainability report: a table-style report markup, with the theme in the left column and the relevant information in the right, such as the following one:

Table 8 Illustration of text extraction issues for non-running text.

ACTION	Enhancing our code of conduct.
DESCRIPTION	The Ethics and Conduct Regulation for employees, which gives legal support to our Code of Conduct, was approved in 2006. Since then, there has been much development in corporate responsibility and regulatory changes in this area. [...]
INDICATOR	The presentation of a proposal for updating the Regulation to the Ethics Committee and subsequently to the Executive Committee.

While it would have been possible, if labour-intensive, to refactor the left columns into headers in order to extract the right column, such an approach would have been neither labour-efficient, nor authentic to the initial layout, as the authors of this report clearly chose a non-running text structure. Reconfiguring the already well-tested regular expression would have also increased the risk of errors, such as false positives in table form, in other texts. As we could simply discard the resulting text, which was too short to be suitable for our analysis (we set the minimum number of tokens at 200 after cleaning), we consider omitting a single text a reasonable trade-off for a low rate of false positives and fairly low rate of false negatives.

With regards to false positives, we did also encounter a few outliers where texts consisted mostly of run-on lists separated by semicolons. As semicolons do not, strictly speaking, end the sentence, we processed these as single sentences, as opposed to separate sentence fragments. Treating them as full sentences may also be more representative of their cognitive load. Such lists were sufficiently rare throughout most texts that they did not disproportionately affect the analysis, save for a few outliers we were able to capture and eliminate in an initial exploratory analysis of the data. We extract one example of such a sentence from Sandfire Resources' 2012 Annual Report:

(6) To achieve these aims the Company:

- Communicates regularly with stakeholders, the community, our employees, and regulatory authorities;
- Integrates environmental considerations into all aspects of the Company's business including exploration, planning, development, operations, rehabilitation and decommissioning-closure activities;
- Develops and implements effective management systems that encourage proactive environmental management and continuous improvement of environmental performance;
- Designs and develops new facilities with regard to environmental sensitivity, and where practicable, seek to reduce the impact of operations on the environment through the efficient use of energy and water, as well as responsible handling of waste and other materials;
- Strives to outperform statutory requirements in all areas of operations including, but not limited to, management of hydrocarbons, tailings, saline water and non-process waste;

Progressively rehabilitates areas in a responsible manner;
Ensures all employees and contractors are environmentally aware and are accountable for their individual and corporate environmental responsibilities; and
Actively seeks innovative and sustainable solutions to meet environmental needs.⁶

If we do not consider the semicolon a sentence boundary, the above is a 154-word sentence. As we found it difficult to determine conclusively to what extent such list-based sentences increased the text's cognitive load as much as a more conventional sentence structure would at this length, we opted for the moderate approach of preserving those texts where these sentences exist amongst other sentence structures, and eliminating those texts that consist almost entirely of list-based sentences. The above excerpt, for instance, came from a text our initial exploration placed at a 32.5 Gunning Fog score, or a grade level requiring 20 years of education beyond secondary. As such fringe cases of text less suited for formula-based analysis might have lowered their overall accuracy, we eliminated the three texts above a 28 Fog Score during the aforementioned exploratory analysis.

3.3.4 Corpus Processing

For aspects related to automatic readability estimation, we mainly processed the corpus using the Stanford NLP Group's CoreNLP toolset (Manning et al. 2014), which its creators describe as "an extensible pipeline that provides core natural language analysis". CoreNLP automatically annotates plain text (hence the extensive plain text conversion process) with various kinds of linguistic information. Its most relevant functions ('annotators') to this study include:

- **Tokenisation**, which separates a text into individual words and, optionally, punctuation elements. It deals with such issues as splitting 'aren't' into two tokens representing 'are' and 'not' and is especially important in ensuring consistency, i.e. that every instance of a same word, even if it is difficult to tokenise such as 'aren't', is consistent throughout and between texts (Manning, Raghavan & Schütze 2008).
- **Sentence splitting**, which delineates sentences within the input. As the relative complexity of the above regular expression patterns may have indicated, delineating sentence boundaries is not a trivial task. CoreNLP approaches sentence splitting somewhat similarly to the above regexes: after the tokenisation process, it detects sentence boundary punctuation – a full stop, exclamation mark or question mark – and examines whether it occurs together with other characters

⁶ While these sentences arguably include enough 'rest points' to keep cognitive load manageable, this example also registered to Microsoft Word 2016 as a 'long sentence', further underlining how humans and NLP often deal differently with text.

in a single token (as might happen with a number or abbreviation). If it does not, it considers it a sentence boundary token, although it might count trailing elements such as quotation marks or brackets as part of the same sentence despite them occurring after the sentence boundary.

- **True case detection**, which attempts to discover the intended casing of a token where that has been lost. Despite our own regular expression-based attempts at normalising casing, this is a useful additional step, as it for instance normalises the case of fully capitalised words (Manning et al. 2014), which the regex-based normalisation does not; in this respect, the two synergise.
- **Part of speech tagging**, which indicates for each token which part of speech (e.g. noun, adjective, verb, adverb, etc.) it functions as in the sentence. (Toutanova et al. 2003)
- **Lemmatisation**, which annotates every token for its base form (where applicable), such as ‘buy’ being the base form of ‘bought’ or ‘good’ being the base form of ‘best’.
- **Named entity recognition**, which recognises those tokens or token sequences that represent people, locations, organisations or other named entities as well as numerical entities such as money, numbers, dates, time expressions, durations or sets (Chang and Manning 2012).
- **Parsing**, which analyses the syntax of sentences into constituents and dependency trees. This enables insight into how the different components of the sentence interrelate.
- **Coreference**, which is likely the most difficult task out of the above. It describes how concepts re-occur throughout the text, both as (proper) nouns and pronouns representing them. As coreference, contrary to any of the other annotators, can pass over sentence boundaries, determining coreference gets computationally more demanding the longer the text is.

The CoreNLP suite’s output not only enabled us to calculate the various ‘classical’ readability formulae, but also the more NLP-intensive, deeper-level metrics (such as the depth of the parse tree or number of subordinating elements). While both types of metrics were calculable from CoreNLP’s output, neither type was transparently available; we obtained the linguistic variables (such as the formulae, and lexicosyntactic variables) this chapter uses by processing the output again with De Clercq’s (2015) Python-based feature extraction pipeline.

3.3.5 Statistical Analysis

We obtained most of this study’s results (we explicitly note the exceptions) by building general linear models for each of the various outcome variables on a per-genre basis. Using general linear models enabled us to combine continuous variables (e.g. performance scores and company size) and categorical variables (e.g. region and

industry)⁷. These analyses include all the independent variables (see section 3.4.2), as well the interaction between industry and region, in a model simultaneously.

We set the alpha level (i.e. threshold of significance) to the standard .05 (but do indicate p values below 0.1). That is, we consider a result significant when probability of results occurring if the null hypothesis were true is 5% or less, but will highlight those cases between 10% and 5% probability. We also indicate percentage of variance explained, both adjusted and unadjusted R^2 (adjusted between brackets) for the models where available. This is a measure of how well the model predicts the data, and thus higher values are more desirable. Adjusted R^2 applies a penalty for additional independent variables with low explanatory value, in order to penalise overfitting; the adjusted R^2 value is thus more representative of how the model would perform on unseen data. For individual differences or predictors, we include another effect size measure, typically Partial Eta² or Cohen's d. Because independent variables can show a highly significant correlation with the dependent variable while having little to no explanatory power, it is also important to report how much variance the individual independent variable explains; independent variables with a higher effect size will have a larger impact on the explanatory power of the model.

3.3.6 Limitations

As is the case for almost any study, a number of the methodological choices come with their downsides, the most important of which we will explore before moving on to the analysis proper.

Paratext

As we have previously noted, this study systematically discards paratextual elements such as headings, tables, graphs, figures, illustrations in order to ensure maximal compatibility with the CoreNLP and suite and De Clercq's (2015) feature extraction pipeline that analyse the remaining textual data. This approach creates the least amount of noise. While these are considerable advantages, we must also remain mindful that eliminating paratext makes the content and even number of such elements unavailable to us – or at least unavailable to the computer. This elimination comes at the expense of useful information. Cho et al. (2012a, b) for instance, attest a number of obfuscation

⁷ Performing the same analyses with linear regression models that encode the categorical variables as dummies yields functionally similar results, albeit at times slightly more pronounced due to less conservative p-value correction for the subcategories than general linear models' Bonferroni-corrected post-hoc analyses offer. These same analyses indicated fairly little risk of multicollinearity issues, with very few of the Variance Inflation Factors for the independent variables exceeding 3, and none exceeding 4.5.

strategies present in corporate report graphs. As we consistently removed paratext throughout all the documents the impact of that removal on analyses should be minimal; even where it is not, it remains necessary; Farewell, Fisher & Daily (2014) note the same necessity and similarly remove non-full-sentence elements. We should, however, note that as text and figures tend to reinforce one another in corporate reporting, some companies in the corpus may have chosen to use highly illustrative paratext (such as graphs) to aid comprehension in those cases where the content they are trying to convey contains a great deal of (potentially irreducible) complexity⁸. In other words, simultaneously considering text and paratext will inevitably give a more complete impression of the content and its difficulty.

Furthermore, this limitation restricts us to minimal or no paratext in respondent-based studies. Even where the NLP tools' restrictions are not immediately relevant, such as when having experts or representatives of a general audience assess texts, we must, for the sake of comparability, keep even potentially relevant paratext out of the material we present those respondents. Given that the base system is unable to deal with paratext, introducing it for human assessment would create a confounding variable: would participants respond as they do due to the text, or the paratext? In order to mitigate this problem, we opted to offer only headers where necessary to delineate the text, and no other paratextual material.

Causality

Another limitation of the study lies in its chiefly empirical design, which enables it to survey a relatively large amount of data, but focuses on observation of real-world data over experimentation. That makes it substantially more difficult to control for the influence of certain variables, and in doing so accurately infer causal effect. Consequently, this study's quantitative approach is mostly limited to describing associations and correlations rather than causation. For instance, where we see an association between language and performance, we might assume a causal relationship. Because the performance that the report describes will almost always precede the report's dates of

⁸ In order to verify to what extent this might occur, we composed three linear models (one per subgenre) in which we attempted to predict the Flesch score for a given text through the amount of paratext we deleted and the textual density for that document. We expressed the former as the number of tokens after 'cleaning' divided by number of tokens before, and the latter as the number of tokens after 'cleaning' divided by the number of pages in the original document. Textual density proved a significant ($p = 0.007$) predictor for financial LtSs, albeit with a fairly low effect size of 0.033. The only other predictor with a significance lower than $p = 0.256$ was amount of noise removed from sustainability report body text at $p = 0.074$ and an effect size of 0.02. As these indicate minor and inconsistent effects, if any, we expect this removal of paratext not to interfere with the analyses, although the interplay between extent of paratext and readability or understandability might nevertheless merit future research.

writing and publication, we might assume a causal relationship as the language cannot somehow retroactively influence the performance. However, because this association does not occur in a controlled environment, we are considerably less able to make strong inferences about the causality behind this association, as both phenomena might be the result of a correlated – but not measured – third variable, such as economic climate.

Therefore, to complement the larger-scale empirical enquiries, future chapters introduce a number of more qualitative ones that ask for respondents' input. These include two annotation tasks – one for readability, and one for sentiment – and a questionnaire on the effects of reading difficulty on company and text perception. As the first two allow for some measure of introspection – participants had a limited ability to motivate their judgments – we can assess causality with a higher degree of confidence, although we must remain mindful that even when participants assess their own perception we cannot be entirely certain that what they indicate as a causal relation is in fact such. The questionnaire-based experiment, which compares and contrasts a real-world text with versions of the same text manipulated to match those difficulties is therefore best positioned to ascribe causal relations to the observed changes. The study, however, has its own limitations: it is insufficiently granular to determine which simplification (for instance more active language compared to shorter sentences) impacted which aspect of perception.

Human Judgment

At various stages, components of this study relied on human judgment in order to gather actionable data. While we took care to approach every step systematically, that means outcomes are far less deterministic than they would be for a fully automated process. For instance, the manual text conversion process meant manual reordering of text elements in collapsing a non-linear PDF down to a linear plain text file, as well as manual rejoining of paragraphs assisted by a regular expression-based search algorithm. While a purely automated approach may have been possible, we expect the gain in accuracy from human intervention where necessary to considerably outweigh potential inaccuracies from human error. In other cases, such as the respondent-based aspects of the follow-up studies, human judgment was an inextricable component of the experimental design.

Scope

Due to the considerable time demands of manual, page-by-page corpus compilation, the size of the corpus is relatively limited. While the 2.75 million tokens that make up the corpus' usable full sentences still contribute to a viable specialised corpus, the demands of the compilation process made it considerably less viable to expand the corpus in more dimensions than a fully automated process might have done. Expanding, for instance, to more English-using countries (for instance Canada), industries with different

sustainability (in)sensitivities, or multiple financial years could have enabled a richer set of analyses.

A diachronic corpus, especially, that captured multiple (consecutive) financial years could not only have enabled a richer corpus in terms of data points, but would have also enabled us to compare the evolution of language and performance, and the extent to which the two interrelate. As the corpus already contains multiple countries (and, consequently, varieties of English) and industries, but only a single fiscal year, adding comparable texts for preceding or subsequent fiscal years and circumstances (for instance pre- and post-crisis) would make it considerably more versatile. Although such an expansion would most enrich the corpus, it would also be the most time-consuming to implement. While adding more regions becomes more time-consuming as the number of industries expands and vice versa, every fiscal year added would, in attempting to provide comparable documents for each one already present in the corpus, likely again take the same amount of time the initial corpus compilation process did.

In spite of the above, the assembled corpus of corporate (sustainability) reporting is, to the best of our knowledge, unique in its size and attention to minimising noise, as well as its variety (in terms of language and, to a lesser extent, industry). The latter especially mitigates the risk of potential bias, region-specific or otherwise, by drawing on a wide set of language varieties and legislative contexts. It is, in short, well-equipped to answer this study's research questions and likely myriad others.

3.4 Analysis and Discussion

With these methodological aspects firmly established, we are ready to begin exploring the sample. We reiterate the number of hypotheses this analysis will attempt to investigate:

Our first hypothesis (H1):

Sustainability report content will be more readable than financial report content.

Our second hypothesis: (H2):

Corporate (sustainability) report readability will vary along language variety (regional) lines.

Our third hypothesis: (H3):

Corporate (sustainability) report readability will vary along industry lines.

Our fourth hypothesis (H4):

Corporate (sustainability) report readability will correlate with corporate performance.

As sustainability reports integrate several pillars of corporate performance, we can subdivide H4 into:

H4a: Corporate (sustainability) report readability will correlate with financial performance.

H4b: Corporate (sustainability) report readability will correlate with environmental performance.

H4c: Corporate (sustainability) report readability will correlate with social performance.

H4d: Corporate (sustainability) report readability will correlate with governance-related performance.

We recall that while extent of passivisation merits examination as a facet of readability, it can also be an indicator of defensive attribution:

H4e: Extent of passivisation in corporate (sustainability) reporting will correlate inversely with corporate performance.

Compared to financial reporting, attempting to detect obfuscation at a text level in sustainability reporting has the drawback of sustainability reporting's multiple perspectives (i.e. they contain information on social aspects, sustainability aspects, etc.). As such, text-level analysis is less able to isolate sustainability topics and their associated performance from other topics and performances. Chapter 9 investigates how future research could benefit from avoiding these drawbacks, and Chapter 5 investigates the association between positivity or negativity surrounding an aspect and its associated performance closer to a per-sentence basis.

3.4.1 Exploring the Sample

Table 9 summarises minima, maxima, means and standard deviations (SDs) for each independent variable, on a per-genre basis. Formulae are rounded to three decimals, syntactic features to five for additional precision.

Table 9 Per-genre summary of minima, maxima, means and standard deviations (SD) for dependent variables.

Genre	Variable	Min	Max	Mean	SD
Sustainability Reports (n = 157)	Flesch	-3.692	43.973	26.797	6.651
	Kincaid	11.182	21.363	15.277	1.463
	Fog	14.761	26.266	19.15	1.695
	Lexical Density	0.5926	0.7173	0.64335	0.02072
	Parse Tree Depth	7.6	12.91905	10.39074	0.81328
	Subordination	0.16364	1	0.44419	0.12996
	Passives	0.058182	0.58824	0.28647	0.09602
Letters to Shareholders (Financial Annual Reports) (n = 217)	Flesch	11.839	56.945	34.672	8.185
	Kincaid	9.697	20.897	14.556	1.941
	Fog	12.911	24.906	18.226	2.1686
	Lexical Density	0.54178	0.70166	0.62735	0.02542
	Parse Tree Depth	8.10638	15.31579	10.93707	1.22899
	Subordination	0.09375	1.33333	0.51535	0.19823
	Passives	0	0.47619	0.20805	0.085908
Letters to Stakeholders (Sustainability Reports) (n = 88)	Flesch	5.274	48.537	28.11	9.225
	Kincaid	9.998	23.171	15.215	2.172
	Fog	13.03	27.21	18.736	2.357
	Lexical Density	0.53882	0.68533	0.60691	0.0289
	Parse Tree Depth	8.48485	15.81818	11.00723	1.37827
	Subordination	0.18519	1.45455	0.62161	0.27355
	Passives	0	0.63636	0.18147	0.09634

3.4.1.1 Readability Formulae

If we recall Flesch's previously explored conception (1979) of the 'bands' of readability within the FRE formula, the upper bands of 'easy' reading between scores of 70 and 100+ appear irrelevant to this corpus; the highest outcome out of any text in the corpus falls just short of even the optimal 60-70 'Plain English' band, at an FRE score of 56.945. Overall, the FRE formula places the most-readable texts within the corpus within the 30-60 'Difficult' band, with the more difficult texts in the 0-30 'Very Difficult' band, with a few sustainability reports descending below 0, into what Flesch (1979) calls the "virtually unreadable." We can also note that the FRE mean for sustainability reports and their accompanying LtSs exists in the 'very difficult' band, with only the mean for LtSs from financial reports exceeding that threshold, although the FRE's dividing line between difficult and very difficult writing does lie within a standard deviation's distance for each of the genres.

We have also previously discussed how the FKGL and GF both express their results as an estimate of a US educational grade level, albeit with different weightings and different conceptions of word length; counting the number of syllables (FKGL) or using a cut-off for syllable count (GF). In the case of corporate reporting, the GF score assesses reports and LtSs as resoundingly more difficult than the FKGL score does, with estimated required grade levels approximately four higher across minima, maxima and means for the Gunning Fog Index, although even the more conservative FKGL-based estimates place the most readable text in the corpus (with a score of 9.697) above the threshold for general readability of 8th-grade writing. In that respect, both formulae align with the FRE score (as they logically should, given all three formulae are functions of word and sentence length).

Based on the FKGL scores, we might assert that these corporate reporting (sub-)genres are, on average, readable at the undergraduate level. While this casts corporate reporting as a specialised genre that benefits from specialised (or at least advanced general) education if its readers are to fully decode it, an average of undergraduate-level readability simultaneously discredits a more general-purpose, diversified-audience interpretation of the sustainability report. While the most readable instance of its most widely-read section (the Letter to Stakeholders) moves closer to general readability, an FKGL score of 9.998 no longer implies an undergraduate-level reading ability, but neither does it accommodate an 'average' 8th-grade reader, even before accounting for the effects of language variety as a complicating factor.

The GF score is substantially less optimistic still; it places the mean score across all three (sub-)genres well into the postgraduate range. We might argue over precisely how

to interpret ‘full understanding’⁹ upon asserting that FKGL and GF scores indicate educational experience required for full understanding of a given text. It seems, however, highly implausible that the reader of an average sustainability report or LtS will require a postgraduate level of education to achieve any of their goals with the text (some of which may not require full understanding). In that respect, the FKGL score seems better suited to this study for readability estimation purposes, although we will carry forward all three measures to ensure maximal comparability.

Finally, this also recalls the now-familiar caution that readability formulae are an estimate rather than a precise instrument. Given the substantial divergence in users’ potential goals with a given text, a single number derived from word and sentence length can only *approximate* how likely those users are to succeed on average. Nevertheless, the outcomes above indicate that sustainability content may be *less* readable than even financial content; there is little doubt that the average reader would struggle when decoding these reports. While who exactly companies envisage that average reader to be will depend on the company itself, any non-experts’ comprehension would likely benefit from more accessible writing.

3.4.1.2 Lexicosyntactic Features

Section 2.4.2.2 previously explored how lexical density can help quantify readability, but also underlined how, contrary to the readability formulae, its relationship with readability is unlikely to be strictly linear: Castello (2008) indicates an approximate dividing line of under and over 40% for spoken and written language, respectively. Gibson (1993) does indicate, at least for abstracts, that lower lexical densities can create greater ease of reading – although, in accordance with Castello (2008) we might expect diminishing returns, or even the inverse, close to lexical densities of 40%. All of the above makes it considerably more difficult to analyse what a given percentage of lexical density implies.

Nevertheless, we can make a few key observations in terms of lexical density. First of all, across all three genres, we can observe lexical densities between ~53% and ~72%, with the mean lexical densities ranging between approximately 60% to 65%, with standard deviations ranging between 2% and 2.55%. In other words, we can observe that all the texts within the corpus are well above the 40% threshold. However, at the same time we have reason to believe that the approximately 7% difference between the most dense sustainability report and the ~65% mean for sustainability reporting may be a larger impediment to readability than that between LtSs’ mean of ~62% and their minimum of ~55%, due to potential diminishing returns when lowering lexical density.

⁹ This statement also draws on the tension between readability and understandability; we refer to section 2.4.1 for further detail.

Given that every single text in the corpus is at least five standard deviations removed from the minimum lexical density of 40% for written text, it appears – in spite of how difficult it can be to define a precise or ideal target for lexical density – that corporate reporting is unlikely to begin suffering from lexical sparsity any time soon. Even the least dense texts are likely to have room for additional function words, though such a change may come at the expense of linguistic efficiency in that they would take longer (i.e. use more words) to convey the same content.

Parse tree depth, in turn, is likely even more difficult than lexical density to assign a meaningful interpretation to. It is useful as an overall gauge of syntactic complexity (Beaman 1984) and a potentially highly informative feature for computer-based readability prediction, but difficult to assign an intuitive meaning to. We can notice, however, a wide range between minima and maxima, with the maximum of approximately 15 average levels per sentence (for sustainability report LtSs) more than double the minimum of 7.6 in the sustainability report body text with the shallowest parse trees. The means of parse tree depths across the different (sub-)genres, however, are fairly consistent, although the two types of LtS appear more variable, and seem to have higher maxima and minima. Section 3.4.1.3 will explore how features differ between subgenres in greater detail.

Extent of subordination, like parse tree depth, is again difficult to interpret as a numeric value, for much the same reasons; additionally, Section 2.4.2.2 has already discussed how and why extent of subordination does not necessarily correlate with linguistic complexity. In spite of that, however, we can note a remarkably high variability in the extent of subordination texts use. Means across the different (sub-)genres vary between an average of ~0.45 and ~0.62 subordinators per sentence. Moreover, the standard deviation approaches up to half of the mean (in the case of LtSs from sustainability reports). Similarly, minima and maxima vary between an average of less than one subordinator every tenth sentence and almost one and a half subordinators per sentence, again indicating high variability. While it may not be the most straightforward syntactic feature to interpret, its variability alone makes it noteworthy.

Extent of passivisation (expressed as number of passives per sentence), finally, is by far the easiest to interpret of the syntactic features; the above scores indicate the average number of passive structures per sentence. Like extent of subordination, it displays a remarkably wide range: while the most active LtSs are short enough to contain no passives (~16% for sustainability reports), the least active one contains ~64% passive-voice constructions (~59% for sustainability reports). Paired with the means for LtSs (~18%–~21%) and sustainability reports (~29%), this relatively high variability suggests that while it is possible to write (predominantly) active-voice corporate reporting text, the majority of corporate reporting language contains a relatively high number of passive structures. These rates are especially high compared to the average rate of passive verb forms in the British National Corpus, which Roland, Dick & Elman (2007) place at 9%.

Although the passive certainly has its uses, we have previously explored how more passives are likely to correlate with higher reading difficulty (see section 2.4.2.2). Based on that assumption, texts that contain two to three times the (relative) number of passives general writing does are more likely to challenge the average reader. Some of this more indirect agency may, of course, be attributable to defensive attribution tactics, i.e. companies seeking to disown unfavourable outcomes through more distant agency framing. Section 5.8.3 will examine the phenomenon in greater detail.

3.4.1.3 Comparison between Genres

In order to have a better idea of how the genres interrelate in terms of formula- and syntax-based readability, we conducted a series of matched-samples T-tests within a subsample of those companies that had all three (sub-)genres present in the corpus: LtSs from both financial reports and sustainability reports, and the sustainability reports proper. This was in addition to the requirements for all other models (GF equal to or below 28, and 200 or more tokens remaining in the document after cleaning). Performing a series of matched-samples T-tests on only the 79 companies that met these requirements effectively controlled for the other independent variables used throughout this study, as every region, industry, and performance metric saw equal representation throughout the three genres, owing to the balance of the subsample. The following table describes that analysis.

Table 10 Summary of genre differences (means, standard deviations and significance of difference with other genres) for main readability measures used in this chapter.

Based on a subcorpus limited to companies that provide all three genres. Includes mean, standard deviation (SD) and significance of difference with other (sub-)genres as determined through matched-samples T-tests. Marginal significance ($p \leq 0.1$) in *italics*, significance ($p \leq 0.05$) in **bold italics** and strong significance ($p \leq 0.01$) in **underlined bold italics**. Means and DS for Formulae are rounded to three decimals; syntactic features to five for additional precision.

	Sustainability Reports (n=79)				Financial (Annual Report) Letters to Shareholders (n=79)				Sustainability Letters to Stakeholders (n=79)			
	Mean	SD	Diff. W. Fin. Lts (p)	Diff. W. Sust. Lts (p)	Mean	SD	Diff. W. SR (p)	Diff. W. Sust. Lts (p)	Mean	SD	Diff. W. SR (p)	Diff. W. Fin. Lts (p)
Flesch	27.841	5.029	<u><0.001</u>	.382	35.163	7.868	<u><0.001</u>	<u><0.001</u>	28.63	9.419	.382	<u><0.001</u>
Kincaid	14.803	1.119	<u><0.001</u>	.572	14.273	1.958	<u><0.001</u>	0.003	15.127	2.24	.572	0.003
Fog	18.448	1.292	<u><0.001</u>	.542	17.885	2.088	<u><0.001</u>	0.013	18.66	2.435	.542	0.013
Lexical Density	0.64494	0.01785	<u><0.001</u>	<u><0.001</u>	0.62566	0.02382	<u><0.001</u>	<u><0.001</u>	0.60675	0.02962	<u><0.001</u>	<u><0.001</u>
Parse Tree Depth	10.35271	0.62141	.002	<u><0.001</u>	10.74029	1.09183	.002	.115	11.00226	1.39904	<u><0.001</u>	.115
Subordination	0.44691	0.09917	0.045	<u><0.001</u>	0.4866	0.16898	0.045	<u><0.001</u>	0.62159	0.27092	<u><0.001</u>	<u><0.001</u>
Passives	0.27079	0.08717	<u><0.001</u>	<u><0.001</u>	0.19124	0.07502	<u><0.001</u>	.671	0.18627	0.09967	<u><0.001</u>	.671

One of the most salient outcomes of these models is that there are highly significant differences between the genres' language in many cases, albeit with a some variability to the size of the differences, i.e. how meaningful they are. This section will explore what those differences are.

All but two of the variables examined see two (sub-)genres cluster together and differ significantly from a third. The exceptions are lexical density and subordination, for which each (sub-)genre differs significantly from the other; in both cases, the distance in means between genres is fairly balanced. Strikingly, which (sub-) genres cluster together in showing no significant difference from each other differs per outcome variable. For the formulae, the sustainability reports and their accompanying LtSs show no significant difference ($p = p > .382$), but are significantly different from the LtSs from financial reports ($p = < 0.013$). For the syntactic variables, sustainability reports show the most shallow parse trees (and fewest subordinators) but the greatest use of passives (with a relative increase of 40% over the latter) as well as the highest lexical density. As such, outcomes are mixed; lexical density and passivisation suggest lower reading ease for report text as opposed to letters, while the syntactic complexity measures suggest greater reading ease.

As the above outcomes explore how sustainability content relates to financial content in terms of difficulty, they allow us insight into H1 (i.e. sustainability content being more readable). Despite the variability in how the three (sub-)genres interrelate based on the different variables, we see no cases in which the LtSs from financial reports are less readable than both types of sustainability content, and only a single case (lexical density) where it exhibits significantly lower readability than those letters from sustainability reports. Furthermore, this single instance of financial LtSs being more difficult than their counterparts from sustainability reporting occurs in one of the more difficult to interpret variables in terms of its linear association with reading difficulty. Conversely, LtSs from sustainability reports appear significantly more difficult than those from financial reporting when considering formula-based readability and extent of subordination, staggeringly more so in the latter's case. In summary, we are unable to accept H1, and find more evidence for the opposite scenario: based on the comparison between the two types of LtSs, sustainability content appears to be less readable than financial report content.

The comparison with sustainability report body text does offer a measure of nuance in this respect. Sustainability reports and their accompanying LtSs are consistently least readable according to the readability formulae (with the LtSs from sustainability reports slightly less readable still, but the difference is not significant). Sustainability report body text ranks lowest (i.e. potentially most readable) on the Parse Tree Depth and Subordination features. Although it ranks highest (and thus least readable) on the Lexical Density and Passives features, these results are sufficiently mixed that we can neither claim that financial content for the same company is consistently more difficult to read

in all respects, nor that sustainability content is always more difficult to read. For LtSs, although those from sustainability reports consistently (with the exception of Lexical Density) score the less readable result than their counterparts from financial reports, the difference is not consistently significant.

There are a number of plausible explanations for some of the more salient between-genre differences we observe. For instance, the substantially higher rate of passivisation between sustainability report body text and either type of LtS is likely due to the discursive conventions of both genres. Corporate reports tend to provide a narrative overview of past events and future expectations, which is inherently conducive to use of the short passive (e.g. 'Energy use has been monitored and will continued to be monitored') as the implicit agent (the company) remains fairly consistent. The LtS, given a more directly identifiable (group of) author(s) in the signatory or signatories, will see a greater incentive for the text to engage directly with the reader, especially as it often serves as an introduction to the report. Composers can optimise the typical salutation, closing signature, photograph of the CEO and overall tone in order to create the greatest possible rapport with the reader; personal pronouns ('I am proud to present...') can serve to enhance that tone and engagement. As LtSs may also see more frequent changes in the (implied) agent, for instance between the CEO and the company, they may be less welcoming to the passive voice. Chapter 5 explores agency framing in LtSs in greater detail, and Chapter 6 delves deeper into tone and engagement.

We might similarly attribute the higher lexical density of sustainability report body text to that third noun – 'body'. As the Letter to Shareholders/Stakeholders serves as an introduction, it will logically require a slightly higher budget of function words to establish the initially relevant relationships. As we can expect the reports' body text to have a higher informational density, we can also expect a commensurate lexical density: not only do sustainability reports have the highest mean lexical density at 64.5%, but they also show the lowest variability (SD of 1.9% compared to 2.4% and 3% for financial and sustainability-related LtSs respectively).

In summary, a balanced between-genres comparison of corporate reporting readability yields the following observations:

1. Judging by readability formulae, LtSs from financial reports appear systematically and significantly more readable than sustainability reports' LtSs or their body text, although not universally so.
2. Judging by syntax, it is difficult to discern any systematic pattern of readability variation between the three genres, as sustainability report body text sometimes exhibits the lowest reading ease (in the case of lexical density and passives) and sometimes the highest (in the case of subordination and parse tree depth). LtSs from financial reports never exhibit the lowest reading ease, although their scores are not always significantly different from the most difficult (sub-)genre (i.e. in the case of parse tree depth).

3. In some cases (such as higher rate of passivisation and higher lexical density in sustainability report body text), the divergences appear intuitively attributable to genre traits and aims. In other cases (such as the higher rate of subordination in LtSs from sustainability reports), this is less so.
4. The above combines to contradict the hypothesis that sustainability content, given its wider potential audience, will be more readable (H1). Conversely, it appears (although not entirely consistently, given the outcomes for syntactic features) that sustainability content is most often more difficult to read than financial content. As the corpus does not contain directly matchable financial report body text, this is not, on its own, conclusive evidence that sustainability content is systematically less readable than financial content.

It is also notable that Curtis (1998) reports a mean FRE score of approximately 46.225 with a standard deviation of approximately 12.65 for excerpts from Letters to Shareholders taken from reports issued by both well- and poorly-performing Stock Exchange of Hong Kong-listed companies between 1994 and 1995. This implies a highly significant ($p < 0.001$) difference with both the LtS from financial and sustainability reports present in this study's corpus. While we must take considerable caution in comparing results across fiscal years and regions (Curtis indicates that issuing reports in non-native languages should yield additional attention to readability), these results nevertheless indicate that it is possible to issue more readable LtSs than those present in this corpus, although those Curtis analysed are nevertheless fairly difficult.

Curtis (1995) also finds mean Flesch scores of 30.72 and 27.7 for a set of footnotes from annual reports issued by Hong Kong listed companies in 1986 and 1991, respectively. While Curtis does not communicate a standard deviation for these means, the means themselves suggest that the sustainability report body text in this study is not meaningfully easier to read than more comparable text from financial reporting might be. Although a four-part, balanced comparison between financial and sustainability-related LtSs and report body text for the same companies would further strengthen this observation, these outcomes again reinforce the notion that rather than accommodating its wider audience's readability requirements, sustainability content is likely no more readable, and may even be less readable than financial content. Given the frequent highlighting of the wider audience (see section 1.4.6), we must conclude that if companies are trying to address those with less expertise or even command of the language than their shareholders, they are likely failing to do so meaningfully.

3.4.2 Fine-grained Analysis of Company Characteristics

Building on the preceding sections, we can now examine in finer detail the effects of various company characteristics on each of the readability measures with the aim of testing hypotheses 2 and 3 and gaining a better idea of how readability can vary within

the same (sub-)genre. To achieve this, we build a separate set of general linear models for each genre. Each individual model predicts one readability metric (the dependent variable) within that genre. The independent variables, retrieved from Thomson Reuters ASSET4 database through Datastream, are:

- **Region**
 - Whether the text (company) is listed in the US, UK, Europe, Australia or India. Operationalised as a single value per data point (text) containing 'US', 'UK', 'Europe', 'Australia' or 'India' labels.
- **Industry**
 - Whether the text (company) operates within the Apparel, Semiconductors, Mining/Metals or Oil/Gas industries. Operationalised as a single value containing 'Mining', 'Oil', 'Apparel', or 'Semiconductor' labels.
- **Interaction between region and industry**
 - We first built each model with the interaction between region and industry included as a predictor variable. When it proved at least marginally significant ($p \leq 0.1$), we use the model including the interaction between region and industry. When it did not, we generated a new model that excluded this predictor variable.
- **Environmental Performance**
 - Ranging from a minimum of 0 to a maximum of 1.
- **Social Performance**
 - Ranging from a minimum of 0 to a maximum of 1.
- **Governance Performance**
 - Ranging from a minimum of 0 to a maximum of 1.
- **Economic Performance**
 - Ranging from a minimum of 0 to a maximum of 1.
- **Company size**
 - Total company assets; used as a control measure (based on e.g. Rutherford 2003 and Li 2008).

Table 11 Summary of predictors' significance and effect size for general linear models predicting readability metrics for sustainability reports. Region * Industry interaction present if at least marginally significant. Dependents in rows; independents in columns. Marginal significance (p <= 0.1) in *italics*, significance (p <= 0.05) in ***bold italics*** and strong significance (p <= 0.01) in ***underlined bold italics***.

Genre	Independent	Variance Explained		Region		Industry		Region * Industry		Environmental		Social		Governance		Economic		Size	
		R ²	R ² (adj.)	p	Part. Eta ²	p	Part. Eta ²	p	Part. Eta ²	p	Part. Eta ²	p	Part. Eta ²	p	Part. Eta ²	p	Part. Eta ²	p	Part. Eta ²
Sustainability Reports (n = 157)	Flesch	.214	.146	.003	.108	.308	.025	/	/	.149	.015	.135	.016	.31	.007	.249	.009	.643	.002
	Kincaid	.138	.064	.104	.053	.809	.007	/	/	.459	.004	.51	.003	.389	.005	.137	.016	.776	.001
	Fog	.176	.105	.11	.052	.682	.011	/	/	.351	.006	.559	.002	.33	.007	.096	.02	.907	.0
	Lexical Density	.293	.232	<0.001	.252	.326	.024	/	/	.486	.003	.152	.015	.926	.0	.998	.0	.732	.001
	Parse Tree Depth	.144	.07	.063	.061	.444	.019	/	/	.299	.008	.213	.011	.883	.0	.056	.026	.687	.001
	Subordination	.19	.121	.014	.085	.545	.015	/	/	.059	.025	.006	.054	.244	.01	.775	.001	.451	.004
	Passives	.352	.296	<0.001	.187	.664	.011	/	/	.913	.0	.586	.002	.748	.001	.064	.024	.211	.011

Table 12 Summary of predictors' significance and effect size for general linear models predicting readability metrics for letters to shareholders from financial (annual) reports.

Region * Industry interaction present if at least marginally significant. Dependents in rows; independents in columns. Marginal significance ($p \leq 0.1$) in *italics*, significance ($p \leq 0.05$) in ***bold italics*** and strong significance ($p \leq 0.01$) in ***underlined bold italics***.

Genre	Independent	Variance Explained		Region		Industry		Region * Industry		Environmental		Social		Governance		Economic		Size	
		Dependent	R ²	R ² (adj.)	p	Part. Eta ²	p	Part. Eta ²	p	Part. Eta ²	p	Part. Eta ²	p	Part. Eta ²	p	Part. Eta ²	p	Part. Eta ²	p
Financial LTSS (n = 217)	Flesch	.146	.056	<i>.056</i>	<i>.047</i>	.123	.03	<i>.063</i>	<i>.074</i>	.864	.0	.27	.006	.474	.003	.994	.0	.683	.001
	Kincaid	.135	.044	.216	.03	.596	.01	<i>.035</i>	<i>.082</i>	.91	.0	.507	.002	.685	.001	.624	.001	.671	.001
	Fog	.159	.071	.284	.026	.448	.014	<i>.029</i>	<i>.085</i>	.833	.0	.376	.004	.518	.002	.577	.002	.81	.0
	Lexical Density	.222	.175	<u><i>≤0.001</i></u>	<i>.151</i>	.081	.033	/	/	.832	.0	.318	.005	.622	.001	.69	.001	.451	.003
	Parse Tree Depth	.141	.089	<i>.002</i>	<i>.08</i>	.985	.001	/	/	.86	.0	.886	.0	.419	.003	.172	.009	.781	.0
	Subordination	.093	.038	<i>.04</i>	<i>.049</i>	.28	.019	/	/	.79	.0	.192	.009	.23	.007	.508	.002	.808	.0
	Passives	.165	.114	<i>.001</i>	<i>.088</i>	.848	.004	/	/	.395	.004	.993	.0	.327	.005	.169	.01	.781	.0

Table 13 Summary of predictors' significance and effect size for general linear models predicting readability metrics for letters to stakeholders from sustainability reports.

Region * Industry interaction present if at least marginally significant. Dependents in rows; independents in columns. Marginal significance ($p \leq 0.1$) in *italics*, significance ($p \leq 0.05$) in ***bold italics*** and strong significance ($p \leq 0.01$) in ***underlined bold italics***.

Genre	Independent	Variance Explained		Region		Industry		Region * Industry		Environmental		Social		Governance		Economic		Size	
		Dependent	R ²	R ² (adj.)	p	Part. Eta ²	p	Part. Eta ²	p	Part. Eta ²	p	Part. Eta ²	p	Part. Eta ²	p	Part. Eta ²	p	Part. Eta ²	p
Sustainability LtSs (n = 88)	Flesch	.308	.194	.001	.213	.181	.064	/	/	.071	.044	.068	.045	.353	.012	.305	.014	.365	.011
	Kincaid	.248	.125	.045	.124	.435	.036	/	/	.076	.043	.139	.03	.173	.025	.499	.006	.32	.014
	Fog	.259	.137	.033	.132	.37	.042	/	/	.082	.041	.118	.033	.142	.029	.363	.011	.3	.015
	Lexical Density	.285	.167	.027	.138	.08	.088	/	/	.285	.016	.141	.029	.953	.0	.044	.054	.891	.0
	Parse Tree Depth	.165	.028	.712	.028	.383	.041	/	/	.577	.004	.809	.001	.33	.013	.77	.001	.58	.004
	Subordination	.183	.048	.497	.045	.058	.097	/	/	.517	.006	.047	.053	.749	.001	.464	.007	.538	.005
	Passives	.292	.176	.031	.134	.358	.043	/	/	.224	.02	.088	.039	.915	.0	.1	.037	.215	.021

Cursory examination of the above tables reveals that the region variable is most often a significant predictor, and typically has the highest effect size out of any predictor variable in the model, frequently exhibiting medium or even approaching (in the case of sustainability reports' lexical density or sustainability-oriented LtSs' FRE score) a strong effect. No other predictor variable in the corpus exceeds or even approaches the threshold for a medium effect size. Including the two cases where the region variable's interaction with the industry variable exceeds the threshold of significance, the region variable is significant in 15 out of a potential 21 cases. Those two additional cases are also the only ones where the industry variable is (a component of) a statistically significant predictor variable. The size variable, as we might expect given its inclusion as a control, never demonstrates a significant association with readability.

In contrast with the region variable's frequent instances of significance, out of four performance measures we see significant ($p \leq 0.05$) effects in three cases out of a potential 84, with an additional 9 cases of marginal significance ($0.05 < p < 0.1$). Judging by those three cases below the alpha level, we might infer that these are most likely false positives; a finding of only 3.57% (3/84) significant cases is difficult to attribute to more than random variation in the sample. The nine additional marginally significant cases might present a slightly more persuasive argument in favour of an association between corporate performance and readability, and suggest that a larger-scale inquiry still may find *some* meaningfully significant effect. However, the differences in effect sizes alone (with all three significant cases for performances closer to a small than medium effect) indicate that, out of those examined, region is by far the most influential variable on texts' readability. As this is, to our knowledge, the first comprehensive study to consider language variety as a significant predictor for the readability of corporate communications, this is a highly salient result.

In summary, these findings lead us to accept H2 and, to a far more limited extent, H3, in that industry only appears to have a significant predictive effect in its interaction with region; that is, while there are no discernible linguistic differences between industries globally, there may be industry differences within the same region. Given the relative inconsistency of this effect between independent variables, no plausible explanation immediately suggests itself.

However, we cannot accept H4 and its sub-hypotheses; contrary to the majority of findings (e.g. Curtis 1998, Rutherford 2003 or Bayerlein 2010) on financial reporting, these financial LtSs and sustainability reports do not appear to exhibit any form of systematic obfuscation based on corporate performance. While these results do not outright disprove the presence of obfuscation behaviour in the corpus, they do cleave closer to those outcomes of previous studies (e.g. Clatworthy & Jones 2003, Kumar 2014) that found no evidence for the admittedly contested hypothesis. More notably, perhaps, we see less evidence for an association between syntax and environmental performance

than we might have done based on Cho, Michelin & Patten (2010), although it is a marginally significant (i.e. $0.05 < p \leq 0.1$) predictor for sustainability-themed LtSs' formula-based readability.

Given the potential implications of both these findings, the following section discusses both in greater detail, and in the case of language variety supplements the above analysis with per-genre post-hoc analyses for the various readability measures.

3.4.2.1 Language Variety

To gain a more fine-grained understanding of how the region variable associates with readability, we conducted a post-hoc analysis of the different regions in the corpus for every model where region (or its interaction with industry) displayed a significant association with the dependent readability variable. To adjust for the increased risk of false positives during multiple comparisons, we apply the Bonferroni confidence interval adjustment to the subcategories' p-value, bearing in mind its sometimes overly conservative corrections to p-values (i.e. increased risk of false negatives).

The tables below, organised per genre, display the region's mean value for the dependent variable on the diagonal, while the rest of the table reports the p-value of the difference between those means. Tables for the significance of the interaction between region and industry instead report means and standard deviations. These tables only report those dependent variables where the region variable or its interaction with the industry variable crossed the threshold of significance in the preceding models. Appendix 1 includes a full overview of post-hoc analyses, even where the independent variable did not show a significant association with the dependent.

Sustainability Reports

Table 14 Post-hoc analysis for region variable where significant (FRE, Lexical Density, Subordination and Passives) for sustainability report readability. Means for region on diagonal; significance of difference between those means in rest of table. Bonferroni correction applied.

Flesch Reading Ease Score					
Region (p)	Australia	Europe	India	UK	USA
Australia	23.431	0.009	1	0.008	0.833
Europe	0.009	29.85	1	1	0.499
India	1	1	27.746	1	1
UK	0.008	1	1	29.182	1
USA	0.833	1	1	1	26.433

Lexical Density					
Region (p)	Australia	Europe	India	UK	USA
Australia	64.60%	0.193	0.964	0.002	1
Europe	0.193	63.30%	0.002	1	0.002
India	0.964	0.002	66%	0.002	1
UK	0.002	1	0.002	62.70%	<.0001
USA	1	0.002	1	<.0001	65.20%

Subordination					
Region (p)	Australia	Europe	India	UK	USA
Australia	0.445	1	0.086	1	1
Europe	1	0.43	0.05	0.782	1
India	0.086	0.05	0.289	0.005	0.057
UK	1	0.782	0.005	0.494	1
USA	1	1	0.057	1	0.45

Passives					
Region (p)	Australia	Europe	India	UK	USA
Australia	0.315	1	1	1	<.0001
Europe	1	0.3	1	1	<.0001
India	1	1	0.302	1	0.096
UK	1	1	1	0.283	0.003
USA	<.0001	<.0001	0.096	0.003	0.202

For the FRE, Australian reports appear least readable on average, with a significant distance between them and the European and UK reports, which occupy the most readable end of the spectrum. US and Indian reports occupy the middle range, with an insufficiently strong difference with any of the other regions to be significant after Bonferroni correction. These results are roughly consistent with those for lexical density,

where Europe and the UK again show the most readable results, with the UK more readable still, but the difference between the two statistically negligible.

The other two variables, subordination and rate of passivisation, however, show an entirely different pattern. In the case of subordination, the Indian reports display a markedly lower rate. While we have previously (see section 2.4.2.2) explored how rate of subordination need not linearly correlate with reading difficulty, we can estimate that the wide gap between extent of subordination in Indian reports (less than 0.3) and all other regions' (more than 0.4) is likely to impact perception and experience of these reports' language, for better or worse. We note that, in spite of how large this distance between India and other regions is with respect to subordination, the difference is not consistently significant, almost certainly due to a combination of the low (sub-)sample size for Indian reports and the caution inherent to Bonferroni correction. Such syntactic differences between Indian English and other varieties are, however, consistent with Sailaja (2012), who asserts that, while Indian English differs less from other varieties in terms of syntax than, for instance, vocabulary or phonology, there are nevertheless differences in verb patterning and structural influences from the linguistic substrate as well as cultural requirements. However, Sailaja also indicates the exact nature of these differences remains underexplored. As India is a highly linguistically diverse region with starkly different potential substrates, we see sufficient reason not to conjecture any further on these results, but instead recognise the potential of further analyses into within and between-regional linguistic patterns in specialised genres such as these. Language modelling (see section 6.6.2) can add great value in this respect given the aforementioned relative lack of corpus research.

Results for passivisation, in turn, exhibit strong evidence of Precht's assertion that American English is more active than other varieties (Precht 2003a, 2003b). American reports show a substantially lower number of passives per sentence, with highly significant differences with all but the Indian reports, the latter again almost certainly due to the low (sub-)sample size interacting somewhat adversely with Bonferroni correction. The other varieties do not differ significantly from one another with respect to passivisation.

Financial Letters to Shareholders

Table 15 Means and standard deviations for Flesch-Kincaid Grade Level and Gunning Fog Score for financial LtSs, divided by region and industry.
 ‘/’ indicates empty or unavailable subsets.

Flesch-Kincaid Grade Level: Region * Industry								
	Apparel		Mining		Oil		Semiconductors	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Australia	/		15.235	1.956	14.676	1.282	15.081	/*
Europe	13.983	2.594	13.432	1.877	15.768	2.142	15.426	2.59
India	/		13.726	1.697	12.867	1.526	/	
UK	/		14.872	1.211	13.969	1.888	13.633	2.973
USA	13.831	3.313	14.854	1.468	13.901	1.957	14.643	1.862

Gunning Fog Score: Region * Industry								
	Apparel		Mining		Oil		Semiconductors	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Australia	/		19.12	2.1	18.545	1.669	20.104	/*
Europe	17.341	2.948	16.877	1.965	19.523	2.393	19.274	2.463
India	/		17.184	1.75	16.948	1.283	/	
UK	/		18.654	1.297	17.647	1.918	17.222	2.247
USA	17.25	3.41	28.471	1.677	17.3111	2.276	18.169	2.16

* No standard deviation available as there is only a single Australian semiconductor company present in the corpus.

Table 16 Post-hoc analysis for region variable where significant (Lexical Density, Parse Tree Depth, Subordination and Passives) for financial LtSs readability. Means for region on diagonal; significance of difference between those means in rest of table. Bonferroni correction applied.

Lexical Density					
Region (p)	Australia	Europe	India	UK	USA
Australia	62.5%	0.186	1	1	0.115
Europe	0.186	61%	0.014	1	<.0001
India	1	0.014	64.1%	1	1
UK	1	1	0.24	61.6%	0.003
USA	0.115	<.0001	1	0.003	63.8%

Parse Tree Depth					
Region (p)	Australia	Europe	India	UK	USA
Australia	11.357	1	0.27	1	0.005
Europe	1	11.014	0.764	1	0.812
India	0.27	0.764	10.169	0.758	1
UK	1	1	0.758	11.144	0.275
USA	0.005	0.812	1	0.275	10.484

Subordination					
Region (p)	Australia	Europe	India	UK	USA
Australia	0.543	1	0.626	1	1
Europe	1	0.583	0.098	1	0.891
India	0.626	0.098	0.377	0.412	1
UK	1	1	0.412	0.564	1
USA	1	0.891	1	1	0.497

Passives					
Region (p)	Australia	Europe	India	UK	USA
Australia	0.236	1	1	1	0.001
Europe	1	0.228	1	1	0.058
India	1	1	0.234	1	0.808
UK	1	1	1	0.231	0.029
USA	0.001	0.058	0.808	0.029	0.17

Regarding the interaction between region and industry for the FKGL and GF, we can observe substantial variability, with for instance FKGL results for Australian Mining and European Oil and Semiconductors more than a standard deviation removed from the most readable results, such as those of the Indian oil industry. Unsurprisingly, given their shared reliance on text and word length, these outcomes extend into the GF results. While interactions between independent variables require more care in interpretation, these

results do indicate that the industry variable can complement the region variable in predicting readability; they also contrast with the results for sustainability report body text, where Europe and the UK displayed the higher formula-based readability.

For the syntactic variables, we can again observe the UK and Europe exhibiting the lowest lexical density, this time with Europe slightly lower still than the UK, although with differences only significant between both regions and the US, as well as India and Europe. Parse Tree Depth shows a significant difference only for the greatest distance, i.e. that between Australia and the US, which display the shallowest and deepest parse trees, respectively. Subordination again shows markedly lower results for India, although with none of the differences significant due to the aforementioned restrictions of this analysis. Passives, finally, reveal the same markedly more active (although less consistently significant) tendencies for US Letters to Shareholders as we found for sustainability reports.

We see remarkable variation within the behaviour of different readability metrics as, while all potentially able to inform a text’s readability, they measure different aspects of the language. Section 6.6.4 explores in greater detail how these metrics relate to perceived readability; as each metric is able to contribute more information to readability prediction, that also implies they can behave differently.

Sustainability-related Letters to Stakeholders

Table 17 Post-hoc analysis for region variable where significant (FRE, FKGL, GF, Lexical Density and Passives) for sustainability-related LtS readability.. Means for region on diagonal; significance of difference between those means in rest of table. Bonferroni correction applied.

Flesch Reading Ease Score					
Region (p)	Australia	Europe	India	UK	USA
Australia	32.121	1	1	1	0.108
Europe	1	34.004	1	1	0.021
India	1	1	37.56	1	0.363
UK	1	1	1	34.458	0.003
USA	0.108	0.021	0.363	0.003	23.813

Kincaid Grade Level					
Region (p)	Australia	Europe	India	UK	USA
Australia	14.854	1	1	1	1
Europe	1	14.269	1	1	0.342
India	1	1	12.396	1	0.275
UK	1	1	1	14.205	0.123
USA	1	0.342	0.275	0.123	15.964

Fog Score					
Region (p)	Australia	Europe	India	UK	USA
Australia	18.142	1	1	1	1
Europe	1	17.63	1	1	0.324
India	1	1	15.398	1	0.195
UK	1	1	1	17.533	0.103
USA	1	0.324	0.195	0.103	19.475

Lexical Density					
Region (p)	Australia	Europe	India	UK	USA
Australia	61.3%	1	1	0.103	1
Europe	1	59.4%	1	1	1
India	1	1	60.9%	1	1
UK	0.103	1	1	58.7%	0.126
USA	1	1	1	0.126	61%

Passives					
Region (p)	Australia	Europe	India	UK	USA
Australia	0.211	1	1	1	0.765
Europe	1	0.216	0.545	1	0.552
India	1	0.545	0.106	1	1
UK	1	1	1	0.213	0.363
USA	0.765	0.552	1	0.363	0.15

For sustainability-themed LtSs taken from sustainability reports, European and UK text again displays relatively higher readability; although Indian letters are most readable overall, results are predictably less significant for this region, to the point of never exceeding the threshold. Again likely due to the Bonferroni correction's potentially overzealous adjustment, we find US letters less readable based on formula, but only significantly so compared to UK letters (by FRE). We see similar outcomes, i.e. potential overcorrection, for passivisation and lexical density. In other words, differences seem least pronounced between regions in sustainability-related LtSs; this may, at least in part, be due to sustainability-themed LtSs being the least represented (sub-)genre in the corpus.

Overall, the most remarkable observations between varieties are the markedly lower number of passives per sentence in US texts, and the apparently lower number of subordinators per sentence in Indian texts. Although the latter was not consistently significant, we can likely consider the Indian documents a worst-case scenario in terms of significance, due to the low size of the (sub-)sample and the caution inherent to Bonferroni correction. In the case of the European and UK documents, we found

occasional instances of them trending towards the more readable end of the spectrum in terms of formula-based readability and lexical density.

The American texts' tendency towards significantly fewer passives, aside from lending additional evidence to Precht's (2003a, b) findings that American English is more active and direct than British English, is one of the more salient areas where the cross-regional reading experience of these reports may differ considerably based on audience expectations. Given Precht's findings and ours, we might expect a British reader of an American report to find it too direct to be credible as professional communication, for instance. Conversely, an American reader of a British report might find it too indirect and, consequently, potentially evasive. The same may be true of the markedly lower extent of subordination we can observe in Indian report text compared to other regions: readers from other regions used to a different variety of English may find them less cohesive or, conversely, easier to read due to a lower cognitive load. The latter scenario, however, might also cause cross-regional readers to perceive them as less professional if they expect a certain level of complexity as part of the corporate voice. Chapter 4 examines the effect of lower-complexity (higher-readability) reporting on reader perception of the company. Further research may wish to examine to what extent syntactic differences such as use of passives affect cross-regional users of corporate communication in particular.

3.4.2.2 Corporate Performance

We drew on Thomson Reuters' ASSET4 database for aggregate performance measures for the four pillars of financial, environmental, social and governance sustainability. Section 1.3 outlined the key components of each of these performance scores.

Contrary to what many previous investigations of the obfuscation hypothesis (e.g. Curtis 1998, Rutherford 2003, Bayerlein 2010) might have led us to assume, we found no statically significant evidence of obfuscation based on any of the performance measures beyond the number of false positives random chance would suggest given an alpha level of .05. However, the relatively higher number of marginally significant ($0.05 < p \leq 0.1$) indicates that a larger-scale study might find *some* association, although it would presumably be a smaller one than that with the region variable.

Although there is no definitive consensus on whether corporate reporting exhibits obfuscation, it seemed more plausible to expect at least some significant associations between performance and readability especially because sustainability reporting involves more aspects of performance than financial reporting does. However, the multitude of performance aspects may be the very reason that we see virtually no evidence of obfuscation. The presence of multiple areas of corporate performance may dilute the effect of any one pillar to the point of no longer being significant – as the relatively higher number of marginally significant associations may indicate.

Conversely, at least for the sustainability content, the explanation may lie in the voluntary nature of sustainability reporting. Companies are often under more external, for instance legal, pressure to disclose financial results, typically through financial annual reports. As these reports face considerably more regulation and must disclose certain information, even if doing so would disadvantage the company, obfuscation as well as defensive attribution can be highly desirable impression management tools, especially when reporting undesirable outcomes. Companies that choose to disclose sustainability outcomes, which are typically voluntary disclosures, may have more room to manoeuvre in terms of impression management than they would for financial outcomes due to that voluntariness. They may thus be less likely to rely on obfuscation or defensive attribution techniques, as they are more able to outright omit undesirable outcomes. It is likelier that the above two factors affect impression management simultaneously, with sustainability content relying less on obfuscation and defensive attribution techniques, and the different pillars of performance diluting incentive to obfuscate or defensively attribute based on any one performance pillar.

Nevertheless, neither scenario, nor the combination thereof, explains the lack of obfuscation or attribution behaviour in LtSs from financial reports, so there must be another (set of) factors influencing the lack of a significant association between performance and readability in this corpus. Outside of the straightforward notion that obfuscation may be a far less prominent impression management technique than some previous studies concluded, the lack thereof even in financially-themed LtSs may still be explained by the more introductory and general nature of LtSs as a genre, or by the independent variables this study used to build its linear models.

We have already discussed some of the features that set apart LtSs in section 1.4.2. Foremost, it is an introductory genre: it highlights the most pertinent information that the full report will explore in greater detail. As such, LtSs will, like voluntary sustainability reporting, enjoy more opportunities than full, obligatory financial disclosures might to present only that information the composer(s) of the report would consider desirable to report. While desirable information to report does not equate to only positive news (consider, for instance, the common scenario of LtS outright acknowledging areas of weaker performance along with potential solutions), companies can choose which information to underline in the LtS and which to relegate to the full report. Following this logic, LtSs may provide less incentive to obfuscate as composers have more opportunity to curate the letters' contents in an earlier stage of composition – before they need to apply obfuscation techniques to unfavourable information. As Cho, Michelon and Patten (2012b, p. 34-35) put it,

“the choice to issue a stand-alone sustainability report is, in itself, an impression management strategy. Once that choice is made, firm-specific factors may play only a small role, if any, in the use of report-specific tools for influencing user impressions.”

While Cho, Michelon & Patten here refer specifically to standalone reporting, their logic can extrapolate to most documents present in our sample. Moreover, from an impression management point of view, composers of LtSs may choose to include precisely those cases of unfavourable information they can use to build engagement and rapport by simultaneously discussing their plan of action. In such cases, applying obfuscation techniques may detract from impression management potential rather than optimising it.

Finally, the set of independent variables the study uses may also diminish the effect and significance of any one variable in the model. Pillars of corporate performance cannot exist wholly independently of one another: a company concerned with its social impact will likely extend that concern to its environmental impact and vice versa. A company with higher financial margins can devote more of those margins to ensuring sustainable operations, but, conversely, a company focused on sustainable business may face opportunity cost in its financial decision-making. In other words, dividing corporate performance into four pillars is a useful exercise in exploring the wider scope of sustainable performance and sustainability reporting, but may come at the cost of focus – and thereby explanatory power – compared to only considering a (more strongly aligned set of) financial performance measure(s).¹

3.5 Conclusions

This section's main research aim was to describe the readability of corporate sustainability reporting and compare it to more financially oriented content. Furthermore, it aimed to ascertain the influence that language variety and performance had on the readability of this content (in the latter case attempting to test the 'obfuscation hypothesis'). While it obtained some potentially interesting outcomes outside of the above aims, it did not specifically aim to advance theoretical reasoning on

¹ In order to ascertain the potential of a per-topic analysis (e.g. examining text on environmental performance separately from social, financial or governance performance), we used sentiment annotations from Chapter 5 to determine the top quartiles of sustainability-themed LtS in the relative attention they paid to financial, social or environmental performance. These showed no meaningful pattern in performance for the relevant topic being more informative; most notably, while environmental performance is almost significant ($p = 0.057$) as the sole independent variable predicting environment-focused letters' FRE score, it is also almost significant ($p = 0.074$) in a model predicting FRE for financially-focused letters that integrates all performance scores. As such, results are very inconsistent. While this is a very coarse exploration, it fails to provide a strong argument in favour of topic-based obfuscation research.

corporate sustainability. Instead, following from this broad-scope exploration of the genre, the following chapters will continue to investigate the genre's language through more sophisticated means. They will query what the genre's readability means for readers' perception, what the genre looks like on a more fine-grained rhetorical level (especially with regards to positivity and negativity), and to what extent we can optimise readability analysis for (sustainability) reporting content.

The first conclusion from all of the above must inevitably be that the readability of corporate (sustainability) content is a result of more complex interlocking systems than its wider, less expert audience and the previous findings on the readability of corporate (financial) reporting might suggest. For one, the reasoning that sustainability content will be more readable because a company cannot make the same assumptions regarding its audience's expertise does not hold; on the contrary, sustainability-oriented LtSs for the same company and fiscal year appear to be more difficult than those discussing financial content. We might attribute these outcomes to a number of factors, including a lesser extent of regulatory control on these reports, i.e. their largely voluntary nature.

A company stands to benefit from *claiming* to present a stakeholder-inclusive narrative in a corporate voice that evokes legitimacy, even (or especially) in a text too complex for its less powerful (i.e. indirect) stakeholders to decode. The alternative would see them potentially benefit from addressing those stakeholders in a language they can understand, but at the same time incur costs from having to develop an alternative, more accessible corporate voice, and transparently and accessibly disclose potentially unfavourable information. Given those choices, we might argue that it is rational from an impression management perspective for companies to choose the former, rhetorically positioning themselves as inclusive without incurring the costs that inclusivity might entail.

In terms of impression management, however, we must acknowledge that the corpus exhibits little, if any, performance-driven obfuscation behaviour. This is in spite of Cho, Roberts and Patten's (2010) conclusion that environmental reports exhibit impression management strategies based on their performance, i.e. that "corporate environmental disclosures of poorer performing firms appear to emphasize good news, obfuscate bad news, and slant attributions to their advantage in an attempt to manage stakeholder impressions of their corporate performance" (p. 442). These are difficult texts overall, which might be partially due to attempts for sustainability content to derive legitimacy as a genre from similarity to financial content, but that difficulty does not appear to vary as performance does. One potential explanation for that contrast might be that Cho et al. examined obfuscation in environmental reports based on environmental performance; as this study attempts to capture all four pillars present in ASSET4 because all four are present in the more holistic sustainability reporting genre, we might see a diluted predictive effect for any one performance variable. Nor do these effects entirely rule out the presence of impression management; while the corpus does not exhibit variation in

impression management techniques along performance lines on a document level, Chapter 5 will explore LtSs' defensive attribution techniques through agency framing on a sentence level.

Rather than performance, or other company features such as industry or size, this study found language variety to be the most significant and strongest predictor of readability or syntactic determinants of readability. While it is also possible to divide the corpus along regulatory lines (based on Leuz, Nanda & Wysocki 2003), we find significant differences between regions that belonged to the same tier of legislative enforcement, which gives more credence to the notion of variety of English being the relevant difference. The main example of this phenomenon was the difference between passivisation in American and British English, with the former being considerably more direct as we hypothesised based on Precht (2003a, 2003b). As companies operate in an increasingly international market, and face an increasingly international audience to report to, this variation is notable in that it may well affect perception of performance when read across different varieties.

Moving forward from the broad scope that this study employed, we aim to narrow that scope and investigate several questions that arise based on these outcomes. These are:

- To what extent readability variation affects reader perception of the company amongst those with greater or lesser expertise,
- To what extent sustainability reports carry over financial reports' reputation for excessive positivity and defensive attribution, and
- To what extent we can improve on the genre-agnostic approach to readability this chapter employed by examining the (perceived) determinants of sustainability report readability and how well we can approximate human judgments of that readability.

Chapter 4

Readability Manipulation

4.1 Motivation

Chapter 3's broad-scope full corpus analysis observed and highlighted the problematic position of readability in sustainability reporting. First of all, it found a significantly lower readability for LtSs from sustainability reports than for those from financial reports. Both, however, were problematically difficult from a general readability perspective and, in many respects, failed to adhere to Plain English principles set out in such documents as the Plain English Campaign's style guide (2013) or the SEC's Plain English Handbook (1998). For instance, both types of LtSs contain long sentences and considerably more passive structures than the average for written English. Given the linguistic parameters we were able to observe for the genre, we found it highly unlikely that these letters will be as accessible as they need to be to ensure readability and understandability for the majority of the groups of stakeholders that companies may (claim to) want to address. This is in spite of such assertions from e.g. Lonmin, Total or The Adidas Group's sustainability reports for 2012, and in spite of LtSs being the most-consulted section of the report (Courtis 1998).

These findings prompt the question why these reports – or at least the letters introducing and summarising them – are not as accessible as they could and perhaps should be. One of the most straightforward potential answers is that, as the Plain English Campaign (2013, p. 2) indicates, “[writing Plain English] is not as easy as we would like to think”. Certainly, given Rutherford (2003) or Li's (2008)'s inquiries into how organisational complexity relate to the complexity of its disclosures, it may be unreasonable to expect universal readability. There may be irreducible complexity underlying what the company is trying to convey, to the point of potentially making it all but impossible to convey the appropriate nuance that a highly complex organisation

might require in very simple terms. However, The Plain English Campaign (2013, p.2) indicates – and manages to demonstrate throughout their entire handbook – that

[a]lmost anything – from leaflets and letters to legal documents – can be written in plain English without being patronising or oversimplified [and] it doesn't mean reducing the length of your message or changing its meaning.

At the very least, then, while universal understandability might be an overly idealistic goal, we can be confident that these disclosures are less readable than they could be. We have already explored a few potential reasons behind that difficulty. For one, there is the issue of mimetic isomorphism (DiMaggio & Powell 1983) that may be at play at both the level of the decision to report and the report's composition. Companies may choose to report in order to gain legitimacy from imitating leading companies. The latter will typically make the decision to engage in CSR reporting for reasons other than mimesis, such as the pursuit of accountability. Those reporting to gain legitimacy through mimesis (which we might call 'followers') may have few reasons to diverge from the genre conventions they are familiar with from the financial reporting they engage in, or the non-financial reporting practices they observe in other companies. Similarly, on a genre level, sustainability reporting might imitate its parent (or, arguably, sibling genre), financial reporting, to evoke the legitimacy and credibility the latter already commands.

There is also the potential issue of impression management– more specifically greenwashing (see e.g. Hrasky 2012, Boiral 2013) – driving companies' reporting decisions. In a voluntary setting, there is considerable incentive, especially for the aforementioned 'followers', to prioritise the *claim* of issuing CSR reports over their efficacy. That is, a company stands to make a favourable impression amongst high-power stakeholders (see e.g. Bouten 2011), such as shareholders, by claiming to engage in inclusive CSR communication. However, it can mitigate the potential costs of that process (a potentially difficult or expensive transition towards a more accessible reporting language, and potential reputational damage from transparently reporting unfavourable outcomes) by keeping those communications in the relatively impenetrable language of financial reporting. Though perhaps a little cynical, such a perspective makes it unnecessary – perhaps even undesirable – to adjust readability to a more general audience's requirements. When this logic drives companies' decision to publish difficult sustainability content, their decision to publish becomes a prototypical form of greenwashing (Laufer 2003).

Furthermore, as Chartprasert (1993) observed increased credibility for informative prose when it used a bureaucratic writing style, we might expect that, due to the positive aspects of financial reporting that sustainability content appears to evoke through its design, companies might fear that simplifying sustainability reports' language may damage the reports' credibility. Although the Plain English Campaign (2013, p.2) claims that implementing plain English does not make a text "patronising" and "is not an

amateur's method of communication", there are certainly potential risks involved in flouting genre conventions.

Accordingly, this chapter investigates the extent to which the low readability attested for sustainability content in the previous chapter affects readers' perception of that content. As we observed that sustainability content, in spite of its wider potential audience, is likely insufficiently readable for that wider, less expert audience. However, we also explored how that difficulty may be a result of sustainability reporting aiming to emulate financial disclosures; as such, a lower readability – further removed from financial content – may also have adverse effects on readers' perceptions of these reports' credibility or professionalism. What is more, these changes in perception may differ between audiences familiar and unfamiliar with the genre.

Better understanding the dynamics between readability and perception also allows us a deeper insight into how problematic sustainability content's low readability is, as well as how it can be and whether it should be improved, and why this genre is as (un)readable as it is.

4.2 Hypotheses

The two main aims of this chapter are to discover the extent to which changes in a LtS's readability can influence audience perceptions, and to what extent such manipulations might affect laypersons (i.e. those unfamiliar with the genre) or non-laypersons (i.e. those at least somewhat familiar with the genre) differently. We will attempt to achieve these aims by testing a number of hypotheses.

On the most general level, based on e.g. Chartprasert (1993), we can hypothesise that:

H1: Overall, changes in complexity will alter readers' perception of the company and text.

As Lehavy, Li & Merkley (2011) find increased reliance on analysts' reports rather than financial reports as the latter's readability goes down, we can expect the reader's level of expertise to have a considerable influence on how they deal with more or less readable content:

H2: Overall, those familiar with the genre will respond differently to reduction in linguistic complexity than those not familiar with the genre.

We can also test these texts' difficulty as a form of impression management. Building on H2 and referring back to Chartprasert (1993), we can expect that the same characteristics that make the text less accessible will, for the impression management efforts to be

maximally effective, evoke positive associations in those less familiar with genre conventions:

H3: Laypersons will respond more optimistically to higher complexity

If more complex language indeed has a number of positive associations, e.g. in terms of credibility, we might expect those associations to disappear in more readable reports. As indicated in section 4.1, a major contributor to sustainability content's low readability might be a fear on the authors' behalf that an easier to read report or LtS may become less credible, or reflect less professionalism:

H4: Less complex language will reduce readers' perception of professionalism

However, a shift towards less complex language may also have the effect of further reducing readers' already diminishing perception (Townsend, Bartels & Renaut 2010) that these reports are written more out of a self-serving desire for impression management rather than out of a desire for transparency and accountability:

H5: Less complex language will increase readers' perception of transparency (e.g. honesty, trust)

Finally, it is worth investigating whether a simplified piece of corporate reporting would actually register as less complex to an audience interacting with the text. While it is likely that readers would be able to differentiate between a difficult piece of reporting and a simplified one when offered the opportunity to compare both, it becomes less likely that they would register this simplification when faced with a standalone (simplified) text. As the degree to which a text might register as easier could very well impact a company's editing decisions, we test the following:

H6: Less complex language will reduce readers' perception of linguistic complexity

Examining the validity of these hypotheses should offer us additional insight into the effectiveness of companies' current editing processes, as well as into how authors can improve sustainability disclosures for a wider audience.

4.3 Editing Process

To detect the impact that a more accessible writing style would have, we selected a Letter to Stakeholders from a British company, as we intended to poll a British audience for their impressions and wanted to minimise any adverse effects from cross-varietal linguistic biases. We aimed to create two additional versions of the base text, each representing an

increase of one ‘band’ of readability on the Flesch Reading Ease scale relative to the previous version. Specifically, as most texts within the corpus occupy the lowest, ‘very difficult’ band of 0-30, we aimed to take a text from this band, and create two versions with a higher readability: one in the ‘difficult’ band of 30-50, and one in the ‘Plain English’ band of 60-70.

We used [readable.io](#) (2017), an online readability tool, to determine the readability for each document. As, in addition to formula-based metrics, it offers a readability rating (A-E, A being best) and visually highlights textual elements that diverge from Plain English guidelines (for instance, long sentences in red and passive structures in blue) it proved a useful tool for this manipulation process.¹

We chose to use the subgenre of LtSs from sustainability reports, rather than sustainability report narratives, out of concern for length and representativeness. Most sustainability reports or chapters contain more than 2000 words, which reduces the chance that all participants will reach the end with undivided attention. We opted to present participants with full texts rather than excerpts from sustainability reports (which might have otherwise averted the issue of length) as we did not want to make our text selection less representative or introduce selection bias in delineating segments. We thought it better to leave the decision of where to segment texts to the original authors, and thus chose to use the LtSs prefacing the reports.

We selected the Letter to Stakeholders from Gem Diamonds’ 2012 sustainability report as a starting point, which [readable.io](#) places at a 13.8 FRE score, i.e. relatively close (within two points) to the middle of the ‘very difficult’ band of the FRE Index. We further expected the first paragraph, detailing workplace fatalities, to draw participants’ attention sufficiently for the text to engage and maintain interest.

We performed two editing sequences, the first more cautious with respect to the text’s original content and rhetorical strategies than the second, although each sequence brought the document closer to Plain English (SEC 1998; Plain English Campaign 2013). For each of these sequences, one researcher played the role of editor and changed the text, and another, who played the role of reviewer, compared the result with the original text in terms of rhetoric, sentiment, coherence and cohesion, and general

¹ We use a different means of readability measurement from the previously established CoreNLP-based analysis extracted through De Clercq & Hoste’s (2016) pipeline due to the very different purposes and requirements of the full-corpus analysis and this very fine-grained, sentence-level one. While the former is highly suitable to bulk processing, [readable.io](#) is much more amenable to fine-grained analysis due to its real-time updates and colour-coded style markers. It is, however, far less suited to producing quantitative analyses. The exact scores may differ with what the bulk-processing pipeline would have produced (for instance, for the original text the bulk processing pipeline indicates a Flesch score of 12.37). This is due to e.g. different syllabification techniques or other technical nuances, However, the order of readability and approximate distance between scores remain consistent between both approaches.

appropriateness of the changes. The two discussed any remarks or disagreements and sought alternatives until they agreed on what the right change was.

While personal names occurring in the text – those of three on-site fatalities and the CEO – remain unaltered, we changed the name from ‘Gem Diamonds’ to ‘Lustre Minerals’ to keep a similarly gemstone-related company name without introducing potential bias due to any familiarity the respondent may have with the company. We considered the likelihood of similar bias from the personal names included in the letter small enough that we did not alter the names to sound similar to, but be distinct from, the original names. Substituting out the original names might be more likely to introduce adverse effects than prevent them if the substitutes failed to accurately mimic the same dynamics and associations, for instance in terms of culture or ethnicity, that the actual names portray. Additionally, we did not wish to lessen the gravity of the fatalities reported by fictionalising this aspect of the events.

Finally, we attempted not to substitute references to the company or the reader out with personal pronouns in places where the Plain English guidelines might recommend doing so. While these are indeed highly likely to influence a reader’s engagement with the text (see section 6.3), we had concerns that changing agency framing might dilute the effects of altered readability because, as sections 5.7.3 and 5.8.1 will explore, the genre already exhibits fairly complex patterns of agency framing. That is, a wide-reaching change in agency patterning would make it difficult to discern whether the effects we will observe are due to changes in readability or changes in agency patterning.

The following sections describe editing process and the various choices we made in greater detail; the full versions of the three texts are available as appendices, under ‘Appendix 2.’

4.3.1 More Readable – ‘Difficult’ (30-50 FRE score)

For the first editing sequence, meant to produce the ‘difficult’ (30-50) text,² the editor focused on converting compound and complex sentences longer than 30 syllables to simple sentences wherever possible without significant loss of cohesion, coherence or rhetorical effect. This aligns with the Plain English Campaign’s (2013) recommendation of 15-20 words per sentence, but also makes some allowance for optimising the formula-based outcome. For instance:

² This is ‘difficult’ as opposed to ‘very difficult’ (0-30) on the FRE index.

Original	More readable version
We will continue to work hard to continuously improve our systems and eliminate risk in as far as is practicable in our workplace, thereby driving to achieve our target of zero harm.	We will continue to work hard to improve our systems and eliminate risk. As far as is practicable in our workplace, we strive for our target of zero harm.

We did keep coordinating or subordinating conjunctions where the connection between the two sentences was integral to their meaning. For instance:

Original	More readable version
Ensuring a safe working environment for all our employees is of primary importance to us at Lustre Minerals and it is with great sadness that we need to report that three fatal incidents occurred during 2012 at our operations.	Lustre Minerals prioritises a safe working environment for all our employees, and it is with great sadness that we need to report that three fatal incidents occurred during 2012 at our operations.

In the above case, letting ‘Lustre Minerals prioritises a safe working environment for all our employees’ terminate in a full stop might have generated more dissonance between Lustre Minerals prioritising safety and the three fatalities, or –worse still – an irreverent contrast between the two propositions. We kept the coordinator ‘and’ because what the author intends to express derives from acknowledging the contradiction between the two propositions. In this case, the coordination is part of the document’s rhetorical strategy, and aligns with the Plain English Campaign’s advice on apologies (i.e. combining directness and sympathy).

We also activated passive voice where we could, as those same guidelines suggest. While this mostly required adding an agent where the text did not specify one (such as in the above case), there was always a logical (if not necessarily specific) agent to infer. For instance:

Original	More readable version
Maintaining the highest levels of product integrity and ensuring that all diamonds recovered are certified under the most stringent ethical standards.	Maintaining the highest levels of product integrity and ensuring that those certifying diamonds we recover obey the highest ethical standards.

Where appropriate (without significant loss of meaning), we also further activated and simplified verbal constructions, often through positioning a non-human agent as the subject. For instance:

Original	More readable version
Our sustainable development framework is our response to three key business drivers.	Our sustainable development framework responds to three key business drivers.

Where appropriate, we created parallel constructions in order to aid understanding, sometimes arguably at the expense of rhetorical considerations. For instance:

Original	More readable version
<ul style="list-style-type: none"> • Retaining our social licence to operate • Continuing to attract high quality customers • Continuously improve our reputation 	We must... <ul style="list-style-type: none"> • Keep our social licence to operate • Keep attracting high quality customers • Keep improving our reputation

We also tried to find shorter synonyms to convey the meaning for difficult or long words (e.g. those with more than four syllables), and split them into more than one word when appropriate. We often preferred active verb forms. For instance:

Original	More readable version
We believe that sustainability demonstrates our adaptability to a changing socioeconomic and bio-physical environment.	We believe that sustainability shows we can adapt to a changing social, economic, biological and physical environment.

Where complexity or length of individual words were not an issue, we tried to shorten sentences (as readability formulae prefer) by using fewer words to convey the same meaning where possible. This occurred multiple times with verb phrases. For instance:

Original	More readable version
Lustre Minerals will remain focused on continually improving its performance and eliminating unacceptable risk to the business and all stakeholders	Lustre Minerals continues to focus on improving its performance and stopping unacceptable risk to the business and all stakeholders

In some cases, linguistic ‘fuzziness’ – especially in the case of adverb- and adjective-heavy constructions - made us less sure of the author’s intended meaning. Unfortunately, these same passages most required simplification to achieve the desired readability scores. When this occurred, the editors inferred the intended meaning to the best of their abilities. For instance:

Original	More readable version
Ensuring an operationally intelligent and productive workforce by implementing appropriate strategies to develop and retain our employees.	Ensuring a productive workforce that knows how we work . We must implement the right strategies to develop and retain our employees.

These changes resulted in an FRE score (as calculated by Readable.io) of 36.6. This placed the text within the desired range (within five points of the band’s average). Readable.io’s overall ‘Readability Rating’ (which is a proprietary combination of several readability features) went up from ‘E’ to ‘C’.

4.3.2 Most Readable – Towards ‘Plain English’ (60-70 FRE score)

We aimed to bring the second version as close to Plain English as possible, judging by the FRE score. We strove to eliminate sentences longer than 30 syllables (the final version

contained three) and minimise sentences longer than 20 (the final version contained seventeen). Similarly, we sought alternatives for words longer than four syllables (the final version contained seven). The target meant optimising to the formula in addition to the Plain English guidelines, which implied considerable differences with the previous version. In many cases, this meant sacrificing some nuance compared to previous versions, and resulted in a shorter document overall. For instance:

Original	More readable version	Most readable version
Conducting our business in an ethical, transparent and responsible manner, will help us retain our social licence to operate. This requires a particular focus on managing and controlling risk and consequential impacts through understanding risk drivers and how these relate to our business processes.	Doing our business in an ethical, transparent and responsible manner will help us keep our social licence to operate. We must focus on managing and controlling risk and its impacts. If we understand risk drivers and how these relate to our business processes, we will be better able to control them.	Our business must be ethical, transparent and responsible. Only then can we keep our social licence to operate. We must manage and control risks and impacts. Understanding risks helps us control them.

Or:

Original	More readable version	Most readable version
During 2012, a collective effort across all business units resulted in the conceptualisation of 'The Gem Way', that [sic.] clearly communicates our philosophy of zero tolerance and our commitment to responsible care.	During 2012, an effort across all business units together outlined 'The Gem Way', that explains our philosophy of zero tolerance and our commitment to responsible care.	During 2012, all business units outlined 'The Gem Way' together. It collects our thoughts on zero tolerance, and commits us to responsible care.

In accordance with Plain English Guidelines, this version introduces lists where a sentence contains a coordinated sequence of information. For instance:

Original	More readable version	Most readable version
We believe that sustainability demonstrates our adaptability to a changing socioeconomic and bio-physical environment.	We believe that sustainability shows we can adapt to a changing social, economic, biological and physical environment.	Sustainability means we adapt to different changes: <ul style="list-style-type: none"> - Social; - Economic; - Biological; and - Physical

Compared to the first edit, we more stringently attempted to clarify ‘fuzzy’ language where we expected that most readers might struggle to infer the author’s exact intent. Although we tried to stay as close to the meaning we inferred from the original, such changes inevitably come at the cost of the original form’s nuance and rhetoric. For instance:

Original	More readable version	Most readable version
By actively managing these material aspects in an integrated manner, we aim to minimise harm and optimise benefit.	If all parts of our company bear these key issues in mind, we can minimise harm and optimise benefit.	These points help all parts of our company do the least harm and benefit the most.

We were unable to eliminate every element that increased the text’s complexity, as not every such element had a simpler alternative. Most readability formulae consider ‘unacceptable’, as in ‘unacceptable risk’ a difficult word (which aligns with Flesch’s emphasis on affixes as increasing cognitive load), but we were unable to substitute it with an acceptable, less complex alternative. Writing ‘risks we cannot accept’, for instance, draws undue attention to the company’s decision to accept other kinds of risk. Because of this, we decided not to alter such collocations as ‘unacceptable risk’ or ‘sustainable value’ where we felt substituting them would unduly alter the sentence’s rhetoric (as ‘value that is sustainable’ elicits the undesirable notion that some value is not sustainable).

These changes resulted in a considerably different text from the original and the first edit. However, we saw diminishing returns in readability score compared to the first edit; this version yielded a Flesch Reading Ease score of 47.1, a considerable increase compared to the original version and first edit, but still thirteen points short of the ‘Plain English’ threshold. This was likely due to the aforementioned ‘irreducible’ complexity-introducing elements. The ‘Readability Rating’ did however increase to ‘A’, the highest possible category, over an ‘E’ for the original and ‘C’ for the first edit. We might interpret this as a considerable improvement over the first edit, albeit one not fully reflected in the FRE score. Again, these manipulated texts are available in the appendix, under the ‘Manipulation’ section.

4.3.3 Risks of Writing to Formulae

The latter text especially also merits a caution regarding “writing to formula[e]” (Klare & Buck 1954, p. 139). While the ‘more readable’ text with a 30-50 FRE target prioritised applying the Plain English guidelines and made a few adjustments to attain better formula-based readability, this one, which targeted a Plain English FRE score of 70 or above, attempted to minimise sentence length and word length. In doing so, we

attempted to keep respecting the Plain English guidelines, but did at times prioritise the formula over cohesion and nuance.

Van Hoecke (2018) submitted this same triplet of texts to a panel of ten respondents in a within-subjects experiment and asked them to rank them by readability. Seven out of ten ranked them in the same order of readability the formulae suggest, but two commented on a partial loss of textual flow in the ‘more readable’ text (i.e. the ‘difficult’ one). Four commented on a loss of cohesion in the ‘most readable’ one according to formulae. Van Hoecke notes the latter may be due to a greater shift of cohesive markers from the explicit due to the implicit (e.g. due to greater use of list structures), which aligns with theories on syntactic depth that section 2.4.2 explored.

These results indicated the viability of the experiment in that most of the respondents agreed with the sorting order, but they also warn against the dangers of prioritising a readability score above human assessment. While the editing process attempted to ensure that the text remained sufficiently true to the original, there is clearly some room for discussion on how successfully it achieved that. Furthermore, this outcome illustrates how readability formulae can fail to capture the full complexity of readability: while they are highly sensitive to word and sentence length, they fail to register that when conveying complex ideas in shorter sentences, cohesion must invariably shift to be more implicit, and this can negatively impact readability and understandability. That is, no part of the formula’s design allows it to capture syntactic depth or complexity, especially not between sentences – all it can capture is sentence length, which is a result of (and interacts with) far more complex syntactic dynamics.

4.4 Assembling the Questionnaire

This section first gives a short overview of, and then explores in greater detail, how we attempted to investigate a number of ways in which changes in a LtS’s language along Plain English guidelines might affect the readers’ perception, and assembled a questionnaire in order to gauge those potential shifts in perception.

In terms of design and execution, we used a professional survey company (ProFacts) in order to extend the reach of the study beyond the circles directly available to academia. We aimed to address native speakers of British English, or at least native speakers of English currently residing in the UK. This had the advantage of minimising any linguistic biases based on the variant of English with which they were more familiar. This is also why we chose a British LtS. Furthermore, we requested that the company attempt a balanced population of those not at all familiar with the genre (hereafter ‘laypersons’) and those at least somewhat familiar with the genre (hereafter ‘non-laypersons’).

Chiefly, we wanted to focus on how (and whether) changing the language affected readers' perception of:

- The text's difficulty;
- The text's sentiment (positive or negative);
- The text (and company)'s credibility and professionalism, and the author's expert status; and
- The text (and company)'s performance, both in terms of finances and other aspects of sustainability.

To achieve this, we set up a between-subjects design in which we asked respondents to read one of the three versions of the text; we did not expose them to either of the other variants. We verified their understanding of and attention to the text through two comprehension questions, and gauged respondents' familiarity with the genre by asking them to explain that familiarity where applicable. We then presented them with three sets of nine statements each, querying their impression of the company portrayed in the report (two sets of statements), and the report's composition. In addition, we queried how positive or negative they found the text, and how easy (or difficult) to read. Finally, we asked respondents for some information on themselves and their language proficiency.

The rest of this section explores the questionnaire and decisions behind it in greater detail.

4.4.1 Introduction and Informed Consent

The questionnaire first introduces its design, indicating that it will present the respondent with a chairman's letter that introduces a company's annual report. We elect not to call this an 'annual sustainability report' due to the terminological density already inherent to the sparser phrasing. The questionnaire then lays out how it is structured: the respondent first reads the text, then answers a number of questions about it, and then about themselves. Finally, it thanks the respondent for their participation.

Respondents are unable to proceed without agreeing to the following statement on informed consent (taken from Lybaert 2016):

I give permission to the researcher and any possible future researchers to use the recorded materials and written surveys for scientific research. I agree that my personal information will be processed and used, and I know that I have the right to access and correct this information. The data will be processed anonymously and my privacy will be respected at any time.

4.4.2 Reading and Comprehension Test

The respondent then reads either the unmodified text, or one of the edited versions. Which one of the three they see is random. Respondents are aware that they may refer back to the text as they proceed through the questions, as the questionnaire informs them of that fact. In order to ensure that they have read the text with some attention to detail, they receive two comprehension-based questions: first, as a basic check, the questionnaire asks for the company's name. It then asks the reader in which industry the company is active. We offered four potential answers for the latter: chemical, oil, retail or diamond industries. The correct answer was 'diamonds', although it only occurred as a keyword towards the end of the text. Given the extent of attentive reading required to answer this question correctly, we felt confident that asking the question would either ensure that the respondent read the text attentively, or, in cases where the respondent had not yet done so, indirectly oblige them to. As the outcomes will reflect, this proved less the case than we initially anticipated.

Finally, this section of the questionnaire ascertained respondents' familiarity with corporate reporting, CEO letters and corporate sustainability, respectively. This question used a four-point Likert scale of 'not at all familiar', 'somewhat familiar', 'familiar' and 'very familiar' with the topic, and asked respondents to indicate the origin of this familiarity if they answered other than 'not at all familiar'. Unlike the declaration of informed consent, the biographical data or the content questions, this Likert scale seemed straightforward enough not to have to draw on previously validated constructs.

4.4.3 Company and Composition

We then presented respondents with three sets of nine statements each and queried to what extent they agreed or disagreed with them on a seven-point Likert scale (i.e. representing integer values of +3 through -3, ranging from strong agreement to strong disagreement). The first two sets of questions queried the respondent's perception of the company – the first based on a set of adjectives, the other on statements about the company – while the final set asked the reader how appropriate they found nine adjectives describing the text's composition. Between the second set and the third, the questionnaire briefly polls how positive or negative the respondent found the text, ranging from very negative (-3) to very positive (+3), and after the third set it queries how difficult the respondent found the text to read, ranging from very difficult (-3) to very easy (+3).

For the purposes of drafting the company-related questions, we drew on Chun (2005) for a number of constructs gauging corporate reputation. These included both adjectives taken from Davies et al.'s (2003) corporate character scale (quoted In Chun 2005, p. 103)

and statements in full sentence form (Chun 2001, qtd in Chun 2005, p. 102), the latter of which especially divide fairly neatly along the four performance pillars. Chun presents these constructs as oriented towards a context in which '[companies] have to satisfy the requirements of many stakeholders', which aligns with our aims of measuring stakeholder perception rather than just shareholder perception.

We focused on assessing respondents' perception of honesty, professionalism and sustainability in the adjective-based questions and sustainability, management, profitability and overall impression in the statement-based questions.³ For the adjectives, we selected six adjectives from Chun (2005):

- Open
- Honest
- Sincere
- Responsible
- Trustworthy
- Competent

'Open', 'honest', 'sincere', 'trustworthy' and to some extent 'responsible' are meant to measure the extent to which the text presents the company as reliable and credible: does the reader believe that the company accurately and/or faithfully portrays their financial year and intentions within the text?

'Competent' and 'responsible', respectively, are meant to capture to what extent the company knows what it is doing and acts in accordance with its longer-term impact. We added a further three adjectives to the list to better explore the themes of professionalism and sustainability:

- Complex
- Professional
- Sustainable

'Complex' and 'professional' represent the negative and positive sides of corporate organization, respectively. The former might apply in cases where corporate structure or activities appear too elaborate to understand based on the text. The reader's perception of professionalism, conversely, may suffer if the text (for instance through linguistic simplifications) presents as overly simple or casual, as readers might expect to see the company's organizational complexity reflected in the text. 'Sustainable', finally, is the most straightforward means of capturing in one adjective whether the reader considers

³ We chose these areas of emphasis as Chun (2001) had previously validated full-sentence constructs available for them (in addition to other categories less relevant to this study). While we do insert adjectives not previously validated into other constructs, using untested full-sentence constructions seemed considerably more likely to detract from the study's validity.

the company portrayed in the text viable in the long term; whether linguistic changes affect readers' perception of sustainability is highly relevant given the tension between sustainability reports' low readability but self-proclaimed wide audience. A number of statements can also help capture this perception.

In order to ensure the validity of the full-sentence constructs, we integrated them directly from Chun (2005) without further alterations:

1. I have a good feeling about this company
2. I respect this company
3. I trust this company
4. This company has a clear vision for its future
5. This company is well managed
6. This looks like a company that would have good employees
7. This is an environmentally responsible company
8. This company maintains a high standard in the way it treats people
9. This looks like a company with strong prospects for future growth

Statements 1-3 can help capture (differences in) reader perception of company competence and reliability, while 4a-5 can capture perceptions of both competence and long-term sustainability. 6-8 primarily focus on aspects of sustainability, while 9 can capture both (financial) sustainability and overall perceived competence.

We asked respondents to rate how positive or negative they found the text overall on a seven-point Likert scale (ranging from 'very negative' through 'very positive') in order to ascertain whether our modifications to the text influenced readers' perception of overall sentiment or tone.

Finally, we presented the reader with nine adjective and adjective phrases concerned with the text's composition, intended to capture their perception of readability, persuasiveness and, again, professionalism. We asked them to indicate to what extent they thought the text was:

- Clear
- Readable
- Complicated
- Well-written
- Easy to understand
- Persuasive
- Accessible
- Pleasant to read
- Written by an expert

These mostly capture understandability, but attempt to cast a wide net in attempting to capture whether the text confused the reader at any point ('clear' and 'complicated', 'easy

to understand') and what their opinion of the author (and consequently, albeit tangentially, their professionalism) was ('well-written', 'written by an expert'). Two more adjectives might reflect how difficult it was to obtain information from the text ('accessible', 'readable') and one reflects to what extent the text achieved its implicit goal ('persuasive'). Additionally, it tried to capture to what extent the text was able to generate intrinsic motivation in the reader ('pleasant to read').

Finally, as a more general question intended to capture readers' perception of readability, we asked how difficult they found the text on a seven-point Likert scale ranging from 'Very difficult' to 'Very easy'.

4.4.4 Biographical Data

As a last step, we asked respondents for a number of biographical data, in order to control for them if necessary. We based these constructs on a questionnaire used in Lybaert (2016). We asked respondents for their:

- Gender, offering the following options:
 - o Male
 - o Female
 - o Neither of the above
 - o Would rather not say
- Age
- Native variety of English
- Assessment of their English proficiency on an eight-point scale ranging from 'very weak' to 'excellent'; this may influence their perception of texts' readability. We added a category of 'very good' between 'good' and 'excellent' in the original construct.
- Highest degree obtained so far
 - o We also asked whether they had an English component in their curriculum after secondary school.

4.4.5 Survey Process

The tension between sustainability reporting's broader audience and low readability creates two important variables in terms of the audience for this experiment: linguistic proficiency and expertise related to corporate reporting. As we were dealing with English-language reports, a meaningful percentage of the audience needed to be native speakers of the English language. After spreading the questionnaire through market research agency Profacts, we obtained a total of 242 responses that correctly replied

‘Lustre’ or ‘Lustre Minerals’ to the initial qualification question of replicating the company’s name. This was out of a total of 343 responses (i.e. in addition to 101 erroneous ones), suggesting some difficulty in the survey process. Respondents that did not return the correct answer were disqualified, as were the 6 respondents who were not native speakers of English, resulting in 236 total English-native responses. Appendix 2 contains the full survey.

4.4.6 Description

Amongst the English-native respondents, 152 (64.4%) self-identified their English as excellent, and a further 53 (22.5%) asserted it was very good. 19 (8%) claimed their command of the English language was good without fitting into any of the higher-ranked categories, while 6 (2.5%) assessed their English as quite good and a further 6 (2.5%) as average. No respondents evaluated their English as below average on this scale. As a strong majority called their English either very good or excellent and only 5% of respondents rated theirs below ‘good’, we can assume that command of the language itself, rather than the genre-specific register, should not be an impediment to their ability to gauge the texts. At the very least, as they answered this question after reading the text and answering readability-related questions, a lack of confidence in their own linguistic abilities is unlikely to have negatively affected their judgments.

In terms of experience with the genre, 88 (37.6%) indicated at least some familiarity with corporate reporting, 102 (43.2%) with CEO letters and 88 (37.6%) with corporate sustainability. 67 (28.4%) of the above indicated familiarity with all three. Out of the three, ‘corporate reporting’ was the most general category. Accordingly, we elected to define those who indicated some familiarity with it as ‘non-laypersons’ (relative to those not at all familiar with the genre).⁴ This is a more restrictive selection than one based on familiarity with LtSs, but it is likely the best differentiator. We reasoned that it was possible for respondents at least somewhat familiar with LtSs but not at all with corporate reporting. Where this occurred, it seemed likely that these respondents were laypersons who had encountered a few LtSs but interacted little with the genre of corporate reporting beyond that. Restricting our selection to only those respondents familiar with all three categories, however, seemed slightly overzealous. The sustainability report remains a subgenre of the corporate report, and Chapter 3 previously demonstrated its similarity in language to financial corporate reporting. Because of that similarity, even respondents less familiar with the content of Letter to Stakeholders will have some

⁴ Virtually all non-laypersons indicated the source of their familiarity as a professional one, typically present or former employment with a company that interacts with or publishes report, or shareholdership.

experience with the *type* of language if they indicated familiarity with corporate reporting in general.

We also note that out of the 236 respondents who correctly reproduced the name of the (fictitious) company, 57 (24.2%) did not correctly identify the industry in which the company operated as the diamond industry. Of those 57, 42 (17.8% of total) identified the industry as the chemical industry, while 3 (1.3% of total) answered retail, and 12 (5%) answered that the company belonged to the oil industry. While this is a perhaps surprisingly high rate of error, we previously indicated that the key to a correct answer lies in a single line about diamond suppliers, and it is plausible for the respondent to gloss over that detail and nevertheless have already developed a notion (in this case a mistaken one) of which industry the company operates in. We might expect that the initial references to workplace fatalities might then lead respondents to answer with an industry they perceive as dangerous, i.e. the chemical industry. Nevertheless, as the degree to which the respondent was able to fully understand the text is a variable that will reflect in the questions we aim to answer, it would likely be imprudent to omit those erroneous responses.

4.5 Analysis

To test the effect of the manipulation process, we conducted a number of Mann-Whitney U-tests (to compare between laypersons and non-laypersons) and Kruskal-Wallis tests (to compare between the three different texts) using SPSS version 23. As with the rest of the study, we set an alpha level of 0.05. We also attempt to highlight those cases where the p-value approached significance ($p < 0.1$). The table below describes results for the various scenarios.⁵

⁵ While a few of these statements measure virtually identical perceptions (e.g. 'I trust this company' and 'this company is trustworthy'), the total number of questions was limited due to our use of a survey company (which was necessary in order to reach the target audience). As such, we used statements that measured the same broader categories (such as credibility or performance) but did not choose to conflate them as their specific phrasings and nuances had considerable potential to affect respondents' reactions.

Table 18 Summary of significance of differences (p-values) within respondents' answers, split along text and/or expertise. Responses on rows, categories in columns. Marginal significance ($p \leq 0.1$) in *italics*, significance ($p \leq 0.05$) in ***bold italics*** and strong significance ($p \leq 0.01$) in ***underlined bold italics***.

(The company is...)	Differences between Texts (Mann-Whitney U)			Differences between Expertise (Kruskall-Wallis)			
	Both Groups (n = 236)	Laypersons (n = 148)	Non-laypersons (n = 88)	Any Text (n = 236)	Original Text (n = 79)	More Readable (n = 78)	Most Readable (n = 79)
Open	.940	.963	.703	.449	.253	.702	.844
Honest	.563	.937	.439	.320	.121	.925	.839
Sincere	.507	.343	.735	.513	.236	.755	.736
Responsible	.859	.398	.606	<i>.045</i>	<i>.024</i>	.210	.996
Trustworthy	.609	.716	.281	.164	<i>.046</i>	.517	.836
Complex	.750	.729	.950	<i>.036</i>	.104	.271	.366
Competent	.855	.263	.494	<i>.018</i>	<i>.015</i>	.065	.865
Sustainable	.894	.652	.677	<i>.042</i>	.086	.111	.848
Professional	.496	.352	.687	.373	.259	.375	.664
I have a good feeling about this company	.713	.577	.883	.384	.316	.540	.983
I respect this company	.730	.401	.725	.353	.165	.488	.689
I trust this company	.740	.361	.415	.272	<i>.043</i>	.619	.629
This company has a clear vision for its future	.971	.965	.866	.967	.715	.870	.732
This company is well managed	.639	.272	.872	.203	.170	.260	.745
This looks like a company that would have good employees	.807	.667	.463	.443	.311	.313	.404
This is an environmentally responsible company	.961	.856	.910	.121	.191	.335	.693
This company maintains a high standard in the way it treats people	.681	.313	.868	.226	.100	.415	.794
This looks like a company with strong prospects for future growth	.954	.610	.746	.460	.329	.413	.572

<i>How positive or negative did you find the text?</i>	.617	.136	.604	.012	.009	.092	.992
<i>Clear</i>	.457	.775	.266	.364	.190	.383	.631
<i>Readable</i>	.854	.950	.741	.385	.422	.465	.987
<i>Written by an expert</i>	.720	.539	.952	.730	.777	.975	.387
<i>Complicated</i>	.510	.840	.351	.577	.846	.710	.203
<i>Well-written</i>	.469	.410	.505	.047	.102	.145	.619
<i>Easy</i>	.129	.459	.177	.344	.983	.992	.086
<i>Persuasive</i>	.918	.987	.923	.040	.252	.173	.326
<i>Accessible</i>	.920	.377	.528	.582	.419	.245	.284
<i>Pleasant to read</i>	.770	.309	.446	.198	.213	.072	.470
<i>How difficult did you find the text?</i>	.368	.775	.130	.966	.917	.277	.215

For those cases that displayed a significant difference between groups, Table 19 displays the means:

Table 19 Likert scale means of significant differences between responses.

	Any Text (n = 236)						
	<i>Responsible</i>	<i>Complex</i>	<i>Competent</i>	<i>Sustainable</i>	<i>Positive/Negative</i>	<i>Well-written</i>	<i>Persuasive</i>
Layperson (n = 148)	0.57	0.71	0.43	0.6	0.39	0.74	0
Non-layperson (n = 88)	0.17	0.4	0.06	0.27	-0.1	0.31	-0.35

	Original Text (n = 79)				
	<i>Responsible</i>	<i>Trustworthy</i>	<i>Competent</i>	<i>I trust this companay</i>	<i>Positive/Negative</i>
Layperson (n = 51)	0.8	0.24	0.63	-0.02	0.61
Non-layperson (n = 28)	-0.04	-0.54	-0.14	-0.79	0.29

4.5.1 Differences between Texts

At first glance, the distribution of answers between texts is very straightforward: none of the 29 questions about company and text perception shows a single significant difference between any of the texts amongst the full audience or (non-)laypersons alone. The greatest significance for any difference between texts occurs in ‘Easy to understand’ (at $p = .129$). While we might have expected some effect, it is worth noting that as respondents only read a single text, nothing in the short term is likely to have ‘primed’ them to read the texts relative to a given standard of corporate reporting language. In the case of the non-layperson group, they might already be at least somewhat accustomed to such language, based on, for instance, their professional activities. However, nothing about the experimental design gave the respondents a point of reference for corporate reporting, or helped them (re)gain one. Although doing so would have likely yielded more salient outcomes for this highest-level analysis, not doing so does enable us to divide the responses between those coming from a group with no point of reference whatsoever (the laypersons) and those that do have one (the non-laypersons).

To ascertain whether having that point of reference made a significant difference, we conducted the same analysis between the three texts twice more: once with only the laypersons, and once with only the non-laypersons. Neither group showed a significant difference between texts (with $p \geq .130$) for any of the questions. Again, outcomes might be more salient if respondents were comparing the text they read to another baseline text that every participant had read. However, this was neither logistically feasible nor

necessarily representative of the circumstances in which readers might read such reports.

4.5.2 Differences between Levels of Expertise

While we see no differences between texts if we consider the full group of respondents or the laypersons and non-layperson groups separately, results become far more salient when we attempt to explore the differences in responses between laypersons and non-laypersons. As the previous analysis did not indicate any significant difference between the three texts, we compared the responses the two groups gave overall (i.e. across all three texts). This revealed that laypersons considered the text they read or company described therein significantly more responsible, competent, sustainable, and positive, better written and more persuasive, but also more complex.¹ That is, except for their perception of greater (textual) complexity, laypersons were more optimistic in several key areas of company perception, including sustainability-related performance (responsibility and sustainability) and overall competence. However, the way in which the laypersons' opinion of the text they read differs from the non-laypersons' group is anything but straightforward in that the laypersons consider the texts better written and more persuasive, but at the same time more complex. One crucial question for future research could be whether that perception of qualitative, persuasive writing is in spite of or because of the complexity, i.e. whether it is possible to write in a style that maintains these positive aspects while reducing complexity.

The most pivotal results, however, lie in how the groups' reactions differ when we compare the original version with the more and most readable versions. When considering only the original version – and this effect is not present in the others – we again see significant differences between the two groups' optimism in a number of key areas of perception. Laypersons perceive the (company portrayed in the) text as more responsible, trustworthy, competent, and positive. Analogous to and consistent with their greater perception of trustworthiness, they agree significantly more with the statement that they trust the company, which supports H3 but contradicts H5. Despite not crossing the threshold of significance, perceptions of the company being sustainable and maintaining a high standard in the way it treats people also may trend towards greater optimism amongst the laypersons ($0.05 < p \leq 0.1$). This provides evidence both in favour of H1 and H2, although only partially so in the case of the former.

¹ While this is an aggregated effect across the three varieties, the per-text analysis reveals that most of this difference likely comes from differing perceptions regarding the original text, as the two groups show no differences in how they perceive the other two texts in that same per-text analysis.

The two edited texts, however, show no significant differences between the layperson and non-laypersons' perceptions; although the layperson judgments continue to be the more optimistic in absolute terms, the difference with the non-layperson group is no longer significant for any of the 29 questions. For the more readable text, 'competent', 'positive' and 'pleasant to read' come closest to significance but do not exceed the threshold ($0.05 < p < 0.1$), and for the most readable this is 'easy' ($p = 0.086$), with the non-laypersons judging the text as easier than the laypersons did. While this effect does not exceed the threshold of significance, it is nevertheless notable in how much stronger it appears than the original ($p = 0.983$) and more readable ($p = 0.992$) versions, the former with laypersons marginally more optimistic and the latter with groups virtually tied.

Remarkably, although perceptions of reading ease show signs of being more variable amongst levels of expertise as texts get easier, perception of professionalism does not appear to differ meaningfully between those levels of expertise as text difficulty goes down. That is, while a larger sample might show a significant – in addition to salient – difference in how layperson and non-laypersons perceive difficulty in texts far more readable than the average for corporate reporting, nothing in the data suggests that that would (negatively) influence their perception of professionalism, competence, or any positive qualities a company might want their reports to exude. Thus, we can reject H4

In other words, these results suggest that, although companies choosing to report in plain(er) English may lose the benefits of successful impression management towards readers with little genre experience, there is little reason for them to fear an adverse effect. While it might be initially costly, authors of sustainability reports appear to have considerable linguistic headroom to work towards a more accessible, stakeholder-inclusive language without damaging their credibility. Those familiar with the genre do not appear to think worse of them when LtSs attempt to use simple(r) language, and even those unfamiliar with the genre do not consider a text or the company behind it less professional as it progresses towards plain(er) English. By contrast, a divergence from the characteristics the genre has a reputation for – low reading ease and frequent use of passives – does diminish the positive attributes the layperson group attribute to it. However, steady change in the genre's (linguistic) reputation would likely erode that perception.

4.6 Limitations

The results should be considered in light of two main aspects of experimental design: first, its scope, and second, as already anticipated above, its between-subjects design that introduced no form of priming or common point of reference for any of the respondents.

In terms of the scope, we were logistically limited to approximately 225 respondents (final $n = 236$) of whom approximately one-third ($n = 88$; 37.3%) had some prior experience with corporate reporting. With access to a larger panel of respondents, adding two original – unedited – letters to stakeholders that approximated the formula-based readability of each of the two edited texts would have been a more robust way to verify the validity of the editing process. What mitigates the necessity of these additions somewhat is that outcomes of the study do already appear to suggest the editing process was valid, based on the different responses to the texts and shifts in perceived (reading) ease. On a similar note, just as adding more texts might have enhanced the study, it is self-evident that expanding the size of the respondent group that read every text would have further enhanced the study.

In terms of experimental design, we explored in the preceding section how a within-subjects design in which respondents compared and/or ranked the different versions of the text in terms of the different questions we asked might have yielded significantly more salient results in terms of the effect of readability on perception of the company. However, such conditions might be considerably further removed from real-world scenarios than this study's were. Readers of corporate reporting only see a single, final version of the text, which means readers are comparatively unlikely to consider the possible variations and phrases the author(s) did not choose to use in the final version while they are reading it.

Similarly, this study had to eliminate paratext during text collection due to the restraints other components' NLP-based methodology – and to a lesser extent this one's formula-based approach – placed on it. Without these constraints and with a larger group of respondents, this study might also have benefited from examining what the effects of including or leaving out accompanying visual information would be.

Finally, results would have likely been similarly more salient had we presented respondents with a single, universal (but unrelated) letter as a point of reference, and then asked them to compare an edited or original version of another letter to stakeholders to that baseline. As, for instance, differences in readability grow larger, we might expect that to reflect more saliently in the results in such a scenario. However, such an approach would have two drawbacks: first, the laypersons, while they would by no means merit inclusion in the non-layperson group after reading a single text, could no longer be said to have no experience with the genre whatsoever and would thus become less representative of the worst-case scenario for non-financial disclosures' wider stakeholder audience. Second, it would become far more difficult to determine which effects and differences the examiner could attribute to a difference in linguistic choices between the two texts the respondent reads, and which they should attribute to differences in content. For instance, if the first company reports far more favourable outcomes, that would more than likely influence the reader's overall perception. By contrast, this study did have the advantage of reporting identical content through

different language. Nevertheless, both a within-subjects design and one with priming through a universal point of reference have sufficient merit that we find ample evidence in this study to encourage such approaches in future research.

4.7 Discussion and Conclusion

Drawing back on the hypotheses, we find evidence for laypersons and non-laypersons responding differently to a reduction in linguistic complexity, which supports H2 and partially supports H1. Laypersons show more optimism than their counterparts in several areas when presented the original text, but this effect disappears when readability improves along Plain English guidelines. This effect suggests that the company's potential impression management efforts in the original document were likely successful in ensuring that those readers without meaningful experience with the genre perceived the company and text as positively as they could. However, they failed to have the same effect on those with more experience and a richer frame of reference, i.e. the non-laypersons, which leads us to accept H3. As this effect disappears in more readable versions of the same text, we can likely attribute it to companies' linguistic choices, given how the editing process controlled for content. This outcome suggests that, regardless of whether the company intentionally engineered this, laypersons do respond more optimistically to higher linguistic complexity; the language of the original had multiple desirable outcomes in terms of layperson perception (as a proxy for stakeholder perception). Given that, in the per-text analysis the low-readability original texts evoked more positive perception, we find one potential explanation for why sustainability content exhibits readability comparable to or worse than financial content, in spite of (or perhaps because of) its wider, less expert stakeholder audience: because, from what we can tell, this low readability works. While companies are unlikely to successfully achieve impression management on those familiar with the genre, their linguistic choices do appear to have a positive effect on the perception of those less familiar, and that effect disappears after making the text more readable.

As to why laypersons respond more optimistically to low readability than non-laypersons, the likeliest explanations would appear to lie in their relatively less elaborate frame of reference for the genre, as well as, potentially, a lower ability to deal with (genre-specific) textual complexity. The more experienced the reader, the less likely impression management strategies within a text are to succeed, as Leavy, Li & Merkley (2011) illustrate through the case of shareholders relying more on (more experienced) analysts as financial disclosures' readability declines. Based on that same (relative) lack of experience, laypersons may gloss over those elements that make those with more

experience more skeptical. As section 2.4.3 explored, a sufficiently high threshold for adequate comprehension can dissuade those less equipped to deal with that threshold from attempting to cross it at all. As the edited versions lower that threshold, laypersons' understanding of the text approaches that of the non-laypersons.

Equally crucially, however, we note that applying the Plain English guidelines does not imply a significant reduction in perceived competence, performance, author expertise, or professionalism in any of the situations we have examined, which leads us to reject H4. The simplified versions appear to negate a potentially desirable difference between the layperson and non-layperson perceptions. That is, layperson audiences attribute more positive characteristics to the original text than non-laypersons do. What causes this more favourable perception amongst laypersons might be an engineered (but not perceived, which contradicts H5) lack of transparency towards them or the beneficial effects of evoking association with other genres perceived as authoritative (consistent with Chartpraset 1993). Alternatively, laypersons might not feel empowered to be critical of the text due to its high complexity, and thus answer more positively. In other cases, editing the text towards general readability does not appear to negatively affect company perception, especially not (as companies might fear) amongst those more accustomed to the generally difficult language of corporate reporting – but neither do the simplified versions present to either audience as more readable.

Van Hoecke's (2018) findings that in a within-subjects experiment seven out of ten respondents replicated the readability hierarchy of the formula suggest that one of the main reasons we do not observe a shift in readability is the between-subjects design. That, outside of the vacuum of a single text, readers do perceive differences in difficulty speaks to the manipulation's validity. However, the result of three out of ten assigning a different hierarchy is equally notable: this illustrates the risks of writing to formula rather than holistically considering the audience's needs (which Chapter 6 expands on). This underlines that while the experimental setup is one part of the reason we see no significant effect, it is most likely enhanced by another, i.e. a loss of explicit cohesion and structure as optimising for formulae lowers sentence length. This chapter already found evidence that readability formulae are unable to capture all the nuances that contribute to a reader's experience, which cast additional doubt on the utility of readability formulae, especially for highly specialised genres. Chapter 6 further explores that utility, in addition to demonstrating a genre-adapted alternative (albeit one with its own challenges) in a machine learning-based readability predictor.

Apart from the genre's reputation for often impenetrable text, popular opinion also often perceives it as overly positive and prone to use of passives. Building on a more qualitative or hybrid approach to examining the genre's textual characteristics, the next chapter investigates the balance of positive and negative sentiment in the LtSs from sustainability reports, as well as how this sentiment relates to agency framing (e.g. passivisation).

Chapter 5

Sentiment Analysis

5.1 Motivation

While exploring corporate (sustainability) reporting, we briefly referenced the ‘Pollyanna Effect’ (e.g. Rutherford 2005), which refers to a perceived tendency of corporate reporting – and potentially corporate communications in general – to maintain a (very) positive register regardless of the outcomes it is trying to communicate. Of the UK annual reporting genre, Rutherford (2005, p. 349) states outright that “[t]he genre employs language biased towards the positive (the ‘Pollyanna Effect’) despite authoritative guidance that the OFR [Operating and Financial Review] should be expressed in neutral terms.” The crucial aspect of this “guidance” that Rutherford refers to is that “[the OFR] should be balanced and objective, dealing even-handedly with both good and bad aspects” (Accounting Standards Board 1993, para. 3, quoted in Rutherford 2005, p. 351).

Hildebrandt & Snyder (1981, p. 6) define the ‘Pollyanna Hypothesis’¹ quite broadly, as the notion that “positive, affirmative words are used more often than negative words”; they find this holds true in business communication, as does Rutherford, albeit with a number of caveats. The first of those caveats is that an excessive balance of positive words is unlikely to alter experts’ (e.g. analysts) understanding of the text, as illustrated by results of Chapter 4. A second potential issue is that excessive positivity can cause a (further) inflation of positive verbiage in these reports. That is, as genre conventions for OFR and likely other disclosures already tend towards positive language, not following these conventions can lead to misunderstandings, regardless of the performance reported on. As such, not following these trends can put a company at a disadvantage.

¹ The Pollyanna Hypothesis was named after an unfailingly optimistic Eleanor H. Porter character whose first adventure was published in 1913. Pollyanna sees the bright side in every situation, even when those positive interpretations are internally contradictory.

In other words, although Rutherford find evidence for the Pollyanna Effect, the cause thereof need not (generally) be an attempt at impression management. It may simply be that due to genre conventions, a disclosure stripped of its positivity for the sake of balance and objectivity would register as overly negative, in spite of communicating the same performance. Rutherford (2005) likens this to the expected excessive positivity of descriptions in real estate, where if both the encoder and decoder of the message expect excessive positivity, encoding it as neutral might cause a deficit. We should, however, note that this Pollyanna Effect is not unique to corporate reporting: as for instance Boucher & Osgood (1969) found, this preference is a ubiquitous phenomenon in human (linguistic) cognition. That is, a bias towards positive language in corporate reporting is, quite literally, only human, even if it is at odds with guidelines or regulatory demands. Positivity in reporting language is more problematic when it is at odds with less favourable performance, as this would imply obfuscation.

Rutherford succinctly presents that “[t]he context within which accounting narratives are produced [...] provides both opportunities for, and constraints on, communication” (2005, p. 350), again referring to the many mechanisms for oversight and regulation that exist for financial reporting. However, given sustainability reporting’s greater voluntariness, we have reason to expect differences in how this interplay of opportunities and constraints manifests. On the one hand, from a defensive attribution perspective, we might expect companies to use a veil of positive verbiage when results are poor(er) because less regulatory oversight means fewer potential undesirable (e.g. legal) effects from doing so. Conversely, that same (relative) lack of oversight might incent companies to omit unfavourable information altogether, save for those contexts where not addressing them might be more reputationally damaging (e.g. workplace fatalities). Given the considerable evidence for the Pollyanna Effect in corporate reporting, the different dynamics of the genre (especially its greater voluntariness) vindicate a re-examination in a context of non-financial disclosures.

However, as previous chapters have explored, sustainability reporting is thematically relatively more complex than conventional financial reporting in that it typically addresses at least four main pillars: financial, environmental, social and governance-related performance (see e.g. Thomson Reuters 2013 or Loh, Thomas & Wang 2017).

This diversity of topics threatens to frustrate conventional binary sentiment analysis approaches that consider an event either beneficial or detrimental for its target, given that an event that benefits one aspect might detract from another. Van de Kauter, Desmet & Hoste (2015), for instance, attempted to classify financial news as good or bad for a given target (typically a company). For example, such an approach should interpret the sentence ‘ACME Corp. share prices drop’ as displaying negative sentiment towards ACME Corp. Share prices dropping is, more often than not, an undesirable outcome for a company.

Van de Kauter, Breesch & Hoste (2015) emphasise the importance of sentiment analysis and its relationship with corporate performance on the basis that stock markets fluctuate chiefly based on human reactions to events, rather than those events themselves; that is, emotion can have a stronger impact on markets than rational behaviour does (Thaler 1993). Rather than investigating investor sentiment, they investigated linguistic sentiment, i.e. the presence of linguistic elements with an affective aspect. They used a polar sentiment approach – i.e. attempted to detect good and bad news – in a financial news corpus that covered several Belgian companies. They also extend their analysis beyond explicit sentiment towards implicit sentiment. The former are subjective utterances that contain a positive or negative opinion, while the latter are utterances in which a positive or negative evaluation can be inferred, but is not explicitly expressed. Example (7) illustrates explicit sentiment, while (8) illustrates implicit sentiment:

- (7) “I believe ROC is now moving into an exciting operational period with significant growth potential.” (ROC Oil 2013)
- (8) “The profit after income tax for the 12 months to 31 December 2012 was US\$158.7 million.” (PanAust 2013)

Depending on the context, readers will likely interpret example (8) as containing either positive sentiment (if this is a desirable profit for the company to achieve) or mixed sentiment (if this is less than desirable, in spite of it being a profit). In spite of (8)’s falsifiability and objectivity, very few would argue that this sentence reflects neither positively nor negatively on the company. These implicit sentiment utterances are especially notable in corporate reporting as they are a highly prominent vector for sentiment that nevertheless enables the author to maintain a factual tone, in that they contain no subjectivity markers but nevertheless reflect favourably or, at times, unfavourably on the company. Parsons & McKenna (2005) emphasise the rhetorical importance of such a factual tone in these reports’ rhetoric. Van De Kauter, Breesch & Hoste (2015) report both considerable added value and potential implementation difficulties in quantifying implicit sentiment as well as explicit sentiment in terms of polarity.

However, as this corpus chiefly contains sustainability content, recognising polar sentiment in terms of what is good or bad for the company becomes considerably more challenging. The multiple perspectives inherent to most sustainability reporting inhibits the viability of the above approach, as a single sentence might be more or less positive (or even carry the opposite sentiment) depending on the performance perspective. For instance, the following (fictitious) sentence is positive from a financial perspective (it benefits the bottom line), but negative from a social one (it reflects a disengagement with local communities):

- (9) ‘As a cost-cutting measure, ACME Corp. has terminated its community outreach program.’

While a highly detailed, word-level annotation scheme might recognise that ‘as a cost-cutting measure’ can reflect positive, negative or mixed sentiment on the target of ACME Corp. (depending on the annotator’s interpretation), ‘has terminated its community outreach program’ almost certainly reflects negative sentiment towards ACME Corp.’s social efforts.

In addition to Van De Kauter, Breesch & Hoste’s (2015) efforts towards automating the sentiment classification of news items, there has been considerable work performed in content analysis of corporate report content (Van Den Bogaerd & Aerts 2011). Corporate report content analysis traditionally falls into either a manual annotation approach or an automated one, which is evolving from what Van Den Bogaerd & Aerts (2011) call “naïve, heuristic algorithms” towards more advanced machine techniques. They indicate that the manual annotation approach will almost inevitably exceed the naïve algorithm approach in precision, but is, of course, highly time consuming. Another challenge lies in ensuring consistency amongst annotators. At the same time, naïve algorithms are limited in terms of the accuracy they can achieve and nuance they are capable of portraying.

Some of the more prominent efforts into automatic analysis of environmental reporting include Brown & Deegan (1998) and Neu, Warsame & Pedwell (1998). Similar ventures into content analysis of Letters to Shareholders include Abrahamson & Park (1994) and Abrahamson & Amir (1996). Each achieved notable, influential outcomes, albeit with a coarseness in how they deal with text and content on a quantitative level (typically using word counting techniques) that present-day NLP techniques would be able to mitigate.² Van den Bogaert & Aerts (2011) expanded on this by applying machine learning techniques (see section 2.6.3) to recognise positive and negative news items, similar to Van de Kauter, Breesch & Hoste’s (2015) efforts. The former study had less of an emphasis on implicit sentiment and attest an approximate accuracy of 90% in classifying whether a text contained favourable news regarding a company.

For sustainability reporting in particular, Wen (2014) similarly attests the value of machine learning techniques in optimising a machine learner to help users extract sustainability-related information from a report and (coarsely) analyse reports’ sentiment. Like Van den Bogaert & Aerts (2011) as well as their aforementioned precursors, this analysis was largely lexicon-based; it relied on word lists, which are fairly weak at detecting implicit sentiment. While on a genre-specific basis it is certainly possible to detect implicit sentiment for common items (i.e. ‘profit’ usually carries positive sentiment), a lexicon-based approach is less suited to more nuanced sentiment; for instance, what if, as may have been the case for example (8), there was a profit, but it was less than expected? Alternatively, a company might also report ‘negative profit’ – in such cases, a sentiment analysis system is likely to recognise ‘profit’ as a positive word

² For a more exhaustive overview, see Van den Bogaerd & Aerts (2011).

and ‘negative’ as a negative, which might cancel out, while the actual sentiment value of ‘negative profit’ is highly negative, as it serves as a euphemistic synonym for ‘loss’.

As such, we see considerable value in applying Van De Kauter, Breesch & Hoste’s (2015) sentiment analysis approach to sustainability reporting, especially given the (sub-) genre’s potential for tension between different areas of sustainable performance. As such, this chapter will pursue a number of aims related to sentiment and rhetoric in LtSs from sustainability reports. As a primary aim, it will further explore the Pollyanna Effect and use of agency framing (which ties in with passivisation; see e.g. section 3.4) in these documents. Additionally, it will investigate the viability of annotating sentiment for the different performance aspects simultaneously, which may enable future machine learning-based studies with similar aims as Van den Bogaerd & Aerts (2011) or Wen (2014) to substitute a lexicon-based sentiment approach for a much more nuanced one based on human-annotated ‘gold standard’ data.³ To achieve this, this chapter explores multiperspective sentiment annotation efforts and an analysis thereof for the Letters to Stakeholders (i.e. CEO or Chairman’s Letters from standalone sustainability reports) present in the corpus.

5.2 Hypotheses

Although the above illustrates how ‘good for the company’ and ‘bad for the company’ can take on greater depth and nuance from a multi-aspect sustainability perspective, the attested presence of the ‘Pollyanna Effect’ in more traditional corporate communications (Rutherford 2005) still makes it likely that we will find a considerable amount of positive language. Brown & Deegan (1998) Neu, Warsame & Pedwell (1998) also found results consistent with the Pollyanna Hypothesis in environmental disclosures. However, we must also account for the potential tensions between pillars and, where applicable, reporting guidelines’ requirements regarding balance, such as the GRI’s, which state that “the report should reflect positive and negative aspects of the organization’s performance to enable a reasoned assessment of overall performance” (GRI 2013). Nevertheless, at time of data collection sustainability reporting was a largely voluntary

³ Due to the technical complexity of implementing annotation-based machine learning, this study will not attempt to create an automatic sentiment analysis system; it limits itself only to the annotation process. Creating a trained and optimised sentiment analysis system based on manual annotation is, quite literally, a research project in its own right (e.g. Van de Kauter, Desmet & Hoste 2015).

(sub-)genre and thus offered considerable opportunity for companies to include only favourable information. As such, we formulate with considerable certainty that:

H1: Letters to Stakeholders from sustainability reports will contain more positive information than neutral or negative information.

As, in Chapter 3, we investigated obfuscation on a text level and found little evidence, we might expect that the extent of positive information present in a text will show no meaningful association with the company's performance relative to the same performance aspect (e.g. social or environmental performance). This would be consistent with e.g. Wen's (2014) findings based on a single aggregate sustainability score, but may differ on a per-performance aspect basis. Neu, Warsame & Pedwell (1998), for instance, also call for more research into the relationship between environmental disclosure and environmental performance. More recently, Aerts & Yan (2017) found high use of positive words regardless of actual performance in annual report (i.e. non-sustainability-report) LtSs. As a corollary to H1, we can formulate:

H2: The amount of positivity or negativity around an aspect of performance in Letters to Stakeholders from sustainability reports will show no meaningful association with its performance for the same aspect.

As, during the full-corpus analysis, we found the greatest association between performance and extent of passivisation (see section 3.4), which suggests defensive attribution (see e.g. Aerts 1994) more than it does obfuscation, we have the opportunity to further explore that association in a manually annotated corpus. Rather than relying on extent of passivisation as a proxy, however, we have the opportunity to ask annotators to directly annotate how the author (or company) frame its agency. We hypothesise:

H3: The more positive the message in a given sentence, the more direct (i.e. closer to first-person) its agency framing will be, and vice versa for negative outcomes.

Additionally, we wished to investigate Parsons & McKenna's (2005) assertion that non-falsifiable and promissory assertions were especially prevalent in sustainability reporting compared to factual, falsifiable claims about steps previously undertaken. The latter were still the most frequent, but these promissory statements were a highly prominent part of the company's rhetoric Parsons & McKenna give (*inter alia*) the following examples (2005, p. 603), with the weakly falsifiable element underlined:

(10) Wherever Comalco operates, our vision is to be the preferred partner for communities.

(11) This will mean sitting down with members of the community and other interested parties to listen to their views on what we are doing and how we report our progress.

However, reporting might have evolved in the seven to eight years in between their study and composition of the subcorpus. While their assessment relies too much on qualitative,

subjective appreciation of text content to formulate a quantitatively verifiable hypothesis (as they do not indicate how often such statements occur, in spite of their prominence in the rhetoric), it nonetheless merits further investigation. We aim to further explore the notion by attempting to capture the dominant rhetorical move, temporal framing and level of subjectivity throughout the subcorpus, as detailed in section 5.4. Before doing so, however, we will briefly explore the utility of lexicon-based sentiment analysis.

5.3 Automatic Lexicon-based Sentiment Analysis

Brown & Deegan (1998), as well as Wen (2014) find an imbalance of positive and negative terms in favour of the positive using a sentiment-based methodology. There is considerable evidence for the Pollyanna Effect in non-financial disclosures, but, as previously mentioned, lexicon-based sentiment analysis is substantially less sensitive to implicit sentiment.

Fundamentally, lexicon-based sentiment analysis consists of counting positive and negative words present in a text, the former determined by a lexicon of positive words and the latter by a lexicon of negative words. As previously mentioned, one of the major weaknesses – and a key reason why this study is investigating the viability of multiperspective sentiment annotation – is that how these individual words interrelate is almost impossible to measure with lexicon-based sentiment analysis. That is, the ultimately negative ‘negative profits’ is difficult for lexicon-based methods to capture, as both Van De Kauter, Breesch & Hoste (2015) or Socher et al. (2013) demonstrate.

As such, we first applied a fairly straightforward lexicon-based sentiment analysis to the LtSs that make up the subcorpus for the annotation-based analysis (see section 5.4). We measured the number of positive and negative words in the text, divided by total text length. As a lexicon, we used the General Inquirer lexicon (Stone et al. 1966).

The mean relative percentage of positive words was 7.09%, with a standard deviation of 1.64%; for the relative percentage of negative words, this was a mean of 1.49%, with a standard deviation of 0.62%. This is consistent with previous studies’ findings of more positive words than negative terms, and suggests these LtSs may not adhere to a strict interpretation of balance between the positive and the negative, in that the standard deviation for percentage of positive words is larger than the mean for negative words.

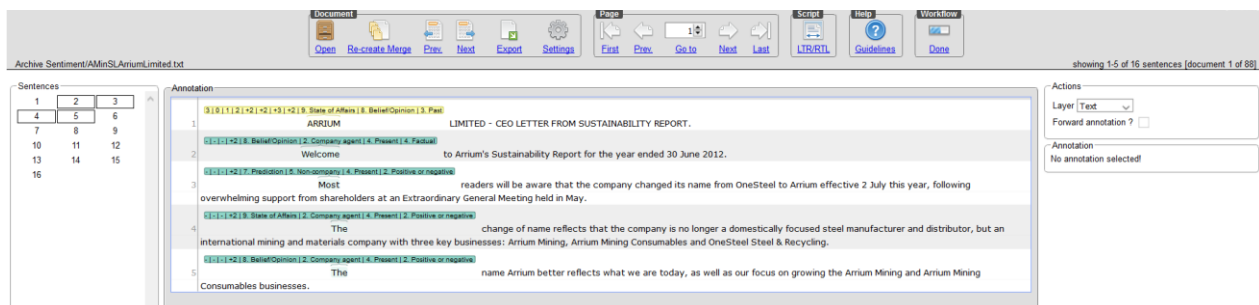
We also constructed two linear models with either percentage as the dependent and verified the impact of the four performance scores, language variety and industry for the relevant document. None of these were significant predictors at an alpha level of 0.05 for the percentages of positive or negative words. In short, performing a lexicon-based analysis added little, if anything to the results of previous studies performing similar

analysis. We see merit in this outcome for a more detailed, human annotation-based analysis; as Chapter 3 found for readability formulae, an in-depth look can expand on fairly shallow results.

5.4 Annotating Sentiment in a Sustainability Context

To assess the extent to which the Pollyanna Effect manifests in sustainability content, we assembled a subcorpus that consisted of the 88 Letters to Stakeholders from sustainability reports present in the full corpus. We presented this subcorpus to three MA-level English student annotators (based, as is the case for Chapter 6, on their analytical insights of the language and ability to emulate an educated non-expert’s perspective). Annotation used WebAnno (version 3, beta 7), assigning approximately one-third to every annotator, with seven shared texts to calculate inter-annotator agreement. Figure 2 gives an example of the WebAnno interface.

Figure 2 Example of the WebAnno annotation interface.



As our aim was to best capture the dynamic between positive and negative information regarding the four primary performance pillars described in sustainability reporting, even a detailed, word-level sentiment analysis approach such as Van de Kauter, Desmet & Hoste’s (2015) had certain shortcomings for our purposes, although these guidelines served as a strong inspiration for the final guidelines. We include this study’s final guidelines in Appendix 4 (‘Sentiment Annotation’). These guidelines also contain examples of annotated sentences.

As previously mentioned, these guidelines place a strong emphasis on extracting both explicit and implicit sentiment. They refer to utterances containing explicit sentiment as *private state expressions* (“internal states that cannot be directly observed by others, e.g. opinions, beliefs, [etc.]”, p. 690). They refer to the latter as *polar fact expressions* (“expressions conveying a piece of factual information that, when relating to a certain target entity or entities, results in a positive impression of that target/those targets”, p. 690). However, they do not treat objectivity or subjectivity as a binary problem, also

allowing ‘mostly objective’ or ‘mostly subjective’ sentences. They also allow for an utterance to express degrees of positive or negative sentiment about a target, which may or may not be present in the same sentence, as well as allowing for a different source than the speaker, for instance in the case of reported speech.

However, some aspects of this approach were suboptimal for multiperspective sentiment annotation. For one, the different perspectives make it more difficult to interpret any given phrase in terms of single-target polar sentiment. We can present equally valid arguments for the phrase ‘a cost-cutting measure’ reflecting favourably on the company, as it will benefit their bottom line, as we can for its need to implement cost-cutting measures implying financial dire straits likely to compromise its long-term survival. Given this tension, we can also argue that ‘a cost-cutting measure’ contains mixed sentiment (i.e. both positive and negative). Apart from being less informative as to the dynamics between the four performance pillars, the third ‘mixed’ option also illustrates the overhead in determining fine-grained, word-level sentiment. While even very coarse-grained, binary sentiment annotation tasks can be difficult and time-consuming, this case provides a clear example of the overhead we can expect when dealing with a fine-grained annotation scheme, especially one that requires annotation at the word level.

As such, we found ourselves better served drafting a new annotation scheme, combining aspects of Van de Kauter, Desmet & Hoste’s (2015) financial market (i.e. corporate)-oriented approach and the more flexible perspectives of aspect-based sentiment analysis (see e.g. Pavlopoulos 2014). We primarily asked annotators to describe the sentiment towards each of the four performance aspects at a sentence level for the sake of efficiency and cognitive load. A crucial disadvantage to word-level sentence annotation is that annotators need to consider an exponentially larger number of options (for every combination of words) rather than being able to answer the same recurring set of questions for every sentence. As the scheme already dealt with some fairly complex concepts, we opted for sentence-level annotation; this also enabled more accurate comparisons between annotators given the smaller number of options.

In order to further explore Parsons & McKenna’s (2005) observations about the content of sustainability reporting, we asked annotators to indicate how forward-looking the sentence was, ranging between positioning itself (mostly) in the future and expressing an intention to being set mainly in the past, present or an atemporal frame. This would help us identify, together with the extent to which the sentence contained an opinion, how falsifiable the content of the sentence was.

To capture the extent to which the sentence contained an opinion or fact, we also asked annotators to place it on a spectrum of subjectivity. This ranged (in descending order) from the sentence presenting an outright opinion, presenting a fact with positive or negative colouring, or presenting a fact but containing some subjectivity, to being almost entirely factual or altogether not assertive. We place emphasis on how the

sentence presents itself; for it to count as presenting an opinion, it would have to highlight the author's own position in the discursive frame, e.g. through parenthetical verbs or comment clauses (such as 'we believe'). Alternatively, the sentence might highlight its containing an opinion through adverbial forms such as 'to be honest' or other means of signalling some distance from the factuality of the assertion, such as the sentence modifying the author's certainty through 'supposedly' or 'allegedly', e.g.:

(12) "We **understand** there are opportunities to improve our data collection processes, especially where information comes from third parties such as contract factories or material vendors that supply to such factories."

We also aimed to discover whether companies were more likely to use defensive attribution strategies when conveying less favourable news. Accordingly, we asked annotators to capture to what extent the company served as the agent in the sentence they were annotating. We considered a first-person pronoun (such as 'we' or 'our') to be the most direct way for the company to mark itself as the agent. This is followed by the company name (or, for instance, 'the company'), followed by use of metonymy (such as 'the steel industry' or 'the HR division'), followed by the company without explicit mention, e.g. in a passive-voice construction. The final two scenarios were an explicit and implicit (or hidden) non-company agent, respectively. The motivation behind asking annotators to describe the agent was that it would enable us to measure whether positive news showed any association with explicit attempts to foreground the company as the agent, and whether the inverse would also hold, i.e. to what extent companies practice defensive attribution on a sentence level.

Additionally, we wanted annotators to indicate the primary type of rhetorical move (i.e. the illocutionary effect the author intends to obtain; analogous to Searle's 1969 concept of speech acts) behind the sentence. This would enable us to better understand how sentences in the subcorpus rhetorically conveyed any excessive positivity that they might contain. We asked annotators to describe which of the following best captured the main rhetorical move contained in the sentence:

- Apology
- Request
- Question
- Expression of gratitude
- Expression of intention
- Expression of desire
- Prediction
- Expression of belief or opinion
- Description of a state of affairs
- (Other)

We might expect, for instance, that companies will be more likely to explicitly position themselves as an agent when expressing gratitude, positive intentions or desires, and less likely to do so when making more unfavourable predictions or describing unfavourable states of affairs. Of the following fictitious examples, the metonymic (13) would be far less direct, and potentially rhetorically more desirable than the explicit first-person (14) :

(13) ‘The steel industry is going through hard times right now.’

(14) ‘We are suffering through hard times right now.’

Conversely, companies might still use high-agency constructions when issuing apologies, even if the overall sentiment is negative; for instance ‘We would like to offer our most sincere apologies’ might be far more effective than the metonymic ‘The industry would like to offer its most sincere apologies.’ This is a potentially more jarring construction both because it creates a more dissonant rhetorical distance between speaker and message and because the company may not have the (moral) authority to speak for the entire industry. As such, while sentiment and agency might show associations, the main type of rhetorical move present in the sentence may also influence agency framing alongside other choices.

For all of the above, it is a fairly straightforward intuition that having information about such elements as agency framing and type of rhetorical move over an arbitrary sub-sentence level span of words (for instance, but not limited to, phrases or subclauses) would enhance the granularity of available information. Given the comparatively high average length of sentences throughout sustainability reporting, it seems more than likely that many complex sentences will, for instance, be capable of making both a prediction and presenting a state of affairs, or presenting the company as an agent through both a personal pronoun and metonymy. However, given the already time-intensive and, according to trial run⁴ annotators, cognitively exhausting annotation scheme, the annotation process would have likely been far less viable, both in terms of labour and time available and in precision on annotators’ behalf. More and finer-grained elements to annotate increases the chances of errors, which increased fatigue from a higher cognitive load may further have exacerbated.

As such, we made a deliberate choice for sentence-level annotation, and attempted to mitigate the primary negative consequence – a relative lack of granularity – by using a

⁴ Trial runs consisted *inter alia* of investigating which questions it was possible to ask and still have annotators maintain what we considered an acceptable speed as well as consistency between annotations. For example, an earlier version of the guidelines asked annotators to assess the sentence’s subjectivity (similar to Van De Kauter, Desmet & Hoste 2015) in addition to the extent to which the authors presented it as an opinion. As gauging extent of subjectivity is a very difficult task, and because we were able to combine the two questions into a single one and still be able to refer to the sentence’s main rhetorical move for further insight, we elected not to explicitly ask annotators how subjective the sentence was.

trickle-down annotation system favouring the most salient result for any of the features to be annotated. In brief, annotators had a ranked list of possible annotations for every feature, The sole exceptions to this ranked order were the ‘very positive’ to ‘very negative’ Likert scales that describe per-aspect sentiment. When deciding how to annotate each feature, annotators went through this list in rank order and, upon finding an appropriate annotation, stopped without considering the appropriateness of the lower-ranked labels, and continued on to the text feature. For instance, the above list of possibly relevant rhetorical moves above is in rank order. As ‘apology’ has the highest rank out of all the possible rhetorical moves, an annotator that tags an apology in the sentence they are annotating need not consider the appropriateness of the other labels; they flag it as an apology and annotate it for the text feature.

The logic behind the rankings expressed two things: the markedness of that particular option, and the extent to which its presence eclipsed the relevance of the other potential labels for that feature. For instance, it is entirely plausible that a sentence that announces a future intention on the company’s part is logically and syntactically connected to a statement about past or present performance, but that future intention is far more likely to be the focal point of the sentence given the genre. Similarly, an apology or expression of gratitude may well contain a statement of fact or opinion, but the apology or expression of gratitude will be a much rarer, and likely more significant rhetorical move in that particular sentence.

The types of rhetorical move were the most difficult to order in terms of trickle-down choices with respect to the relationship between the higher-ranked (i.e. more salient) lower-ranked (i.e. less salient) categories. Generally, it is simply that of a widening net or increasingly less stringent filtering process. For instance, in the case of the ‘subjectivity’ feature, the first-choice option is that the sentence explicitly presents itself as an opinion. This is the most rhetorically salient scenario, but also the most specific; it will almost certainly contain positive or negative colouring (which is the second-ranked choice) and inevitably contain subjectivity (which is the third-ranked choice). By elimination, if none of the three are true, then the sentence must be factual (the fourth-ranked choice) unless it is not at all assertive (the lowest-ranked choice). The same principle holds true for agency framing, with increasingly wider interpretations of the company presenting itself as an agent as the annotator proceeds from higher-priority to lower-priority options.

Finally, when annotators were finished with sentence-level annotation for the document, we asked them to make a number of text-level judgments very similar to the sentence-level ones. For one, we asked them to rank the degree of attention the four performance pillars received throughout the text (leaving out any that received no attention at all). For those aspects mentioned, we also asked them to evaluate the overall sentiment on a seven-point Likert scale between very negative and very positive. In terms of rhetorical moves, we asked them to rank the three most prominent types of rhetorical move (matching those on a sentence level) with the aim of capturing their overall

impression of the text. Finally, we asked which temporal frame the text was most situated in (with options identical to the sentence-level temporal frames).

The reasoning behind this was that while for some of these questions we would always be able to simply count the number of instances of a given answer, not every sentence would necessarily be equally important to the overall message. Asking these text-level questions enables us to get the annotator’s overall impression of the text, which, although inevitably subjective, may be more representative than synthesising an overall characterisation of the text from its sentence-level annotations. However, this also entailed the risk that annotators’ own sensibilities or biases would create greater inconsistency between them (one might e.g. be more sensitive to information on social or environmental performance).

5.5 Inter-annotator Agreement

As inter-annotator agreement calculations in WebAnno would reveal, one annotator had, contrary to indications, left sentiment values blank where there was no sentiment for a certain aspect, rather than selecting the ‘blank’ option. As this impeded inter-annotator agreement calculations as well as extraction and processing, we regrettable found it the better option to discard this annotator’s annotations from the subcorpus.

5.5.1 Text-level Annotations

The table below displays the inter-annotator agreement on the two remaining annotators’ tags for the text-level (as opposed to sentence-level) annotations. We calculated these in WebAnno’s (version 3.0, beta 7) curation interface using Cohen’s Kappa.

Table 20 Inter-annotator agreement scores (Cohen’s Kappa) for text-level sentiment annotations.

Text-Level- Annotations							
Attention to Financial	Attention to Environmntl.	Attention to Social	Attention to Governance	Financial Sentiment	Environmntl. Sentiment	Social Sent.	Governance Sentiment
0.06	0.17	0.48	0.02	0.17	0	0.22	0.22
Primary Rhetorical Move		Secondary Rhetorical Move		Tertiary Rhetorical Move		Time Frame	
-0.04		0		0.18		0.35	

As results were, at best, inconsistent (with the exception to texts’ attention to the social aspect), we judged it better not to proceed with these text-level annotations and, instead,

average them from the sentence-level annotations. Given the reasons for asking for a text-level assessment indicated above, this was unfortunate, but still preferable. Although Cohen’s Kappa is fairly conservative, especially when annotation categories have an order to them, results for sentence-level annotation were sufficiently better to pursue that option.

5.5.2 Sentence-level Annotations

The table below shows the same Cohen’s Kappa-based inter-annotator agreement extracted from WebAnno, this time for sentence-level annotations.⁵

Table 21 Inter-annotator agreement scores (Cohen’s Kappa) for text-level sentiment annotations.

Sentence-level Annotations							
Financial Sentiment	Environmntl. Sentiment	Social Sentiment	Governance Sentiment	Rhetorical Move	Agent	Time Frame	Subj.
0.46	0.49	0.55	0.18	0.49	0.72	0.53	0.27

We observe that the lowest score is for governance-related sentiment, just below the conventional threshold of fair agreement. As governance aspects of sustainability are likely the most nebulous to non-experts, this is a foreseeable outcome. The subjectivity category, in spite of how difficult subjectivity and objectivity are to delineate, still occupies the conventional ‘fair agreement’ category. While we can carry it forward into further analyses, we must do so with the appropriate caution. As annotations for the other categories fall between moderate and substantial agreement, and given the relative complexity of the annotation task (see Appendix 3) and the conservative nature of Cohen’s Kappa in ordinal annotations, we assert that these outcomes support the validity of the annotations and data set.

5.6 Processing

Before analysing the annotations further, we performed a number of processing steps necessary to combine them with companies’ performance scores.

⁵ The annotator who participated in the testing phase also achieved viable inter-annotator agreement scores with our own annotations at a Kappa of between 0.28 and 0.54 for the various categories. This is with the exception of subjectivity (at a Kappa of 0.02), which we later remedied through additional coaching.

The first step was to download the annotations in the WebAnno tab-separated value format (version 3). This produced a (tab-separated) table that lists every annotation span per sentence. As annotators received the instruction that they were free to apply their annotation for the sentence only to the first word, we needed to collapse annotations that spanned the entire sentence down to only the first word by deleting any annotations for each sentence other than the first. We then renamed every file to also indicate the annotator. All these steps used regular expressions, carried out through PowerGREP 4:

1. **Collect** `^(#Text.+|[0-9]+-1\t.+)$`
Match those lines that start with either the ‘new sentence’ marker in WebAnno TSV 3, ‘#Text’, or that contain the annotations for the first annotated element, of which the number sequence at the start always ends in ‘-1’.
2. **Delete** `^(#Text=[1-1.+|[0-9]+-[0-9]+\t[0-9]+-[0-9]+[\^-*0]+)`
Delete the line-initial tags that identify its position in the document (which start with either ‘#Text’ or ‘1-1’ followed by a hyphenated number sequence. This ensures that every annotation takes the format of the original sentence, followed by its tags (separated by tabs) on the next line.
3. **Replace** `\r\n(?![1-9a-zA-Z]) -> /t`
Substitute line-breaks not followed by alphanumeric characters with a tab; this places the annotations (which never start with alphanumeric characters) on the same line as the sentence, separated by a tab.
4. **Replace (within the relative path name)** `^([A-Za-z0-9&-]+)\.txt\\([A-Za-z0-9]+)\.tsv -> \1\2.tsv`
This strips the ‘.txt’ from the path name and collects all the (tab-separated value) files within the target folder and subfolders into the target folder, renaming them as DocumentNameUserName.tsv.

We then concatenated these documents into a single spreadsheet, separated by the file names, separated tag ranks and types into different cells (e.g. ‘9. State of Affairs’ becoming ‘9’ and ‘State of Affairs’), and manually removed sentences that had exhibited errors during the collection process and did not have a full set of annotation tags available, either due to user (annotator) error or errors during the collection process. Causes included:

- **Incorrect sentence splitting in WebAnno**, for instance in cases with colons and semicolons. In these cases we had asked annotators to consider the full sentence, and thus deleted any sentence fragments split off by WebAnno.
- **(Very rare) cases with the ‘Other’ rhetorical move type**, which did not correctly port over to the spreadsheet interface due to WebAnno only allowing for numbered tag categories up to 9; ‘Other’, as the tenth, was not numbered, and thus incorrectly carried over after splitting ranks and types into different cells.

- **Sentence fragments appearing as sentences due to punctuation;** for instance, ‘An interview with Pawet Olechnowicz, CEO of Grupa LOTOS S.A.’ (Grupa Lotos 2013) does not contain a verb, and generally makes it impossible to apply the annotation scheme, which requires (*inter alia*) that the annotator designate an agent.
- **Incomplete sentences.** Where the annotator had not assigned the sentence a tag for every category, this frequently caused issues with the conversion and collection process. In those relatively few cases where a missing tag did not create processing errors, we nevertheless deleted the entire sentence rather than fill in the gaps as the annotator’s interpretation of the sentence may not necessarily have aligned with our own.
- Finally, in those cases where the two remaining annotators had both annotated a text (which we needed to calculate inter-annotator agreement), **we selected the annotations belonging to the annotator who had performed (subjectively) better** during the initial try-out phase.

The final subcorpus after deletions tallied to 2384 unique annotated sentences across 74 texts (which is 84% of all the LtS from sustainability reports in the corpus). For the texts, we counted the total number of sentences with mention of a specific aspect, and average scores for financial, environmental, social and governance sentiment only for those sentences where annotators had assigned a score (for instance, when three out of fifteen sentences contained financial sentiment, we averaged over three, not fifteen). We also averaged the ranks for agency (with lower scores being closer to first person agency) and timeframe (with lower scores being closer to the past).

5.7 Frequencies and Description

As a first exploration of this subcorpus, we will tally annotation results for the various categories and subsequently discuss the outcomes. We counted the various categories using SPSS version 23. The authoritative annotator had annotated 1283 sentences; the second 1101.

5.7.1 Sentiment scores

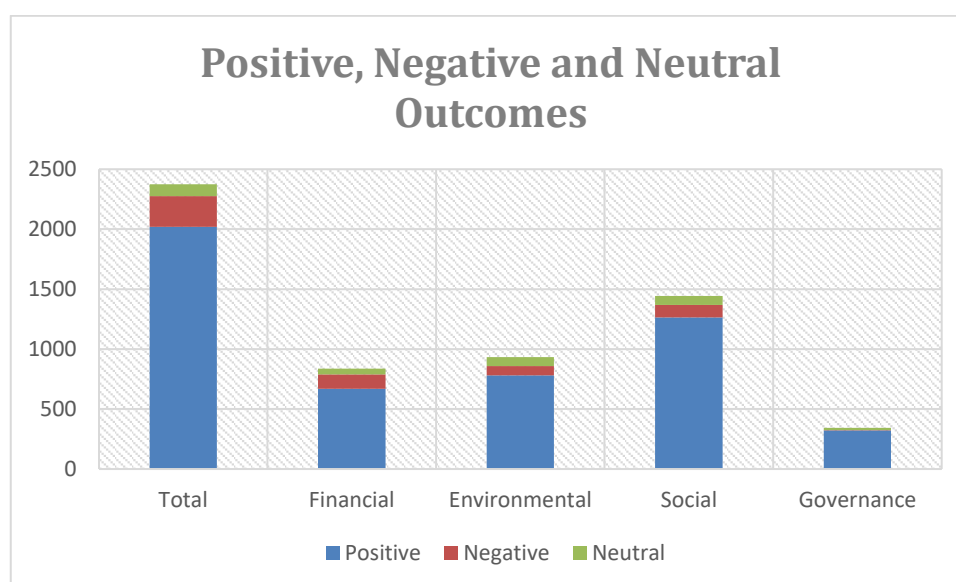
Table 22 Sum total of all negative and positive sentiment in all sentences in the subcorpus. A sentence can contain both negative and positive sentiment.

Negative	Neutral	Positive
255	100	2020

Table 23 Tallies of sentiment scores assigned to all sentences in the subcorpus.

	Sentiment									
	Any negative	-3	-2	-1	0 (blank)	+1	+2	+3	Any positive	
Fin	119	27	76	16	48	1548	93	501	75	669
Env	78	9	60	9	75	1451	62	625	93	780
Soc	104	25	65	14	77	940	77	1011	175	1263
Gov	5	1	1	3	17	2041	12	290	19	321

Figure 3 Stacked column chart representing positive, negative and neutral outcomes per performance aspect. Blanks not included.



As the above tables and figure indicate, out of a total of 2384 sentences, annotators tagged 2020 (84.7%) as containing at least some positive sentiment for one or more of the cases, and 255 (10.7%) as containing at least some negative sentiment. We note that these categories are non-exclusive; the two categories overlapped in 62 (2.6%) cases, which implies that according to the annotators, the entire subcorpus contained 171 (7.2%) sentences with no positive or negative elements for any of the performance aspects.

Cases where a tension existed between positive sentiment for one aspect and negative sentiment for another were typically cases where one aspect benefited at the expense of another. In 36 cases (i.e. over half) these sentences reported on non-financial initiatives incurring expenses, e.g.

- (15) “All mining activity was suspended whilst briefings and training sessions were held across sites.” (Hochschild Mining 2013)
- (16) “Since the exploration, appraisal and development phases all consume cash, transparency is perhaps most critical during the operating phase, when significant revenues accrue to our host governments.” (Tullow Oil 2013)

In another fifteen cases, these sentences represented environmental impacts on financial growth that are (almost) inevitable for the extractives industries, which are most present in the subcorpus, e.g.

- (17) “In August 2012, PanAust completed the acquisition of the balance of tenements (which the Company did not already own) over the Carmen deposit and commenced drilling.” (PanAust 2013)
- (18) “With more people moving into cities, world population rising and living standards improving, all forms of energy will be needed to meet demand.” (Royal Dutch Shell 2013)

Given how delicate the balance between pursuing the different aspects of corporate sustainability can be, it is rhetorically remarkable that only 2.6% of the subcorpus represents any tension between different aspects at all, and over half of that fraction presents a financial disadvantage as a boon to non-financial aspects of performance. While these numbers do not represent tensions beyond sentence boundaries, these proportions do not seem entirely representative of the potential drawbacks of the extractives industry in terms of non-financial sustainability.

The sentences without sentiment in either direction were slightly more diverse; recurring categories included the following:

- (19) The strictly informative, e.g. “The Carmen deposit is located 14 kilometres southwest of Inca de Oro.” (PanAust 2013);
- (20) The interpersonal, e.g. “This is our third stand-alone sustainability report and I hope you find it informative.” (Arrium Limited 2013);
- (21) Metadiscourse, e.g. “We report progress against our CR KPIs below, and provide an update on our CR performance and significant activities in 2012 on page 48 of this report.” (Kazakhmys 2013); and
- (22) Rhetorical flourishes and questions, e.g. “What does the future hold?” (Gem Diamonds 2013)

Given the above proportions, we might also wonder to what extent this distribution of content is balanced, i.e. to what extent it discusses the good and bad in equal, or at least representative measure. The former, we can immediately conclude, it almost certainly does not; for every negative element, we find approximately eight positive elements. It is, of course, worth reiterating that these are accompanying Letters to Shareholders rather than full reports. Balance of content in the reports themselves could be different.

In terms of representativeness, we might use the ASSET4 performance scores for the companies in this subcorpus as an approximation of how well they performed relative to the various aspects. The table below contains means and standard deviations for the four performance aspects.

Table 24 Means and standard deviations for performance scores for companies present in the subcorpus.
Multiplied by 100 for ease of use.

Performance				
Aspect	Economic	Environmental	Social	Governance
Mean	67.26	73.88	79.85	73.71
Standard Deviation	24.76	21.11	15.87	23.86

As most companies present in the subcorpus score above 50%⁶ on the various performance aspects, we might allow for some skew before considering the LtSs unrepresentative in terms of content. In doing so, we must acknowledge that both these performance scores and the tallies of positive and negative elements in the subcorpus are simplified representations of vastly more complex realities. Nevertheless, we might still argue that when companies score, on average, no more than 80% out of a total 100 in the optimal scenario of social sustainability, a balance of four positive elements to every negative might be more representative of four points obtained for every point not obtained than the current one of approximately eight. That is, although it is a somewhat reductive reasoning, a score of 80% implies 20% of potential targets not achieved, which leads to a ratio of four achieved for every one not achieved, which is only half the ratio that the corpus currently exhibits, and most companies scored worse than 80% for most aspects. Although this reasoning does not account for the universal positivity bias in language, eight positive elements for every negative may nevertheless be excessive, especially in contrast with the roughly three or two-to-one ratio Wen (2014) finds for positive and negative lexical items in annual reports from 2010.

The relative balance of themes is equally notable; there are more sentences in the subcorpus that contain positive sentiment regarding social sustainability than there are sentences containing no sentiment at all regarding social performance. As social performance is, on average, the best out of the four, this may be an argument for the LtSs being more thematically representative. Environmental performance follows social performance as a distant second, though no longer eclipsing the number of sentences that do not reflect on it. While governance performance roughly approximates environmental performance, it is by far the least-discussed topic, which meets expectations regarding its status as the most opaque aspect of corporate performance to the wider stakeholder audience; financial performance, in spite of it being poorest on

⁶ Companies within this subcorpus may already be above average performers due to the requirements of being selected for the subcorpus; we might expect that companies that issue a standalone sustainability report and preface it with an introductory letter are generally more concerned with non-financial or sustainability performance than those that do not.

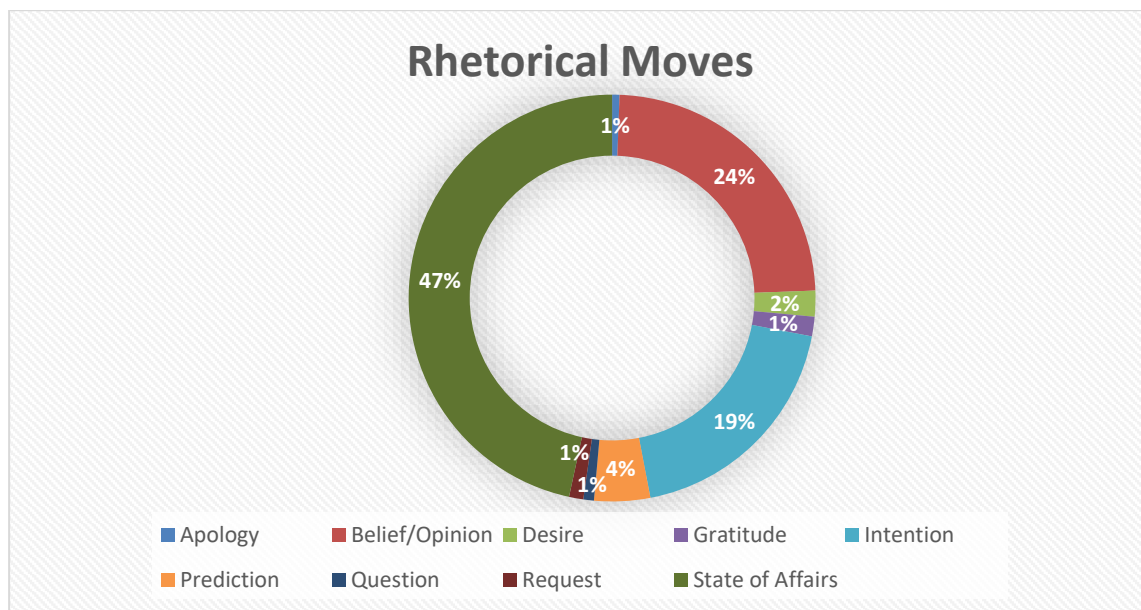
average, is the third most addressed topic. In many sentences, however, this may only be by implication.

5.7.2 Rhetorical Moves

Table 25 Tallies for rhetorical move types for all annotated sentences.

Rhetorical Move								
Apology	Belief/Opinion	Desire	Gratitude	Intention	Prediction	Question	Request	State of Affairs
14	568	49	37	452	106	20	27	1111

Figure 4 Proportions of rhetorical move types.



Based on the table and chart above, we can immediately observe that our annotators perceived that slightly less than half the sentences in the subcorpus (46.6%) primarily reported a state of affairs. As this is followed by sentences expressing a belief or opinion (23.8%), we might argue that although these letters almost invariably strive to maintain a business-like tone, although ‘business-like’ does not necessarily equate to strictly factual. The third most frequent category, declarations of intention (19%), might intuitively appear the more intrinsically linked to these letters, as they frequently signal how companies intend to approach future performance and challenges. It is fairly predictable in that respect that predictions (4.4%) are the fourth-most frequently tagged category, although they are a distant fourth, and the final perceived rhetorical move to occur more than 50 times throughout the subcorpus.

We do, find examples of every sentence type; in descending order of frequency, these are a few examples of prototypical sentences for the categories occurring fewer than 50 times throughout the subcorpus:

- Desire, e.g.

(23) “We want to make lasting improvements to our operations to increase the safety of our employees and contractors and are stepping up our efforts in this regard.” (Eurasian Natural Resources Corp. 2013)

(24) “As our vision states: We want to see ourselves, and be seen, as an industrial group made up of people who live and promote a culture of safety through our daily actions.” (Saras, 2013)

- Gratitude, e.g.

(25) “I would like to take this opportunity to thank all our stakeholders for their support in 2012 and I look forward to working with them in the future.” (Oz Minerals 2013)

- Requests, e.g.

(26) “As we move up the learning curve, I look forward to your valuable feedback to advance our sustainability performance and make this progress more holistic.” (Oil & Natural Gas 2013)

- Questions, e.g.

(27) “So is everything perfect in the world of the Adidas Group?” (Adidas 2013)

- Apologies, e.g.

(28) “Before discussing any other subject matter, we want to express how deeply saddened we are by the Underground QMS Training Facility tunnel collapse in May 2013 that resulted in 28 fatalities and serious injuries.” (Freeport-McMoRan 2013)

(29) “In early 2013, ExxonMobil Pipeline Company responded to a crude oil spill in Mayflower, Arkansas, a regrettable event for which we are deeply sorry.” (Exxon Mobil 2013)

The final example, (29), is notable in being the only direct apology in the subcorpus; a further three express their “condolences”, and a further eight express “regret”. Similar to what outcomes for agency will indicate, these outcomes support the notion that companies will want to rhetorically distance themselves from unfavourable results. In the case of the sole ‘sorry’ in the subcorpus, we might assert that the magnitude of the event might have been such that while not acknowledging or addressing it (or trying to obfuscate it) would have been too costly to be pragmatic, this is a rare instance where the same would have been true of not apologising directly. This example is consistent with Aerts, Peng & Tarca (2013), who assert that causal explanation can be more costly to a company, and thereby generate additional credibility; it is difficult to imagine a more costly causal attribution for a company to make than claiming direct ownership of an environmental disaster.⁷ Accordingly, the company can expect readers to perceive its

⁷ We might argue that of all the events addressed in this subcorpus, the aftermath of the Marikana miners’ strike was likely the most impactful and greatest corporate social responsibility crisis, even more so than the Mayflower oil spill. Given the legal and social complexity of the event, however, it may have been unwise for

apology as highly credible. A more optimistic reading of the situation is, of course, that the company or author(s) of the letter feel sufficiently responsible that they feel the need to apologise regardless of impression management concerns.

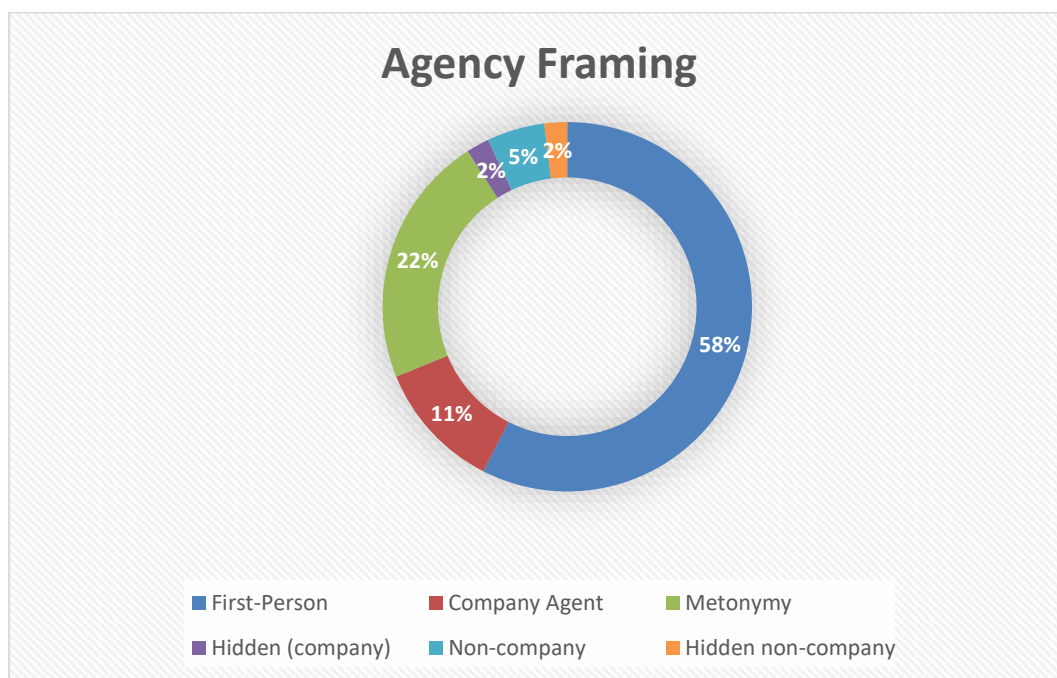
5.7.3 Agency

The table and chart below tally the various instances of companies expressing (their or others’) agency throughout the subcorpus. In cases where there were multiple agents, annotators chose the option closer to first-person agency.

Table 26 Tallies for agency types for all annotated sentences.

Agency					
First-Person	Company Agent	Metonymy	Hidden (company)	Non-company	Hidden non-company
1295	255	498	45	114	45

Figure 5 Proportions of agency framing types.



First-person agency ('I' or 'we') occurs in the majority (54.3%) of sentences according to our annotator's tags, with metonymy a distant second (20.9%) One prototypical example of a company-through-metonymy agent would be:

Lonmin to apologise directly rather than expressing regret as doing so may have been construed as an admission of culpability, and the credibility generated may not have weighed up against other (e.g. legal) consequences of admitting fault.

(30) “Guiding this sense of responsibility is ROC’s Sustainable Practices Framework.” (ROC Oil, 2013).

The company (name) directly serving the agent is about half as common (10.7%):

(31) “Over the coming years, Paladin will continue to appropriately upgrade and expand its sustainability reporting as part of our commitment to stakeholder communication.” (Paladin Energy 2013)

(31) is a typical sentence for this category. According to annotators, slightly fewer than 6.7% of sentences used an agent other than the reporting company, be it an implied (1.9%) or, far more commonly an overt one (4.8%). Typical examples here would be:

(32) “This is reflected in the continuing levels of external recognition for our work.” (Anglo American 2013; hidden non-company agent)

(33) “GRI verified this year’s report at the B+ Application Level.” (Premier Oil 2013; overt non-company agent)

(Parts of) the company as a hidden agent tied as the rarest (at, again, 1.9%); one example here is:

(34) “The likelihood of major safety and environmental incidents is minimised.” (ROC Oil 2013).

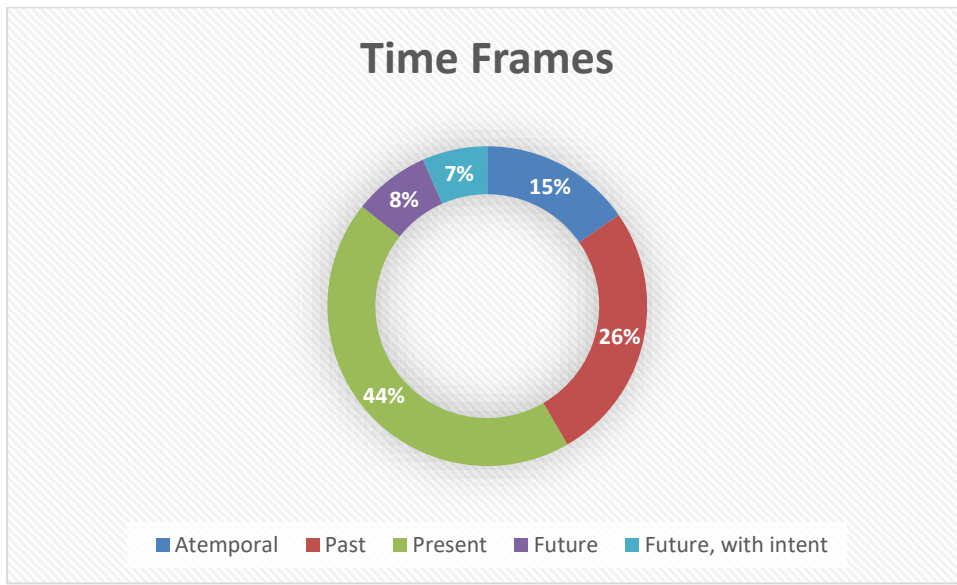
If we look purely at the frequencies, from a defensive attribution and impression management point of view, the observation that the majority of sentences contain a first-person company agent aligns with most sentences containing (exclusively) positive sentiment. That is, these findings do not contradict the notion that companies will attribute favourable outcomes to themselves and unfavourable outcomes externally. Of course, these data do not necessarily show that first-person agency and positive sentiment will tend to co-occur just because they are the largest groups. Section 0 will examine the extent to which the presence of positive or negative sentiment in a sentence can predict its agency framing.

5.7.4 Time Frame

Table 27 Tallies for time frame types for all annotated sentences.

Time Frame				
Atemporal	Past	Present	Future	Future, with intent
366	627	1050	184	157

Figure 6 Proportions of time framing types.



We asked annotators to indicate the dominant time frame in every sentence they encountered; the table and chart above summarise the results. They perceived just under half (44%) of the sentences to employ a chiefly present time frame, such as in:

- (35) “The name Arrium better reflects what we are today, as well as our focus on growing the Arrium Mining and Arrium Mining Consumables businesses.” (Arrium Limited 2013)
- (36) “This sale is consistent with our strategy to pursue long-life, low-cost operations.” (Newcrest Mining 2013)

The next most frequent time frame is the past (26.3%), including such statements as:

- (37) “I believe that over the last financial year we have made significant steps in the area of sustainability and 'Licence to Operate' and commend this report to you.” (Newcrest Mining 2013).

The next most frequent time frame was the future (14.3%), of which slightly less than half indicated a clear intent on the company’s part. Examples with and without such signalling respectively include:

- (38) “The important thing, however, is that we are committed to tackling these challenges and that they spur us on to reach our sustainability goals.” (Adidas Group 2013)
- (39) “The 2013 financial year is expected to be challenging.” (Aquarius Platinum 2013)

Lack of any demonstrable time frame, finally, is the least common, with (sentences presented as) general truths such as the following prototypical examples:

(40) “Sustainability is an integral part of Santos.” (Santos 2013)

These results suggest that annotators’ perception of the dominant time frames throughout the subcorpus place it mostly in the present and past. This is, to some extent, at odds with Parsons & McKenna’s (2005) findings, although we must note that it is still possible to frame promissory statements in the present or atemporally. First, companies may still be expressing ideas that are difficult or impossible to falsify, but in a mostly present-time frame. Second, if we assume that better-performing companies have more reason to discuss actual performance rather than hypotheticals and intentions, the companies in this generally well-performing subcorpus may have less incentive to make such statements. Third, concreteness and falsifiability may have increased in general since Parsons & McKenna’s (2005) study as the genre continues to be a rapidly evolving one.

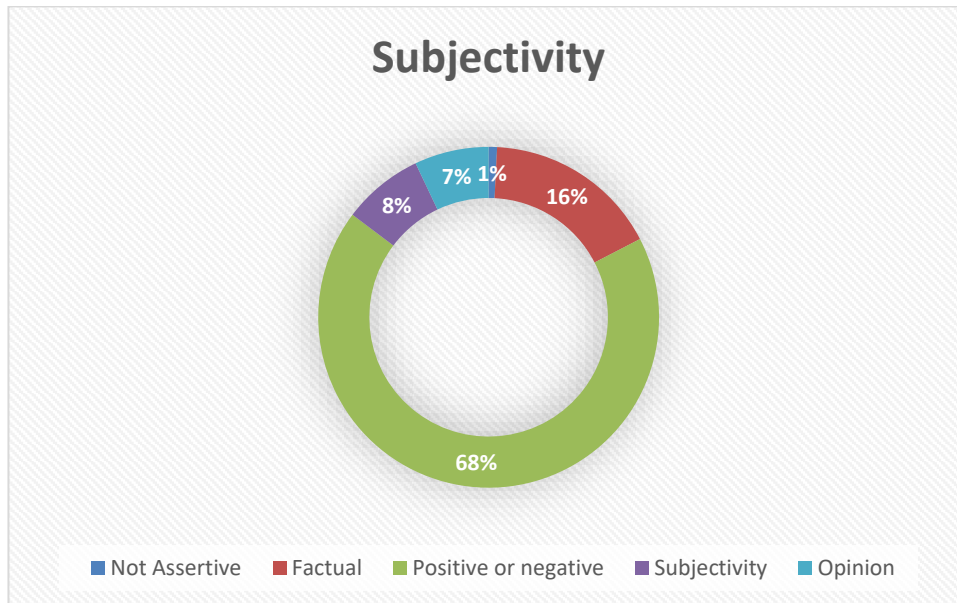
5.7.5 Subjectivity

In line with the above, we wished to investigate to what extent sentences in these letters were subjective, as the concept is closely related to falsifiability but easier for annotators to grasp. Nevertheless, inter-annotator agreement for this category was the lowest out of the categories to which we asked annotators to assign a tag. While the choices they made throughout the subcorpus still merit discussion, this indicates that we should only do so in very tentative terms, and will not pursue further (e.g. regression-based) analysis. The table and chart below tally annotators’ perceptions of sentences’ subjectivity.

Table 28 Tallies for subjectivity types for all annotated sentences.

Subjectivity				
Not Assertive	Factual	Positive or negative	Subjectivity	Opinion
19	396	1618	183	168

Figure 7 Proportions of subjectivity types.



While we must account for considerable variability for these tags, it is nevertheless striking (though within expectations) that a considerable majority of sentences in the subcorpus (67.9%) are presented as factual but coloured positively or negatively. Examples tagged as such constructions include:

- (41) “Further, our two core values of ‘safety’ and ‘customer’ remain at the forefront of everything we do.” (Arrium Limited 2013)
- (42) “The 2012 Financial Year was a period of unprecedented growth for Macmahon, with a record profit, revenue, order book and the largest number of direct employees in the Company's history.” (MacMahon 2013)
- (43) “This sale is consistent with our strategy to pursue long-life, low-cost operations.” (Newcrest Mining 2013)

These sentences illustrate the continuum between subjectivity and objectivity that annotators will likely have struggled with. For instance, ‘safety’ and ‘customer’ are simply ‘good’ words with positive associations in a CSR reporting context, but it is difficult to assess the extent to which these can objectively be said to be at the forefront of operations. Conversely, Newcrest Mining may objectively have a strategy to pursue long-life, low-cost operations (that is, there are likely strategy documents available), but their perception of whether a given sale aligns with that strategy remains subjective, however well qualified they might be to make such a judgment. Conversely, Macmahon may be able to prove based on previous years’ records that 2012 was unprecedented in terms of

the variables they indicate. This statement is entirely falsifiable, in spite of containing strong positive colouring.

The next most frequent category (16.6%) was the one that annotators perceived as purely factual, i.e. without positive or negative colouring. Examples include:

- (44) “In August 2012, the PanAust Board declared the Company's maiden interim dividend of three cents per share and in February 2013, the Board declared a final dividend of four cents per share.” (PanAust 2013)
- (45) “To date, 68 people have obtained permanent positions at Prominent Hill through this program, of which 70 percent are Aboriginal people.” (Oz Minerals 2013)
- (46) “The depressed uranium price in the wake of the events in Japan in March 2011 has put significant pressure on our Company, concerned our investors and compelled us to more closely watch and prudently manage our costs, debt and resources.” (Paladin Energy 2013)

While the first two examples are fairly clear examples of purely factual states of affairs that may nevertheless – through implication – reflect favourably on the company’s financial or social performance respectively, the third example draws closer to the category of ‘positively or negatively coloured’. While the core idea that the uranium price was depressed is falsifiable, other ideas may not be entirely so. Few would dispute that flagging sales compel a (competent) company to do prudent business, and fewer would argue that external events lowering sales prices will concern (an intrinsically coloured word in terms of sentiment) that company’s investors. Fewer still would dispute that the aforementioned events were not responsible for lowering uranium prices. Nevertheless, these causal connections are, by nature, far less falsifiable than they are easy to understand and agree with intuitively. This potentially weak contrast with the ‘positively or negatively coloured’ tag is a good indicator of precisely why this aspect of the annotation task was so difficult; Van De Kauter, Desmet & Hoste (2015) encountered similar issues.⁸

Less frequent still (7.7%) are sentences that annotators perceived to contain subjectivity but no rhetorical devices that mark the sentence as such. Some of the less ambiguous examples include:

- (47) “Our corporate social investment (CSI) programmes in 2012 were extensive as were our training and development programmes.” (Aquarius Platinum 2013)
- (48) “We are accountable for our decisions and in communicating our actions to our stakeholders.” (Eurasian Natres Corp., 2013)

As there were relatively few patterns in the results, however, this appears to be a tag with which the annotators struggled especially, which again partially explains the

⁸ See Van De Kauter, Desmet & Hoste (2015) for a more abstract interpretation of subjectivity, which we chose not to implement.

comparatively low inter-annotator agreement scores. In contrast with annotations for the ‘explicitly marked as opinion’ tag (7%), we find more consistency, e.g.:

- (49) “I would also like to make special mention of entries received for the Safety Excellence Awards from Moly-Cop, Chile, one of which was highly commended.” (Arrium Limited 2013)
- (50) “To combat this, I believe that we should play our part, along with other stakeholders, in formulating a new informal ‘social compact’ for business that encourages greater transparency, better governance, a shared understanding of the role and value of business to society, and accountability for our actions - of which this report forms part.” (Anglo American 2013)

These outcomes (tentatively) indicate that overall, these companies used fairly few overt subjectivity markers such as ‘I/we believe...’ in spite of the texts’ substantial amounts of positive (and to a lesser extent negative) elements regarding the four aspects. However (and more crucially), we chiefly find that annotating factuality is a difficult task, but that giving annotators very binary instructions (e.g. only assigning a certain tag if a specific element is present, such as a subjectivity marker like ‘I believe’) does enable more systematic annotations. While this is not in the least surprising, it remains an important consideration to carry forward into sentiment annotation tasks on corporate reporting, especially ones that try to account for multiple aspects (e.g. the different performances) or implicit sentiment, as this one attempted to.

5.8 Analysis

While the preceding section already provided considerable support for H1 (more positivity), we conducted a number of further analyses to investigate H2 (no relation with performance) and H3 (relation with agency framing).. We performed all analyses with SPSS version 23, with alpha level set to 0.05, and in analyses 5.8.2 and 5.8.3 controlled for the annotator as a random variable.

5.8.1 Sentiment and Performance

In a first analysis, we calculated the correlations between ASSET4 score for a given performance aspect as the dependent, and the following variables as predictors:

- Number of sentences in text
- Number of sentences relevant to performance aspect
- Proportion of relevant sentences relative to all sentences in text
- Average score across relevant sentences
- Average score across all sentences

Sentences ‘relevant’ to a performance aspect in this analysis were those to which annotators had assigned a sentiment score for that aspect. Table 29 summarises these correlations.

Table 29 Pearson correlations between ASSET4 performance measures and sentiment measures for that aspect.
‘Relevant’ sentences are those with a sentiment score between -3 and +3 assigned for that aspect by annotators.

ASSET4 Performance Score:	Financial		Environmental		Social		Governance	
	Coeff.	p	Coeff.	p	Coeff.	p	Coeff.	p
Measures:								
Total number of sentences (regardless of topic)	0.192	0.108	0.251	0.035	0.341	0.004	0.032	0.793
Number of relevant instances	-0.019	0.875	0.28	0.018	0.227	0.057	0.206	0.085
Proportion of relevant sentences	-0.148	0.218	0.012	0.918	-0.254	0.032	0.285	0.016
Average sentiment across relevant sentences	0.02	0.868	0.048	0.693	0.038	-0.756	-0.126	0.351
Average sentiment across all sentences	-0.065	0.593	0.004	0.973	-0.237	0.047	0.271	0.022

Remarkably, while there are some significant correlations we can observe that there is never a significant correlation between the ASSET4 performance score and the average

sentiment across sentences relevant to that performance score, which is consistent with the Pollyanna effect, especially considering the high degree of positivity throughout these documents.

However, we can note a number of significant correlations, chief amongst them those between the performance measure and sentiment across *all* sentences (i.e. also counting those not relevant to the performance aspect) for social and governance performance. Both show medium-small correlations, but what is especially remarkable is that social performance correlates negatively with attention to social performance (i.e. proportion of social-themed sentences) and social sentiment averaged across all sentences. Equally notably, social performance correlates positively with longer letters. This suggests that socially responsible companies may issue longer LtSs in their sustainability reports, but may not pay commensurate attention to or discuss positive outcomes involving their social performance. This particular finding interacts strangely with the Pollyanna Effect in that it suggests that good social performers may strive to avoid excessive positivity. This finding merits further inquiry, especially through more qualitative means similar to e.g. Crilly, Hansen & Zollo's (2016) exploration of motivations behind CSR practices.

Similarly, for environmental performance we find that longer letters and more sentences (in absolute terms) dedicated to environmental performance correlate with better environmental performance. Remarkably, neither the average environmental sentiment nor proportion of attention paid to the topic exhibit that same correlation, which aligns with the Pollyanna Effect. The main point of evidence against the Pollyanna Effect is the medium-small correlation between governance performance and governance sentiment across all sentences. However, this too deserves the nuance that attention to governance performance (i.e. proportion of relevant sentences) itself shows a positive correlation with governance performance (and further analysis indicates that attention to governance sentiment and average governance sentiment across all sentences correlate with a coefficient of .985). As such, while these correlations provide interesting incentives for future analyses, they do not discredit the Pollyanna Effect. These findings do, however, only lead to a partial acceptance of H2.

5.8.2 Industry and Region

We also constructed a general linear model with each of the documents' sentiment scores (i.e. financial, social, environmental and governance) as dependent variables that included region and industry, as well as their interaction, as independent variables. This analysis would help us detect patterns in use of sentiment along industry or region lines.

We found a single significant predictor (which suggests it is likely a fluke) in semiconductor LtSs exhibiting more positivity regarding financial sentiment than either of the extractives industries did. The significance of the overall association between

industry and financial sentiment was 0.005, at a small to medium effect size of 0.198. After Bonferroni correction, we saw a significant difference between semiconductors and both mining ($p = 0.019$) and oil ($p = 0.024$). As there was no significant association between sentiment and performance (and thus no reason to attribute this to a performance gap between the two industries), this effect may well be a random occurrence in the data. As neither region nor industry proved significant predictors in any other cases, we see little reason to expect an association between language variety or industry characteristics and positivity or negativity. While this might be a viable avenue for future research, we see little reason to explore it in greater detail in the present study.

5.8.3 Agency

Subsequently, on a text level, we investigated to what extent the performance underlying a text could predict its average agency rank (expressed as a continuous variable). While performance for governance registered below the threshold of significance at $p = 0.033$, drawing the regression line through a scatterplot revealed that this was likely a fluke, as there was no visual pattern to the data, nor is there a likely reason why governance should be more influential than one of the other performance measure in influencing agency framing.

On a sentence level, however, we fit an ordinal logit model that used three variables as predictors:

- The presence (or absence) of any positive information
- The presence (or absence) of any negative information
- The type of rhetorical move in the sentence

The table below describes the outcomes. A lower rank implies agency framing closer to first person, the presences of both sentiment types are relative to their absences, and rhetorical move types are relative to depicting a state of affairs.

Table 30 Summary of ordinal (logit) regression predicting agency patterning for n = 2384 sentences.

A positive coefficient means more first-person agency framing if the variable is present.

Marginal significance ($p \leq 0.1$) in *italics*, significance ($p \leq 0.05$) in ***bold italics*** and strong significance ($p \leq 0.01$) in ***underlined bold italics***.

Variable	Significance	Coefficient
Positive sentiment	< <i>0.001</i>	<i>0.603</i>
Negative sentiment	0.145	-0.227
Apology	<i>0.001</i>	<i>2.587</i>
Belief/opinion	< <i>0.001</i>	<i>601</i>
Desire	< <i>0.001</i>	<i>1.227</i>
Gratitude	< <i>0.001</i>	<i>1.327</i>
Intention	< <i>0.001</i>	<i>1.238</i>
Prediction	0.06	-0.346
Question	0.069	-0.76
Request	<i>0.003</i>	<i>1.257</i>

As all but three input variables exceed the threshold of significance, we can safely say this model is much more informative in explaining agency patterning (on a sentence level). The presence of positive sentiment proves a significant predictor for agency framing closer to the first person, as H3 predicted, but the presence of negative sentiment does not have the opposite effect⁹.

Equally notably, with the exception of predictions and questions, shifts away from rhetorically depicting a state of affairs also shifted the agency framing closer to the first person. This aligns with the earlier observation that attempting to practice defensive attribution when, for instance, making an apology might create an undesirably jarring message. Accordingly, these outcomes indicate that as LtS highlight the discursive frame more (e.g. through expressing gratitude compared to depicting a state of affairs) they become more inclined to use the first person.

5.9 Discussion & Conclusions

In summary, we find evidence in support of all three hypotheses, although only partially so for H2 and H3 (i.e. agency framing). That is, LtSs from sustainability reports contain

⁹ We note that in a model including only the presence or absence of both negative and positive sentiment (i.e. omitting the rhetorical move as a predictor), significance for the ‘negative sentiment’ variable was 0.042, and the estimated change for its presence was 0.302.

more positive than negative information, and the balance thereof shows no significant association with company performance. However, the relationship between positive sentiment and defensive attribution tactics (see e.g. also Aerts & Yan 2017) is considerably more complex than the third hypothesis managed to capture.

Regarding H1, we indeed found that the presence of positive sentiment far outweighed that of negative and no sentiment. While we can partially attribute this to companies in the subcorpus performing well overall, a ratio of eight positive elements to each negative one still aligns with the Pollyanna Effect, suggesting that it is indeed present in sustainability reporting. Especially in combination with more than half of sentences containing positively or negatively coloured elements (but chiefly the former), it seems fair to assert that sustainability report Letters to Stakeholders do not appear to limit themselves to neutral language.

Whether these LtSs are balanced in the sense of being representative is another matter. Given that most companies present in the subcorpus performed above average (that is, 50/100) on the ASSET4 performance scores for the same aspects we collected sentiment data on, we might argue that a greater ratio of positive information is representative of actual performance. However, as we saw no significant association between the presence of sentiment information and performance for a given aspect (or even the length of the letter), and although some of it is attributable to a positivity bias in all language regardless of genre, we can conclude that this 'excessive' positivity appears to be regardless of actual performance. This leads us to accept H2 and expect to see little variation in LtS sentiment balance based on performance during the year reported on, with the potential exception of extreme or highly notable cases, i.e. reputational crises such as the deaths at Marikana or the Mayflower oil spill.

The same effects that make reputational crises an interesting avenue of study motivate our partial (rather than full) acceptance of H3. Both the sentiment of a message and its rhetorical positioning appear to affect that message's agency framing, and positive sentiment (possibly due to its greater presence in the corpus) appears to be more influential than negative sentiment in this respect. From a sentiment perspective, the presence of positive sentiment appears to bring its agency framing closer to the first person. Interpreting this from a defensive attribution perspective, we might consider this evidence of companies strategically – arguably opportunistically – attributing positive outcomes to themselves. Conversely, if we take into account the the main rhetorical move being presented, the opposite does not hold true with statistical significance. That is, although we can observe that, on average, companies will distance themselves from first-person narrative when reporting negative outcomes, the difference between two scenarios is not sufficiently large to be statistically significant when controlling for the sentence's primary rhetorical move.

Regarding the influence of the sentence's rhetorical move type on agency framing, the most remarkable (and statistically influential) type is likely the apology. While we might

expect that a company issuing an apology would rhetorically distance themselves from the matter through agency framing (with the most typical example being the cliché of ‘mistakes were made’), there is no more powerful predictor of first-person agency framing in the model than it containing an apology. What is more, based on the outcomes for the other potential rhetorical moves, we find that reporting companies in LtSs most distance themselves from the subject matter through agency framing when portraying a state of affairs. In causal terms, there is likely a feedback loop: Companies wishing to distance themselves from unfavourable information will typically need to combine both the most factual possible rhetorical framing and the most distant possible agency framing. That is, in the following (fictitious) examples, the defensive attribution achieved by sentence (51) is rhetorically greater than the sum of its parts:

(51) Difficult conditions forced the industry to carry out layoffs.

(52) Due to difficult conditions in the industry, layoffs were necessary.

(53) Difficult conditions forced us to carry out layoffs.

As previously alluded to, the same reasoning might explain why apologies, though infrequent, appear to nevertheless be the most powerful predictor of first-person agency framing. Given the extent to which an apology highlights the interaction between the author and the audience (as opposed to depicting a state of affairs, which does not), the damage is likely already done in terms of defensive attribution, with the consequence that presenting sincerity through first-person agency has very few downsides.

Finally, regarding Parsons & McKenna’s (2005, p.603) findings of “[unverifiable] promissory statements with no declared timeframe”, it is notable that statements framed in the future (with declaration of intention or otherwise) and atemporal framings are distant third and fourth places to the most frequent present timeframe and second most frequent past. While several of these problematic sentences Parsons & McKenna refer to might also fall into the ‘present’ timeframe, it is notable that the ‘past’ timeframe – which is likely the most vulnerable to challenge – is the second most frequent (as their own study also finds for the single report they investigate). In other words, it appears that, while these LtSs certainly do convey ideas that might be difficult or impossible to challenge (as the preceding example sentences can illustrate), they are certainly not limited to such content alone.

The interpretations and conjectures above are, of course, preliminary, based on this exploratory study. Nevertheless, given the complex interplay of elements this study was able to discover through a combination of qualitative methods further underlines the importance of continuing inquiries into how corporate reporting frames its positive, and especially its negative outcomes.

Chapter 6

Human Assessment and Machine Learning

6.1 Motivation

While formula-based readability is a cornerstone of this study, the formula-based approach is, out of all the readability assessment techniques we have access to, the most limited in explaining *why* it assigns a score to a piece of text. A technically proficient user knows that a lower formula-based readability will stem from a higher average number of syllables, a higher average number of syllables, or both, but the scores do not identify which. Especially because these are ‘shallow’ textual features, a readability score does not explain why the text is more or less readable. For example, a text with many polysyllabic words, and generally short sentences amongst a few extremely long ones might yield approximately the same score as a text with short words but generally long sentences. Measuring syntactic features can be more informative, but is less intuitively understandable and less suitable to comparison between texts and genres. More passives (and, we hypothesised deeper parse trees) make a text less readable because these features directly increase cognitive load.

The significance of differences in scale is easier to gauge than that of differences in kind: we can presume that a text with twice as many passive structures as another one may be more difficult to understand, even if it contains five per cent deeper parse trees. Imagine, however that one text contains 20% more passive structures and 20% shallower parse trees than the other. Which is easier to understand? Such a comparison is almost impossible to answer in a vacuum. We are better equipped to answer it if we understand the particularities of the genre and its audience’s expectations. This section explores how we set out to further investigate and operationalise the readability variation we found in the full corpus analysis (Chapter 3) in greater detail.

Our approach fed into two of this study’s aims:

- supplementing formula-based readability by creating a genre-specific readability measure through machine learning;
- gaining insight into which aspects of corporate reporting make it less readable or understandable.

The latter objective contributes to the former: if we know why some texts are easier or more difficult to read or understand than others, we are better able to fine-tune the readability measure. This approach does, however, come with a crucial caveat that prompts us to explore machine learning system and the scores and comments that it relies on as at least partially distinct areas of inquiry. That caveat is that there is nothing ensuring that those features that (human) readers describe as contributing to or detracting from readability match those that best help a computer analysing a piece of writing predict or infer the ease with which a human could read it.

As an example, as the readability formulae's long history as a readability metric can illustrate, it predicts human reading ease with enough precision to be fit for purpose. However, none of the readability formulae we have explored have access to more than sentence and word length in order to make that prediction. By contrast, as section 2.4 explored and section 6.3 will illustrate, a reader who finds a text difficult will likely do so for different – or at least more – reasons than the average length of words and sentences (see e.g. Kleijn 2018). Rather, word and sentence length mainly show meaningful associations with those aspects of the text that increase perceived difficulty. This makes it crucial not just to report on the performance of a machine learning-based readability predictor, but also on those aspects that motivated the scores that system is trying to predict. We expect the most influential factors in either to at least be partially disjunct. That is, the strongest reasons why humans might attribute a particular readability to a text are not necessarily those that best predict the readability they will assign from a statistical point of view when quantified. Including both perspectives better enables authors to optimise for the people reading their work, rather than optimising for automatic assessment systems.

6.2 Scoring

6.2.1 Implementation

As a first step towards achieving both objectives, we submitted excerpts from the sustainability report body text in our corpus to a group of language experts in training and asked them to rate the difficulty of each text on a scale from zero (easiest) to one hundred (most difficult). This provided training material for the machine learning goal. We also asked participants to (briefly) explain the reasoning behind their score, to further the second goal of knowing what causes one report to be more difficult than the other according to human perception.

6.2.2 Participant Selection

We selected third-year BA students, as well as MA students, belonging to Ghent University's Applied Linguistics and Linguistics and Literature tracks in equal measure. Outside of being highly viable candidates from a practical point of view, they were also a sensible choice from a design perspective. Ideally, we required participants with a linguistic proficiency and familiarity with the subject matter comparable to a wide cross-section of sustainability reporting's audience, as well as the linguistic insight and acuity to meaningfully describe their assessment of the text's readability. Experienced second- or third-language users of English may be more linguistically proficient than the average user of sustainability reporting and therefore less representative, but offer an attractive trade-off in their technical and theoretical insight into how language works. Mixing Applied Linguistics and Linguistics and Literature students enables a more heterogeneous level of experience with business communication, as the Applied Linguistics curriculum at time of writing included business communication whereas Linguistics and Literature did not.

Before selecting any participants for the final stage of the experiment, we invited them for a 45-minute exploratory session. This consisted of 15 minutes of instructions and explanation, which offered the participants the opportunity to ask questions, and 30 minutes of scoring as a test exercise. Appendix 4 (Human Assessment & Machine Learning) contains a few examples of such texts. Based on their output, we selected 24 participants out of a total 28 (ten during a first stage and fourteen during a second) who we felt best carried out the task for the first phase of scoring. They proceeded into the second, more extensive phase of scoring. One candidate did not wish to proceed into the second phase, and we eliminated the remaining three based on what we judged to be a relatively poor speed, accuracy and/or insightfulness relative to other candidates. That is, given the same amount of time these candidates scored fewer texts and assigned fewer or more shallow comments than others did.

6.2.3 Subcorpus selection

We selected the excerpts for the experiment from the sustainability report body text; i.e., this experiment did not draw from the financial or sustainability-oriented LtSs. In doing so, we slightly favoured the comparability that similar-length excerpts offer, over the representativeness that using full-length LtSs would have done. This is the opposite choice from our sentiment analysis study (Chapter 5) as well as the readability manipulation experiment in Chapter 4. The latter only needs to ensure comparability between one set of three texts rather than arbitrarily large sets drawn from several hundred texts, while completeness and context are much more crucial to per-sentence

sentiment annotation. Furthermore, per-sentence annotation already ensured greater comparability. As LtSs can vary in length from a few hundred words to several thousand, asking participants to compare the readability of texts between the two extremes would be less valid than asking them to compare texts of roughly similar length.

We tried to extract two excerpts per text where possible: the first thematic whole that fit our criteria from the text's initial pages, and the first after the 100-page mark, or counting back from the end of the document when it contained fewer than 100 pages. De Clercq et al. (2012) found that in terms of readability, a sufficiently long excerpt adequately represents the full text, and we strove for a balance between more general introductory material (excerpts from the beginning of the document) and more detailed, fine-grained narrative on one specific aspect of the company (excerpts towards the end).

We aimed for an excerpt length of 300 up to 500 words, i.e. between one to two long paragraphs and a page of text; short enough not to exhaust the reader, but long enough for them not to decide based on a few sentences. We prioritised thematic consistency (the excerpt discusses one thing, or several things that belong to the same theme) within each excerpt as much as possible within those 300-500 words. We favoured shorter excerpts where possible, but did occasionally exceed those limits in favour of including the full sequence of text that covered a specific topic, although seldom by a margin of more than 10%. The other reason for this target number was a technical restriction of the scoring interface. The frame the excerpt appeared in did not support text scrolling (but did support browser text zoom-out), and would thus cut off before the end of the text if it ran too long. We thus needed to ensure that text would still be legible on the furthest necessarily level of zoom-out participants would need on the lowest monitor resolution they would be likely to use. We estimated this at 80% zoom on a 1280x720 resolution. These restrictions made text longer than 550 words impractical.

Figure 8 provides a screen capture of the scoring interface:

Figure 8 Screen capture of the scoring interface. Text on the left, score and positioning relative to other texts in the batch in the middle columns, score entry and comment boxes on the top right, and text functions on the bottom right.



The interface in question was already in place thanks to the efforts of De Clercq & Hoste (2014, 2016), who conducted a similar experiment using a diversified general corpus that contained four types of text, each with more readable and more difficult instances per subcategory:

- Administrative (entailing some corporate reports as well as survey or policy documents);
- Informative (entailing news writing and encyclopaedia articles);
- Instructive (entailing user manuals and guidelines); and
- Miscellaneous (entailing various other text types, including “very technical texts” and children’s literature; De Clercq & Hoste 2016).

Such a broad dataset illustrates the need for genre-specific retraining. The system, when trained on a wide, ‘generic’ corpus would certainly be capable of assigning a score to each text present in this study’s subcorpus. However, this non-genre-adapted implementation would face the same problem as readability formulae do: a much lower resolution in terms of nuance between the scores it can assign. That is, while the non-adapted system would likely be an accurate one from a general readability perspective, it is almost certain that most texts in this study’s subcorpus would be amongst the most difficult that the generic learner had encountered, just as we found for the readability formulae. As it is fairly intuitive that corporate reporting is less readable than (most) user manuals, we would encounter the same problem we do with readability formulae: compared to the greater granularity a genre-specific readability prediction system would ideally enable, a general-purpose yardstick has insufficient resolution to differentiate between these texts

beyond calling them ‘very difficult’. As such, this study’s machine learning implementation makes few attempts to push technical boundaries. Rather, it attempts to advance the state of the art of research into corporate reporting through methods that, while tried and true, are substantially more advanced than the de facto toolset of readability formulae.

6.2.4 Scoring and Annotation Process

As De Clercq & Hoste (2014, 2016) did, we asked participants to score a text’s readability from 0 (easiest) to 100 (most difficult) and explain why they assigned that score, but without telling them how to define readability. Rather, we encouraged them to make their own, open-ended assessments and tell us via the comment box which linguistic elements stood out to them during the reading process. We were able to observe that a single participant typically uses a consistent set of variables (such as vocabulary or syntax) to explain the score they assign, and that the sets of criteria most participants use overlap meaningfully.

De Clercq & Hoste (2016) recognised three main categories of comments on a text’s readability: those related to vocabulary, to structure, and to coherence, respectively, with a fourth ‘other’ category of comments for those that did not belong to those three categories. These categories fairly comprehensively cover the main types of comments we might expect assessors to give when assessing a generic corpus. However, as this study deals with a specialised (sub-)corpus with a reputation for both technical jargon and fairly indirect verbiage, we added two additional categories of comments.

The first additional category was that of comments involving knowledge, i.e. how the extent to which the text requires or assumes background knowledge and disseminates its information. We can justify separating out this category based on (*inter alia*) Bean (2011 & Weimer, p. 136), who presents “background information, allusions [and] common knowledge that the author assumed that the reading audience would know” as a crucial contributor to readability. As sustainability reporting is a specialist genre, we expected issues related to specialist knowledge to occur frequently enough to belong to a separate category.

The second category we added represents comments on the interpersonal aspects of the text or on the rhetorical style which the author employs. While studies ranging from Davison (1985) to Kleijn (2018) have pointed out that style is one of the easier aspects of readability to capture as an author’s choice of words is fairly quantifiable, their ability to create engagement is far more intangible. Nevertheless, those intangible choices can have a significant impact on how a reader engages with the text and how engaging they find it. Bean & Weimer (2011, p. 135-136) indicate “difficulty in appreciating a text’s rhetorical context [and] seeing themselves in conversation with the author” as two impediments to

readers' ability to fully process a text. Consequently, we might argue that those elements that enable readers to appreciate a text's rhetorical context and feel connected with the author belong in a separate category. As a corollary, precisely because participants knew they would be assessing a specialised corpus of corporate communications (a genre with a fairly impersonal and factual reputation), we expected them to be more aware of the genre pragmatics involved. As, based on the aforementioned studies, we did not want to co-opt the label of 'style', we labelled this category as 'tone'.

For each of these categories, we distinguished three types of comments: those indicating that the participant found the text easier because of this feature, those indicating that they found it harder, and those commenting on a feature without indicating its impact or indicating that it simultaneously made the text easier and harder to read.

The following table summarises our criteria for marking a comment as referring to one of the categories, in addition to providing examples:

Table 31 Categories, criteria and examples of scorer comments.

Category	Criteria	Example
Vocabulary	References to word choice and idiom; frequently though ‘vocabulary’ or ‘lexicon’, occasionally as ‘difficult/easy <i>language</i> ’ when used in addition to other categories. Can overlap with ‘knowledge’ category in cases of industry-specific terminology.	‘Very vocabulary-laden, especially the last paragraph was difficult to read and understand because it was incredibly dense as to names and abbreviations.’
Structure	References to textual composition at sentence level and purely structural (but not flow-of-information) choices at paragraph level. Includes references to sentence complexity and length as well as specific issues such as passive voice.	‘The third paragraph in particular was very complicated in both content and form, and the structure of the entire text was at times hard to follow’
Coherence	Outright references to coherence, as well as flow-of-information, as well as paragraph and sentence order. Does not include references to information outside of text, which belongs to the ‘knowledge’ category.	‘This text is a list of disparate items that had to be in there, but there is no real coherence other than the general theme.’
Knowledge	References to delivery of information (e.g. ‘the information present in the text is easy to grasp/recall’) and demands on the reader’s non-linguistic or subject expertise, such as familiarity with company operations, or industry- or company-specific terminology, where it can overlap with ‘vocabulary’ (see below).	‘The second paragraph of this text is difficult to understand because of the numbers and the ISO terms. The other paragraphs are not so difficult.’
Tone	References to engagement and pragmatics, such as formal/informal language and register, level of engagement (‘lively’ or ‘boring’), and general attitude towards reader.	‘Straight-forward but difficult to plow through because of rather dry subject matter’
Other	Anything unrelated or only tangentially related to the above.	‘The first words of the three last paragraphs are not capitalised but they appear to be a part of the list of features, in which the first two do are capitalised.’

The decision whether to count specific terminology as an issue of vocabulary or knowledge was a pragmatic one resolved on a case-by-case basis. It is often impossible to discern whether an informant found the word (group) itself complex or difficult to process, or was able to process the word or word group but not its meaning in the specific context. The former would be a vocabulary issue, while the latter would be a knowledge issue. We most often count abbreviations and acronyms as knowledge issues as the

informants typically indicated they did not have the expertise to process what these meant in context.

In order to balance representativeness and labour efficiency, we randomly selected just over 20% (804 out of 3987) of comments and manually annotated them for the above categories, tallying whether comments indicated greater ease, greater difficulty or some other impact – for instance that it simultaneously facilitates and impedes reading - for those categories¹. Based on both the diversity of comments and their relative consistency in falling into the available categories, we concluded that the 20% sample struck an appropriate balance between representativeness and viability of manual annotation and verification.

We do note, however, that our manual tallying of scores measures how many of these five aspects of readability a text mentions, but it does not measure their intensity. For instance, the (fictitious) assessments of ‘structure has slight room for improvement’ and ‘structure makes text almost illegible’ would be equivalent, although the latter is a far stronger assessment. From both an objectivity and time-efficiency standpoint, the quaternary division between ‘easier’, ‘more difficult’, ‘both/unclear’ and ‘neither’ reduced subjective bias and processing time in simply asking the two binary questions of ‘does the comment say or imply that this aspect improves readability’ and ‘does the comment say or imply this aspect reduces readability’? Expanding those questions to include intensity would have made for a far slower and potentially more erratic annotation process.

6.3 Outcomes

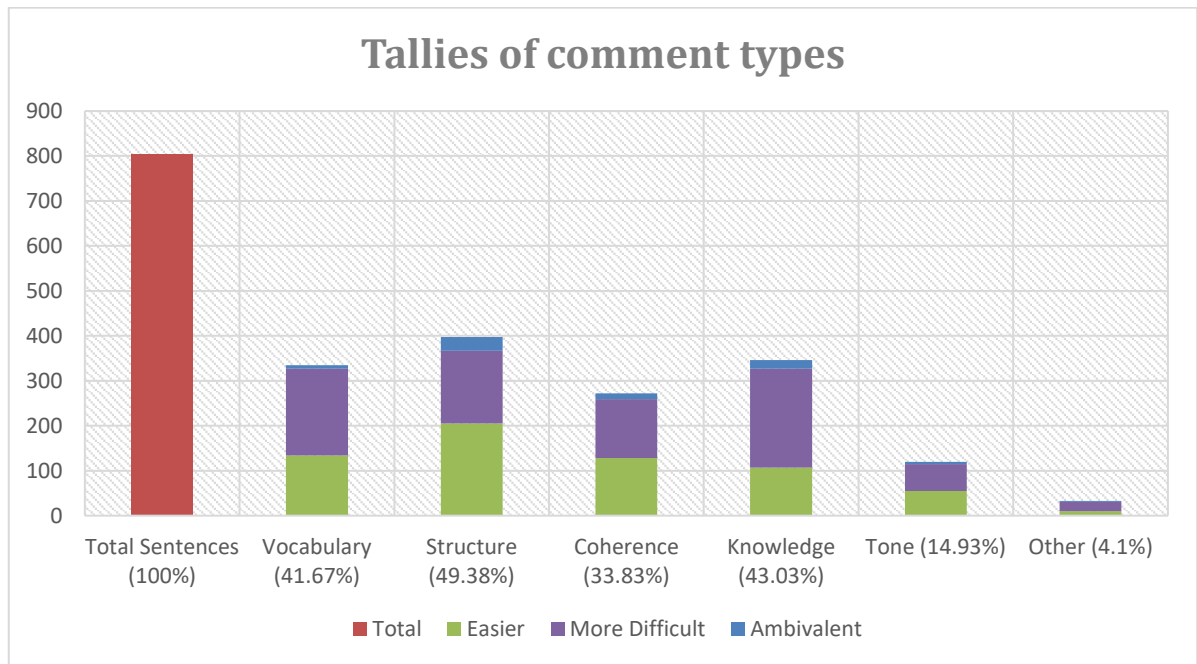
The following table summarises the distribution of positive, negative and ambivalent or neutral comments about the above elements. We expected the latter to prove the rarest; they did. Note that as assessors could draw on multiple elements in order to explain their score (and most did), totals across categories exceed sample size. The graph underneath summarises outcomes.

¹ We conducted this annotation in Excel 2016. We randomised using its ‘RAND()’ function, assigning a random value to each row and sorting them sequentially. After annotation of the comments, we extracted the annotated sample to another spreadsheet, and tallied ‘e’s, ‘d’ and ‘x’es, for ‘easier’, ‘more difficult’ and ‘neither’ respectively, by filtering out the other categories using the Filter tool.

Table 32 Tallies of comments about improved, lowered or ambivalently affected reading ease based on aspects of language and content. Absolute percentages are relative to the full selection of 804 texts; relative percentages are relative to the aspect of language.

Total sentences:				804
	Easier	More Difficult	Ambivalent	Total
Type	Vocabulary			
Count	134.	193.	8.	335.
Relative %	40.	57.61	2.39	100.
Absolute %	16.67	24.	1.	41.67
Type	Structure			
Count	205.	162.	30.	397.
Relative %	51.64	40.81	7.56	100.
Absolute %	25.5	20.15	3.73	49.38
Type	Coherence			
Count	128.	131.	13.	272.
Relative %	47.06	48.16	4.78	100.
Absolute %	15.92	16.29	1.62	33.83
Type	Knowledge			
Count	107.	220.	19.	346.
Relative %	30.92	63.58	5.49	100.
Absolute %	13.31	27.36	2.36	43.03
Type	Tone			
Count	55.	60.	5.	120.
Relative %	45.83	50.	4.17	100.
Absolute %	6.84	7.46	0.62	14.93
Type	Other			
Count	10.	21.	2.	33.
Relative %	30.3	63.64	6.06	100.
Absolute %	1.24	2.61	0.25	4.1

Figure 9 Stacked column chart representing comments indicating greater reading ease, lower reading ease and ambivalent effects.



While participants comment on text structure most often, they indicate knowledge (27.4% of sentences) and vocabulary (24%) as inhibiting factors more often than they do structure (20.1%). Participants most often found structure (25.5%), vocabulary (16.7%), coherence (15.9%), knowledge (13.3%) and finally tone (6.8%) to contribute to reading ease. This places the three middle categories close together, with the structure category considerably ahead of this middle of the pack. After offering a general analysis, we will explore these categories in order of frequency.

The comment annotation process revealed that assessments generally aligned with conventional linguistic wisdom. With few exceptions, participants found greater reading ease in shorter, more common words, simpler, more common structures, more coherent text construction, less demand on outside knowledge and expertise, and a less formal tone. We discuss the most typical and salient categories of comments with dual goals. The first is identifying which elements of the text will likely be strong predictors for the machine learning system (and may subsequently merit addition to that system). The second is enabling us to observe potential divergences between important factors for human readers and highly predictive features for machine learning.

Going forward, one point of nuance lies with the participants' aptitude for self-reflection. While offering participants an open, rather than restricted means of indicating how the text was more or less readable enables them to focus on those areas they deem important, other factors they are not (fully) aware of may also influence their assessment, but be less obvious to them when they rationalise their assessment. While we selected candidates to have better than average insight into the linguistic process that tie into readability and the reading experience, we are still limited to an indirect – as it is self-

reported – understanding of what influenced their assessment. That is, what causes participants to assign a score may still differ from what they *report* as causing them to assign a score.

6.3.1 Initial Exploration

In order to nuance the above caution, we will first explore how assessors' comments interrelate with the scores they assigned. We will then explore which facilitators or inhibitors to readability participants indicated, and attempt to sort them into thematic clusters². Where relevant, we will explore how we are able to quantify these inhibitors and facilitators using NLP. A more qualitative approach is not only crucial to optimising the machine learner for the genre, but also provides an important complement to that machine learner, as the best predictors of readability may be different from those that actually cause the most reading difficulty, and the latter are more important to authors seeking to optimise readability. This also compensates for one of the disadvantages of using machine learning: the techniques we will employ are a black box – that is, while they can achieve stellar accuracy, it is extremely difficult, if not impossible, to ascertain which features were the most important contributors to that accuracy.

Intuitively, we would expect assessors who mention more positive or negative features to find the text more or less readable, respectively, except potentially in those cases where they mention positive and negative features in equal proportions. In other words, the more skewed the balance of positive and negative comments, the more we might expect assessors to assign a more positive or negative score to how readable they found the excerpt. We hypothesise:

The more aspects negatively affect readability according to assessor's comment, the lower the assigned score will be, and vice versa.

As we do not measure intensity of assessments per aspect, we can only measure general intensity of positive or negative comments through the balance of aspects that receive a positive or negative readability assessment. That means we are unable to capture the nuance of cases where, for instance, the assessor indicates minor issues with multiple categories but expresses that another aspect more than redeems them in how well the author(s) implement it; they would simply register as easier based on one category and more difficult based on others. We can, however, generally expect assessors' positive or negative comments to reflect in the score (although we lack an intensity to assign as a weight), and as we expect few comments will contain such contrasts between breadth and

² As we inferred these thematic clusters from comments after the annotation process, we do not have absolute numbers for how large each cluster is.

intensity of comments, attempt to validate the basic assumption that the score should reflect comments in either direction.

For every score of the 804 we annotated, we count the number of aspects for which the assessor indicates a beneficial effect on readability, as well as the number of aspects they indicate detract from readability. Using SPSS version 23, we fit two separate regression models to predict the assigned score as a dependent variable. The first was a multiple regression model with the number of positive comments and the number of negative comments as two independent variables, and the second was a sum-based model with the number of aspects with a positive comment minus the number of aspects with a negative comment as the sole predictor. We set the alpha level to 0.05, and calculate partial Eta² as an effect size measure for significant independent variables, as well as reporting R² (adjusted) as the effect size for the model. The tables below describe the outcomes.

Table 33 Summary of significance and effect sizes for a general linear model predicting normalised scores based on number of facilitating and impeding factors; separate model.

Normalised score as dependent; number of comments indicating greater difficulty and number of comments indicating greater ease as independents. Strong significance ($p \leq 0.01$) in ***underlined bold italics***.

Model	R ²	R ² (adj.)
		0.353
Dependent	p	Part. Eta ²
Number of impediments	<i><0.001</i>	<i>0.2</i>
Number of facilitators	<i><0.001</i>	<i>0.041</i>

Table 34 Summary of significance and effect sizes for a general linear model predicting normalised scores based on number of facilitating and impeding factors; sum-based model.

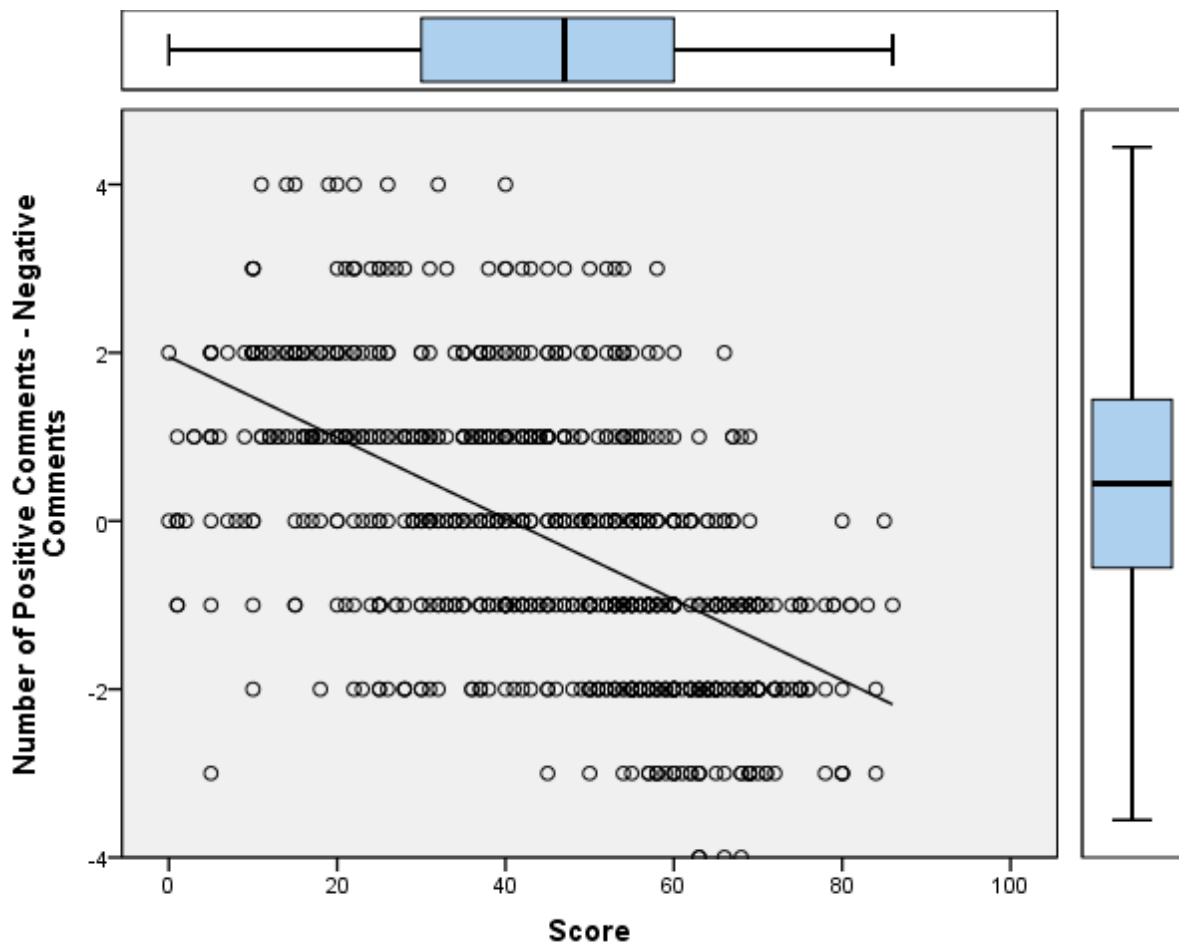
Normalised score as dependent; number of comments indicating greater ease minus number of comments indicating greater difficulty as independent. Strong significance ($p \leq 0.01$) in ***underlined bold italics***.

Model	R ²	R ² (adj.)
		0.333
Dependent	p	Part. Eta ²
Sum of facilitators and impediments	<i><0.001</i>	<i>0.333</i>

As we might expect given their similarity, results for both models are virtually identical: each of the variables proves significant ($p < 0.001$) in their models, with the models themselves showing values of approximately 0.34 for R²: 0.353 (0.351) for the multiple

regression model, and 0.333 (0.333) for the sum-based model. Effect sizes for the variables themselves do differ in the first model: the number of aspects with negative comments shows an effect size of 0.200, while the number of aspects with positive comments shows a relatively smaller effect size of 0.41. This suggests that the number of ways in which a text impedes readability is much more informative than the number of ways it facilitates it. Figure 10 shows the regression line for the sum-based (positive minus negative) model.³

Figure 10 Relationship between number of positive comments minus negative comments and assigned score.



Based on both models, we can accept our hypothesis. This outcome helps justify our initial assumption that assessors' comments as to why they assigned a certain score would be consistent with the score they assigned. Although fairly coarse-grained, these results suggest that any potential discrepancy between what assessors claim influences reading ease and what actually influences it is unlikely to invalidate the scores they assign. We do

³ As the results are virtually identical, we do not show the regression line for the multiple regression model.

notice that the breadth of negative comments seems to have a larger effect than that of positive comments. Perhaps, as we will also explore next in the per-aspect analysis, assessors have an easier time distinguishing where the text differs from a hypothetical ideal than where it matches it, which might lend those perceived differences a greater impact on their assessment than the similarities. While it might be slightly reductive to claim that assessors appeared to focus on the negatives, that claim appears to be consistent with the results. If readers are indeed more inclined to notice obstacles to readability than facilitators, that may be an important consideration to carry forward into reader-conscious writing.

6.3.1.1 Lower reading ease

The outcomes neither straightforwardly affirm nor contradict how ‘shallow’ readability formulae approach readability as a function of word and sentence length. When participants needed to rationalise their numerical assessment of a text’s readability, they most frequently noted structural aspects, followed by knowledge aspects and matters of vocabulary, in that order (see Table 32). Issues related to textual coherence and especially the tone of the text occurred substantially less often.

Consistent with the readability formulae, participants underscore the importance of word choice and sentence structure: those that indicated vocabulary as a contributing factor to low readability mentioned long, complex, unusual or unfamiliar words as detracting from reading ease. Similarly, readability formulae incorporate text structure as a variable in having sentence length detract from readability. When participants indicated a problem with structure, they frequently mentioned length, either at the sentence or paragraph level. We will first explore the vocabulary and structure aspects as core components of the conventional formulae, then examine the other aspects in order of frequency.

Vocabulary

While participants’ commentary does stress the importance of word and sentence length, the core components of the FRE, FKGL and Fog scores (see section 2.3), these are only one aspect of comments that pertain to the vocabulary and structure aspects. For instance, vocabulary issues do not necessarily imply long words, but can also encompass complex word groups – often subject-specific terminology. For instance, one participant indicated the following as an especially detrimental word group to understanding:

(54) ‘1U form factor rack boot device’

As this word group contains no words longer than two syllables, it is unlikely to affect formula-based readability from a vocabulary point of view, although using six words instead of one (e.g. simply ‘device’) will affect it from a sentence length point of view.

Nevertheless, the 'shallow' formulae are unable to detect this as difficult vocabulary. We must also remain mindful that word length and rarity do not overlap entirely. That is, it is possible for a word to be 'long', but common (e.g. 'company') or short but rare (e.g. 'lest'). The rarity of words in an excerpt appears to affect readability more than their length does (e.g. the relatively rarer 'dovetail' or 'encompass' serving as inhibitors to understanding, as one participant indicates). Other issues participants found with the vocabulary outside of length included:

- Unusual elements that impede reading 'flow', e.g.

(55) 'excessive use of product names and acronyms, especially as the subject of the sentence: renders the text impersonal and inhibits readability' or 'a lot of acronyms, complicated concepts and numerical codes/units/data'

- Unusual or inappropriate expressions, e.g.

(56) 'I don't think that "food for thought" is an appropriate term in this register'

- Vague word choice, e.g.

(57) 'this text relies heavily on abstract terms, euphemisms and "typical" business language registers' or 'too many corporate fluff words'

Structure

The structural problems that participants found were not solely limited to length issues, although these were particularly frequent (e.g. 'the bulk of sentences are compound, and therefore rather long'). Participants however indicated that shorter is not always better at a paragraph level, e.g:

(58) 'short paragraphs of varying lengths made the text a bit harder to read'

Participants also referred to multiple other aspects of structure as detracting from readability, which may correlate with sentence length, but would not directly register to conventional formulae:

- Complexity, e.g.

(59) 'explains how the system works, yet uses sentences that are more complex than needed'

- Lack of clarity, e.g.

(60) 'structure is unclear'

(61) 'in some cases, [sentences] don't seem to make much sense; or they seem to deliberately avoid being concrete'

- Monotony, e.g.

(62) 'construction of sentences is too similar throughout the text, making the reader bored'

- Inconsistency, e.g.

(63) 'the penultimate paragraph was challenging to understand, because it appeared that the sentence structure shifted'

- Use of specific structures, e.g. passive constructions or:

(64) 'enumeration of facts, which makes the text harder to read'

- Fragmentation, e.g.

(65) 'summary-style structure: a lot of sentences don't have a conjugated verb'

Knowledge

Participants' approach to readability, or in this case understandability, diverges from the formulae in their emphasis on issues of their own knowledge, and the information present in the text. After structure, this topic was the most frequent in participants' comments. While it is a very difficult aspect of readability or understandability to measure (perhaps most efficiently, if still only partially, captured by the cloze test), informational density and required knowledge proved highly meaningful to participants in determining how difficult they found a text. The conventional readability formulae, however, ignore this aspect almost entirely, only tenuously entertaining it in longer words (which may be context-specific terminology) penalising readability. Formulae (such as Dale & Chall 1948) that consider texts with a higher percentage of 'uncommon' words more difficult (for varying definitions of 'uncommon') address this more directly. We note that considering texts in terms of how difficult they are to understand for a given audience or even reader moves us closer to 'understandability' as we have defined it.

Issues of context- or genre-specific terminology, which blur the line with issues of vocabulary, commonly impede understanding, according to participants (e.g. 'field-specific terminology makes the specific meaning vague and difficult' or 'use of abstract nouns which refer to things I have no idea of'). The issue is not necessarily that the reader cannot decode a given word (for example because it is unfamiliar). Rather, the reader lacks the expertise to decode how words interrelate in the specific context (e.g. 'I was able to understand every word but the overall difficulty of the content of the text makes it more difficult to understand the overall meaning'). Even when the individual words make sense, the cognitive load is sometimes too high for the reader to process the text effectively (e.g. 'the first half is rather dense with information and the reader does not immediately grasp what it is really about'). Such comments highlight the importance of considering lexical density – as a proxy for informational density – in a more fine-grained approach to readability or understandability.

Participants also found texts more difficult to understand if they contained more references to information not present in the text (e.g. ‘a European Commission resolution, requiring further reading’). We note that some instances of such difficulty may be a result of the experimental setup: the smaller the excerpts we use are, compared to the full texts they were a part of, the more an excerpt can refer to content that the full text would have included – especially excerpts drawn from further into the document.⁴ As the example above illustrates, however, not every instance of participants struggling with outside references is a result of the subcorpus composition process. Furthermore, regardless of why these outside references are present in the text, noting that it lowers understandability is valuable for our purposes. Other complicating factors related to (background) knowledge and information include:

- Unexplained abbreviations and acronyms, e.g.

(66) ‘The text refers to ‘CO2 EOR’ several times without explaining this term.’

- Numerical information, e.g.

(67) ‘The many numbers and abbreviations in the text, make it harder to understand.’

- Reader’s distance to subject matter, e.g.

(68) ‘The reference to some chemical processes might make it a little more difficult’ or ‘due to the legal character, the text is more difficult to read’

- Excessively low informational density, e.g.

(69) ‘Little is explained’

Coherence

Relatively fewer comments addressed textual coherence and flow of information compared to structure, knowledge and information issues, or vocabulary issues, but at one mention in every three comments, coherence still played a substantial role in participants’ reasoning. We interpret coherence in the widest possible sense here and, consequently, it also includes the closely related concept of cohesion. Halliday and Hasan (2014, p. 4) define this as “relations of meaning that exist within the text, and that define it as a text.” In other words, while ‘structure’ captures comments related to organisation on a sentence level, ‘coherence’ largely captures how ideas connect between sentences.

⁴ While we attempted to minimise the number of cases in which an excerpt referred to other parts of the text, it was not always possible to eliminate these cases altogether.

How new and old information interrelate as the text builds up, and how the writer addresses different topics throughout the text's different paragraphs, are again difficult to measure using conventional 'shallow' readability formulae. Even Natural Language Processing techniques have difficulty accurately extracting the schemata that lie beneath a text's surface form (see e.g. De Clercq 2015). If we conceptualise this broad-sense quality of coherence as a text's surface form aligning with and supporting the underlying thoughts, then we would have to know – or infer – what those underlying thoughts are. Human brains remain, at least for the time being, significantly better at such tasks than computers, as we also explored when comparing the computational ease of readability formulae with the completeness of a human editing process.

In terms of how issues of textual coherence reduce readability, participants chiefly highlighted paragraph division and organisation, e.g.:

(70) 'awkwardly divided paragraphs'

(71) 'first and second paragraph could be together- fourth paragraph on its own is odd, should be linked together with third one'

In many of these cases, participants' comments seemed to imply or even outright stated that they could imagine a better way for the author to divide information into paragraphs (e.g. 'you expect a longer paragraph, but then it is cut short'). While the actual text's difference from an ideal paragraph structure is all but impossible to compute – readability formulae do not even attempt to – participants' other frequent concern is easier to quantify: a lack of linking words (e.g. 'no explicit linking between paragraphs' or 'almost no linking words are used: not within the paragraphs nor to link them'). This is more informative for machine learning purposes. As CoreNLP annotates parts of speech, we can measure the number of, for instance, relative pronouns. Participants indicate few other factors that detract from reading ease, including:

- Lack of overarching topic, e.g.

(72) 'No real structure, just (randomly chosen) facts listed'

- Lack of framing, e.g.

(73) 'No introduction, first sentence seems to be one to follow something previously said'

- Excessive repetition, e.g.

(74) 'The text constantly repeats itself, which makes it rather difficult to read'

- Shallow flow of information, e.g.

(75) 'Not a lot of detail given so text flow is hampered'

Tone

The least commented-on aspect – though still represented in just under 15% of comments – was tone, or attempts to directly address the reader in general. Again, conventional readability formulae are virtually unable to capture this aspect. We might, if somewhat tenuously, argue that a more formal tone might imply a more formal register, which in turn may correlate with more or longer words per sentence, and thus lower formula-based readability. The formal tone of the text's language is the most frequent point that participants raise against a text's readability, e.g.

(76) 'The text is (very) formal [...] the use of formal words may be an obstacle for the reader
'Formal language use and dense with information but structured properly'

Level of engagement – which would be almost impossible to measure directly with a computer – is another recurring issue, e.g.

(77) 'Very "dry" text, which it (sic) makes it harder to follow'

A few participants outright called texts boring, e.g.

(78) 'Construction of sentences is too similar throughout the text, making the reader bored'.

While a text's formality is by far the most frequent concern, participants' assessment of texts' pragmatics was fairly diverse, with key themes including:

- Lack of personal connection, e.g.

(79) 'Excessive use of product names and acronyms, especially as the subject of the sentence: renders the text impersonal'

(80) 'Referring to the company as The Company instead of 'us''

- Lack of professionalism, e.g.

(81) 'the recurrent lack of capitals reduces the professional character of the text'

- Abstractness, e.g.

(82) 'the formal register and abstract view of the situation diminish readability' or 'long text filled with generalities and promising remarks, yet little examples'

Regarding the comment that texts were 'boring', we must acknowledge that participants in this study potentially had less intrinsic motivation than those consulting a report to extract specific information. As section 2.4.3.4 explored, a less engaging text might lead to a less motivated reader and thus alter the threshold of difficulty at which a reader decides to stop reading the text (with full focus). However, as section 6.3.1.2 found that participants also commented on those cases where a text created more engagement, the

continuum between less and more engaging language should remain valid in spite of assessors potentially having fairly low intrinsic motivation to engage with these texts.

Other Impediments

Finally, the ‘other’ category encompasses those elements participants felt detracted from readability that we could not straightforwardly cluster with the preceding topics. These include:

- Macro-level layout issues, e.g.

(83) ‘A lot of elements that slow down the reading process: titles, enumerations, dates,..’

- Text/excerpt length, as opposed to sentence or paragraph length, e.g.

(84) ‘The text is very lengthy’

(85) ‘Big piece of text’

- Inconsistency in readability, which may itself lower readability, e.g.

(86) ‘Started off as an easy read but the last paragraph is more difficult’

- Orthographic errors, e.g.

(87) ‘Little defects such as using the (space forgotten)’

(88) ‘I do not know if this was a writing error in the fourth paragraph [...] but it is very distracting and looks erroneous’

- Unidiomatic language use, e.g.

(89) ‘Although the text itself is well-structured, the sentences appear unnatural. I’m even hesitating whether the author is a native speaker’

As with most of the preceding categories, none of these themes interact directly with the conventional formula-based approach to readability. For one, the most common readability formulae assume correctly formatted running text free of errors⁵. They are entirely too shallow to measure how idiomatic a piece of writing is. Overall text length, however, may tie into conventional readability: it neither consistently increases nor decreases readability, but rather lends a lower weight to any one sentence or word in the text, as there are more sentences and words over which to average. Nevertheless, none

⁵ While corporate reports, more so than many other genres, should be well formed, it is nevertheless possible for errors to make it through the editing process. Somewhat more likely in many of these cases is that Optical Character Recognition errors occurred during the conversion process to plaintext.

of these these miscellaneous comments meaningfully inform feature selection for a machine learner.

6.3.1.2 Increased reading ease

As the conventional readability formulae consider more syllables per word and more words per sentence to lower readability, it follows that they consider fewer syllables per word and fewer words per sentence to increase readability; difficulty is a continuous function of both variables. This does not necessarily hold true for human brains; the presence of a linguistic feature (for instance, proper capitalisation) might escape a reader's notice entirely, while its absence would inhibit the reading process. By a similar logic, readers might explain their perception of high reading ease using different criteria than they would reading difficulty. The elements that determine reading experience for humans are not necessarily linear or continuous; for instance, as we have found for lexical density, it is possible for extreme scenarios on either end of the continuum to negatively affect reading ease, and for other aspects of a text it may be altogether impossible to quantify them. At best, a linear function – or almost any heuristic – can only approximate human perception.

That is why this section considers reported determinants of reading ease separately from reported determinants of reading difficulty (above); we might expect considerable overlap, but not perfect alignment. This imperfect alignment, for instance, reflects in the different ratios of facilitating and impeding factors between the various categories: participants indicate structure as a facilitating factor considerably more often than as an inhibiting factor.

Structure

Participants frequently indicated that they found a text well or clearly structured, e.g.

(90) 'structure is more or less clear'

(91) 'structure is fine'

However, assessors rather infrequently explained how or why the structure agreed with them. There is a potential explanation in that many participants who signalled problems with the flow of information phrased their comments around the better ways they could imagine to structure the text at a macro level. In cases where structure contributes to readability, they may simply find that the structure the text has aligns with the structure they feel it ought to have, and be less able to define the similarities than the differences, as our initial regression analysis also suggested. Those that expanded upon their positive appraisal of the structure indicated a diverse host of reasons:

- Straightforwardness, e.g.

(92) 'Syntax is easy'

(93) 'Active voice'

- Format supports flow of information, e.g.

(94) 'The questions ensure a very clear text structure'

(95) 'Clear text, "bullet points" or rather subtitles really help to provide an overview'

- Brevity, e.g.

(96) 'The sentences are very short'

(97) 'Easy text, partly because of the short paragraphs'

- Variation in sentence length, e.g.

(98) 'The sentences are alternately short and long(er)'

(99) 'Good variation in sentence structure'

- Departure from running text where appropriate, e.g.

(100) 'Regardless of the rather difficult introduction, the remainder of the text is easy to follow with a list format adding to the readability'

(101) 'Bullet points help structuring the text (sic), increasing readability'

- Topic-focused sentence structure, e.g.

(102) 'Subject of the paragraph in bold at the beginning of the actual paragraph'

In the case of structural elements, we find that these themes do align almost perfectly with the impediments participants pointed out where such elements lowered reading ease. Participants processed active structures more easily than they did passives; they preferred varying sentence length over monotony, shorter sentences over longer ones, and topic-focused sentences over erratic structures. That last element, contrary to the previous ones, might prove more difficult to operationalise in automatic readability prediction than the others, as it entails a qualitative rather than a quantitative judgment of the text, and therefore less suited to defining readability as a mathematical function. Similarly, we might easily measure paragraph length (although delineating a paragraph is not entirely unambiguous), but participants do not indicate any standard for optimal paragraph length – the measure itself is continuous, but readers' ideal length is difficult to judge. In other words, shorter is not always better.

Participants' appreciation of bullet-point and list structures, however, is most remarkable in terms of readability prediction: as we have already addressed, conventional readability formulae assume running text, and are poorly equipped to deal with other text structures. Here, participants indicate (as the Plain English guidelines,

2013, also claim) that flouting the requirement of running text⁶ may improve readability. Recall, for instance, that the text-preparation regex for the full corpus had to strip away any lines of text that did not terminate in sentence delimiters (including the semicolon). While a bullet-pointed list with items not separated by sentence delimiters might make the text more accessible to readers, readability formulae will interpret these as excessively long run-on sentences, resulting in an inappropriately low readability score. In other words, we must be careful how we translate perceived contributors to readability technologically.

Vocabulary

As with the structure category, a strong majority of participants favoured a clear or simple vocabulary without specifying how that simplicity or clarity manifested. Others negatively defined vocabulary that contributed to readability as an absence of difficult vocabulary in general, or specifically an absence of jargon, e.g.

(103) 'Jargon absent'

(104) 'Everyday language'

(105) 'Limited use of expert vocabulary'

Sufficiently advanced NLP techniques can measure the amount of specific terminology in a text, which enables the relative presence or absence of terminology and, by proxy, jargon, as a (continuous) predictor of readability. Other types of comments were limited in number and infrequent. They included:

- 'Effective', 'efficient' or 'functional' use of vocabulary or terminology; most participants did not specify beyond this; and the most specific example was:

(106) 'The language has not been complicated where this is not functional, resulting in a straightforward and easy-to-predict chunk'

- Sufficiently concrete and precise vocabulary, e.g.

(107) 'Superfluous fluff words don't impede understanding'

(108) 'The text is candid and reads easily because of [...] precise wordings'

- Sufficiently low lexical density, e.g.

(109) 'Lexical items are not dense'

⁶ It is possible for NLP techniques to process text not formed like running text, but doing so requires either normalisation (for which the approach can differ for every type of source material) or specific training on the non-well-formed source material (Schulz et al. 2014, Van Hee et al. 2017).

Judging whether the author uses vocabulary effectively, efficiently or functionally is, again, more the purview of an editor than of a formula, as it is a qualitative rather than quantitative judgement. However, we might argue that lexical density – which participants also highlighted – can be an appropriate, if imperfect proxy for efficient language use. Lexical density, however, only indirectly – if at all – measures concreteness, in that a lexically sparse text may offer too little content to be concrete, and an excessively dense text has fewer function words to explain how its contents interrelate. In summary, although such qualities as ‘functional’ or ‘precise’ language use may be difficult to capture, including measures of terminological and lexical density can help the system judge texts as human readers would. Lexical density is more difficult to operationalise than many other measures we will use, however, because it does not appear to offer a straight, continuous, linear relationship with reading ease: a text must be neither excessively lexically dense, nor excessively sparse. In this respect, the comments affirm ideas explored in section 2.4.2.

Coherence

How the text achieved coherence (in the widest sense of the word) was the next most frequent contributor to reading ease and, in contrast with the categories of vocabulary and structure, participants frequently made clear how the text achieved coherence in their eyes, although such evaluations as ‘paragraphs make sense’ or ‘clear [...] use of paragraphs’ do occur. Overall, we continue to see that participants evaluated texts more favourably where the text’s organisation aligned with how they believed it should be organised. Explanatory comments clustered around the following themes:

- Alignment of structure with flow of information, e.g.

(110) ‘First sentence introduces subject and main focus of text’

(111) ‘The text is well structured on a macro level’

(112) ‘Paragraphs are linked via content’

- Deliberate use of paragraphs, e.g.

(113) ‘The last three paragraphs are more hands-on [...] they clearly center around one idea’

(114) ‘Paragraphs are spaced out logically’

- Explicit cohesive markers, e.g.

(115) ‘Many linking words’

(116) ‘This text is well-structured with for example the use of ‘however’”

Again, the main themes align with the impeding factors that participants suggested where they found lower reading ease: they register, for instance, both the use and lack of framing techniques such as introducing a topic at the start of a paragraph. They signal a

clear logic in paragraph divisions just as well as an illogical one, and indicate absence or presence of cohesive markers. Again, then, the issue becomes that it is difficult to quantitatively express how well a text's structure aligns with the informational schemata underlying it, because the latter are almost impossible to infer computationally compared to how well the human brain does so.

We were previously unable to discern an optimal paragraph length; shorter is not always better, for instance, as readers indicate that they prefer paragraph structure to support flow of information, and different topics require different amounts of attention. Determining how paragraphs (should) interrelate is no less difficult, although not entirely impossible, thanks to the still-evolving (at time of writing) NLP technique of coreference resolution (see e.g. Clark & Manning 2016), which focuses on how sentences within a text linguistically refer to (elements of) one another. As De Clercq & Hoste's (2016) generic system already integrated coreference features, we are able to carry it into training the system, although we must note that given the complexity of performing accurate coreference resolution, implementing coreference features may only have modest gains compared to their computational requirements (De Clercq 2015).

Another proxy for cohesion might be the number of linking words, which we can simply count in order to have a quantifiable, if again imperfect, measure of how cohesive a text is – or at least attempts to be. We could count conjunctions, different types of pronouns and adverbs, etc. (with single-word linking words easier to measure than multi-word lexemes), to better understand how much attention the text devotes to connecting the information it expresses.

If we had no access to automatically resolved coreference, we might further refine such a measure by comparing the number of linking words to the text's lexical density, i.e. the percentage of content words compared to function words. This might inform us which percentage of the text that expresses relations expresses explicit logical relations. The case for linking words also helps explain why lexical density is not on a continuous scale of more (or less) readable: a text with relatively more dense content has fewer words to achieve cohesion or signal logic.

Knowledge

That the corporate reporting genre informed most participants' expectations concerning background knowledge and expertise is hardly surprising; it is the reason we added the 'knowledge' aspect to account for genre-specific readability issues. We also see a straightforward continuum between the primary detractors from and contributors to readability within this aspect. Participants find texts with less specific terminology and fewer demands on subject-specific expertise easier to read, just as they found texts with more such demands harder, e.g.

(117) 'Use of clear, every-day language'

(118) 'Even outsiders or non-experts are able to read this text'

As with the coherence aspect, themes clustered together fairly tightly, with few thematic outliers:

- Ease of assimilation, e.g.

(119) 'Feel like I remember most [of the text] after reading once'

(120) 'The gist of the text remains after reading'

- Necessary knowledge available in text, e.g.

(121) 'All information required is present'

(122) 'Little reference to outside information'

- Use of examples and explanations, e.g.

(123) 'The writer explains every piece of information'

(124) 'This text takes the time to illustrate abstract principles and concepts (with elephants, if need be)'

As we have already addressed under the heading of the overlapping category of vocabulary, estimating the number of genre- or topic-specific terms that a text contains using NLP techniques is a feasible task. It might be more difficult to determine, without an editor, whether a term would be problematic because of its obscure or technical meaning, or because the word form itself is more difficult to process (consider, for instance, how 'get' can mean the same thing as 'acquire' but 'acquire' can still be slower to process). How useful such a distinction would be remains to be seen, as we can reasonably assume that longer words and more terminology means a more difficult read, regardless of how they interact. We might expect that for shorter domain-specific words the experiential distance between word and reader would be the larger obstacle, while for longer terms it could be that distance, the word itself, or both.

Whether a text contains all the knowledge the reader needs in order to fully understand it depends at least in part on the reader; authors may need to keep the lowest plausible level of expertise amongst their prospective audience in mind while they craft their text. The aforementioned coreference measure may be a viable, if indirect proxy for the presence of prerequisite knowledge: the longer a text's coreference chains are relative to its total length, the more early text may scaffold later text, and later text may draw on earlier text. It is an indirect proxy at best because there are other potential indicators that a text gradually builds up the required information, and other ways to achieve that build-up. We may, however, use number of hyperlinks as an indirect proxy of outside references in web-based content.

Ease of assimilation, finally, provides something of a chicken-and-egg problem: did participants find the text more readable because the information within proved easy to assimilate, or did they find it easier to assimilate because it was more readable? This ties

into the contrast between readability and understandability, and we must note in this respect that as respondents were language students, they may have been somewhat more focused on form rather than content than the average reader might be.

Tone

The issue of tone ties into that same contrast between readability and understandability: how the text engages with the reader is seldom a purely text-internal issue, and preferences are likely to differ more between readers than they do for vocabulary or structure. The issue of formal language provides an excellent example: we might have expected participants to find informal language easier to process, just as they generally found more formal language more difficult to process. Most do, e.g.

(125) 'Straightforward and pretty informal language'

(126) 'Lots of dry information but told in an informal/personal way'

However, some participants indicated a preference for how the text deployed formal language, e.g.

(127) 'A formal register effectively used'

(128) 'Formal language is not obtrusive, but functional'

(129) 'Switches between formal and informal registers to keep a pleasant cadence'

In other words, we typically see a continuum between informal and formal language in how they influence readability, but at the same time are aware that exceptions can occur. In as much as NLP might be able to detect how formal a text's writing is – and while it is possible to detect specific formal and informal turns of phrase, quantifying overall tone is difficult - translating a numerical expression of the text's tone into a measure of understandability could be challenging. If nothing else, this study's participants inform the caveat that more formal language is not always less readable.

Participants also pointed out a few other contributing factors to readability or, arguably understandability (see section 2.4.1), within the category of tone, although, like some of the previous categories, comments clustered together fairly tightly:

- Textual flow, e.g.

(130) 'Nice flow in text'

(131) 'Its cadence [...] makes it easier to process'

(132) 'The text is very easy to read because it is fluently [written]'

- Narrative, e.g.

(133) 'Facts somewhat embedded in "story"'

(134) 'A lot of narrative elements'

(135) 'Geared towards the reader because it tells a "story" rather than state facts'

- Engaging language, e.g.

(136) 'Well told'

(137) 'Very "vivid", lively and of an optimistic nature'

- Personal identification on behalf of speaker, e.g.

(138) '[Being] written from a first person perspective tends to make texts easier',

(139) 'Use of person's name, "we", "our (stakeholders)" as a subject improves readability'

(140) 'We-perspective makes it easier (to me, at least)'

This final example again illustrates how experience of tone is more reader-dependent and less text-internal than, for instance, text structure or vocabulary use; the participants themselves are aware of how subjective these judgements can be. Somewhat ironically, however, this comment applies to the most text-internal, quantifiable and reader-agnostic theme out of various tone-related contributors to readability. Compared to some of the other contributors to and detractors from readability, we can very easily count the number of (first-person) pronouns a text uses.

The other themes, again, are more subjective to the reader's experience than they are intrinsic to the text. While authors may craft a text to engage, tell a story or seem fluent to the largest possible part of their audience, these are qualitative aspects of a text, and belong firmly to the realm of the editor rather than the formula. Participants' observations here inform decisions authors may wish to make to achieve readable writing, but are less suited to optimising readability formulae.

Other facilitators

Virtually all of the participants' comments fit into the preceding categories; we placed very few positive remarks on readability in the 'other' category. Those remarks primarily included very reader-specific comments, such as their finding the text particularly interesting or having encountered it earlier in the experiment. However, we also find total text length as a variable relevant to participants, forming a continuum with participants' previous negative remarks on long texts, e.g.

(141) 'Very short text, easy to concentrate and easy to read'

(142) 'Short and not too difficult to understand'

We have already explored how text length interacts with formula-based readability; we note that some participants notice a text's length both positively and negatively, and that all cases that explicitly mention text length in the sample associate longer texts with lower readability, and vice versa. We did, however, encounter cases in the previous categories where practices that contributed to readability, such as detailed explanations of expert concepts, would invariably make the texts longer than if authors did not apply

them. Future studies may wish to examine whether readers generally prefer shorter texts, or whether these comments are more indicative of what a reader believes an author ought to do or have done with the amount of text they present, i.e. whether the issue is truly length or, as previous categories indicate, conciseness.

6.3.2 Implications

The above analysis has two major implications: first, the number of qualitative aspects of a text relevant to participants' reading experience makes clear that mainstream NLP techniques cannot replace an editor with the ability to make subjective judgments, empathise with the audience, and deal with difficult-to-quantify data (see section 2.6.1). That is not to say we cannot approximate such an editor in some respects – and that is precisely what the rest of this section will attempt – but given the vast array of different variables, the editor's position as the 'gold standard' of determining how readable or understandable a text is, for now, remains secure.

Conversely, this study also reinforces how there are significantly more variables that contribute to reading ease (or a lack thereof) than simply average word and sentence length, many of which are still quantifiable. While participants' comments also indicate that word and sentence length remain relevant, the additional variables become especially important when dealing with domain-specific text and how it contrasts with general text. Participants indicated several additions we can make to a metric of genre-specific readability.

- Use of terminology. One of the participants' most frequent remarks was on the relative presence and absence of specialist terms and concepts. Readability metrics can estimate this using term frequency (see section 2.4.2) or similar measures.
- Use of rare words. Participants associated uncommon words and unusual registers with lower readability, regardless of word length. How often words used throughout the text occur in a general corpus may be informative, even for non-terms (i.e. generally uncommon words that do not see frequent use throughout the text in question – for more on termhood, see section 2.4.2.1). Texts that use less common words, even without such words being subject-specific terms, may be less readable.
- Vagueness. Participants associated (excessive) vagueness with lower readability. While how vague a text is might seem like a highly subjective judgment, Alexopoulos & Pavlopoulos (2014) demonstrate that it is possible to automatically classify words as vague or non-vague with high accuracy using machine learning. However, we find a more feasibly implementable approach in Brysbaert, Warriner & Kuperman (2014), who provide a 40000-

word lexicon with concreteness ratings. We expect that texts with more words that are vague may be less understandable.

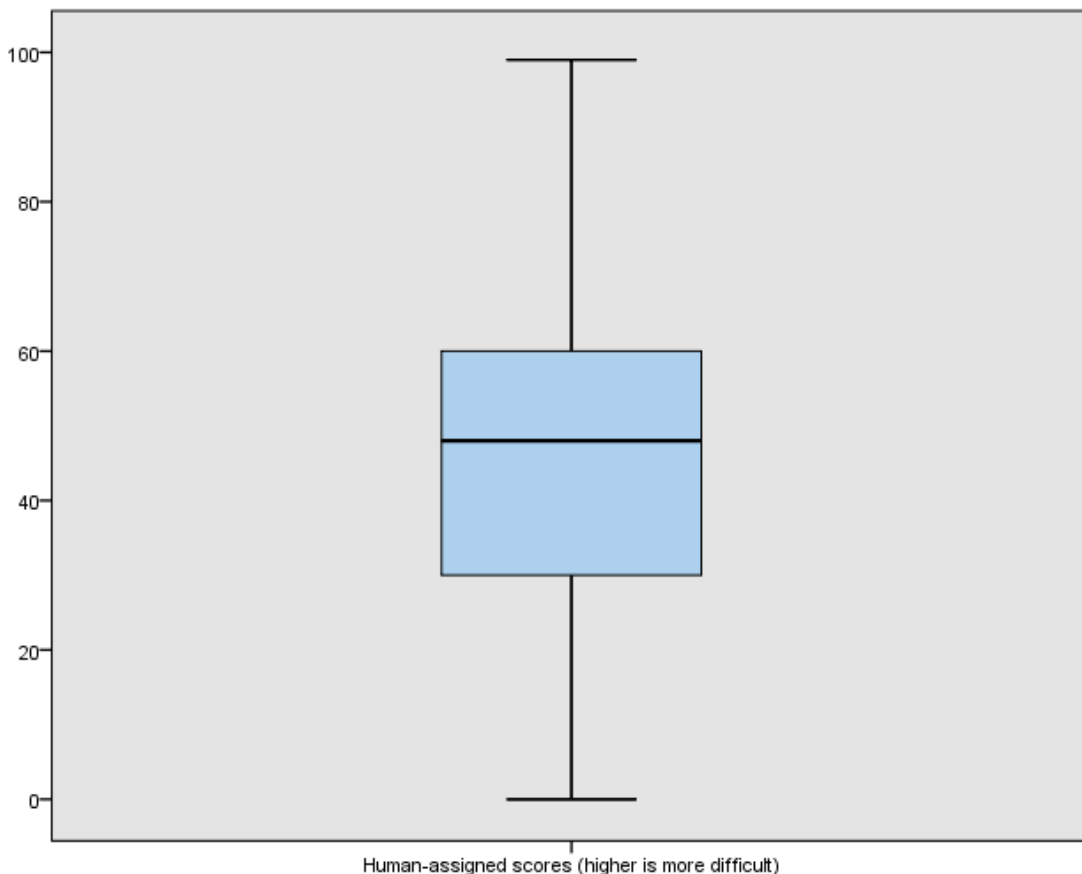
- Informational density. Participants disfavoured texts with extremely dense or sparse content. Lexical density can serve as a proxy, although we do note that either extreme is undesirable; thus, a straight line is unlikely to accurately describe the relationship between lexical density and readability.
- Use of proper nouns, acronyms, abbreviations and numbers. Participants indicated that these, in particular, impeded reading ‘flow’. Counting these elements is a straightforward task.
- Sentence complexity. Participants disfavoured compound sentences, especially those with highly irregular patterns. We can capture sentence complexity through the ‘number of subordinators’ and ‘average parse tree depth’ variables; adding the number of coordinating conjunctions as a variable may offer additional information.
- Use of passive structures. Participants find that the passive voice lowers readability, and the active voice raises it. Our initial exploration already included relative use of passives as a variable (see section 2.4.2).
- References to other parts of the text. Participants preferred it when other information the text refers to was text-internal. When this is the case, we would expect to see more coreference chains. Coreference resolution (see section 2.4.2), however, is a computationally expensive task that may yet not be viable at the required level of quality for texts averaging the full length of a corporate report, as its computational demands scale exponentially with length, and accuracy, while steadily evolving, remains relatively low compared to that of other NLP tasks (Clark & Manning 2016).
- References to other texts. Participants disfavoured external references. In the case of web-based content, the number of (external) hyperlinks may serve as a proxy for the number of external references.
- Use of cohesion markers. Participants preferred texts that explicitly signalled connections between ideas. We can measure the relative presence of various connectives, such as relative pronouns, and expect more such linking words to increase readability.
- Use of personal pronouns. Participants favoured greater use of personal pronouns, as they found them to convey a more personal tone. We can count the uses of personal pronouns, be it any uses or first-person uses specifically.
- Exemplification. As participants responded favourably when authors offered examples, a tally of exemplification markers such as ‘for example’,

‘for instance’, ‘i.e.’, ‘e.g.’, ‘such as’, etc. may help more accurately predict reading ease. While this is an open list in that authors can use an arbitrarily large number of ways to express exemplification, the vast majority of such uses should fit within a limited list, and thus approximate the actual use of exemplification.

6.4 Exploration

To counterbalance these qualitative insights into the data with a more quantitative understanding, we computed the mean and standard deviation for the assigned scores, in addition to creating a boxplot to summarise the data. Contrary to other analyses, we calculated these on the full set of scores (i.e. all 3987). These outcomes are before normalising scoring ranges.

Figure 11 Boxplot of human-assigned scores before normalisation



Participants assigned a mean score of 45.38 (48 median) with a standard deviation of 19.414. As the minimum was 0 and maximum was 99, the initial impression is that they assigned a fairly wide range of scores. As the mean is close to the middle of the range, this suggests the validity of the genre adaptation experiment, as these texts scored towards the bottom end of the readability formulae.

We also wanted to gain further insight into participants' responses and, potentially, the chief causes behind their perceptions of difficulty in addition to which aspects of the text tie into machine learning-based readability. Accordingly, we first investigated to what extent the type of comment participants made could predict the score they ultimately assigned (again only drawing on those data points we annotated for comment type). In order to approximate this before beginning to implement machine learning, yet provide more quantitative insight into the data than the preceding sections, we fit a univariate linear model in SPSS version 23 that used the type of comment for each of the categories (easier, more difficult, other, or none) to predict the score the assessor assigned. We controlled for the participants' own scoring bias by including their user ID as a random variable. Table 35 summarises the outcomes.

Table 35 Summary of significances and effect sizes for type of comment on a given category in a general linear model predicting assigned score. Comment types operationalised as categorical variables.

Association with assigned score (n = 804)		
Variable	p	Part. Eta ²
Vocabulary	< 0.001	0.035
Structure	< 0.001	0.132
Coherence	< 0.001	0.096
Knowledge	< 0.001	0.144
Tone	< 0.001	0.068
Other	0.193	0.006

During a post-hoc analysis, after applying Bonferroni correction, we found that for each of the significant predictor categories except 'knowledge' (i.e. also omitting the 'other' category), the presence of a comment indicating difficulty systematically indicated higher reading difficulty than the absence of a comment or a comment indicating greater ease. The opposite was proved to be the case for comments indicating greater reading ease for a given category. These outcomes were all significant at $p < 0.001$. We were unable to perform a post-hoc analysis for the 'knowledge' category as its ambivalent 'other' tag only had a single case, but see no reason to expect a meaningfully different outcome.

The following table indicates the mean differences between the three options of blank, easier and more difficult; given its nature, outcomes for the 'other' difficulty category were too erratic to merit comment.

Table 36 Mean differences between comment types per category. Values are mean for second category subtracted from mean for first.

Mean Differences				
	Vocabulary	Structure	Coherence	Tone
Difficult - (Blank)	12.06	14.3	11.89	12.41
Difficult - Easier	19.12	20.78	19.79	28.47
Easier - (Blank)	-7.05	-6.48	-7.89	-16.06

The observation that every category of comment shows a highly significant association with the assigned score reflects positively on both the experimental setup and division into categories of comment. Although every predictor except the ‘other’ category was strongly significant, we do see a clear hierarchy between the effect sizes of the different categories; the predictive value of comments related to knowledge (partial η^2 of 0.144) and structure (partial η^2 of 0.0.132) exceeds the typical threshold for a medium effect size, while that for comments related to vocabulary fairly narrowly exceeds the threshold for a small one (partial η^2 of 0.035). Coherence and tone occupy the space in between.

Accordingly, we see that the mean difference between mention of difficult structure and a lack of any comment is the largest of the various groups (14.3). The largest difference overall (28.47), however, is between texts that received a positive comment on tone in terms of reading ease compared to a negative one. Given the ‘tone’ category’s relative lack of frequency in the comments, this suggests that the presence of interpersonal engagement might benefit perceived reading ease more than its effect size alone would suggest. Nevertheless, the ‘knowledge’ and ‘structure’ categories’ relatively high effect sizes (respectively .144 and .132) align with participants’ comments. They likely found it easier to infer the meaning of a specific lexical item than to accommodate for difficulties processing a difficult text’s structure or to compensate for prerequisite knowledge its author (wrongfully) assumed they had.

6.5 Processing

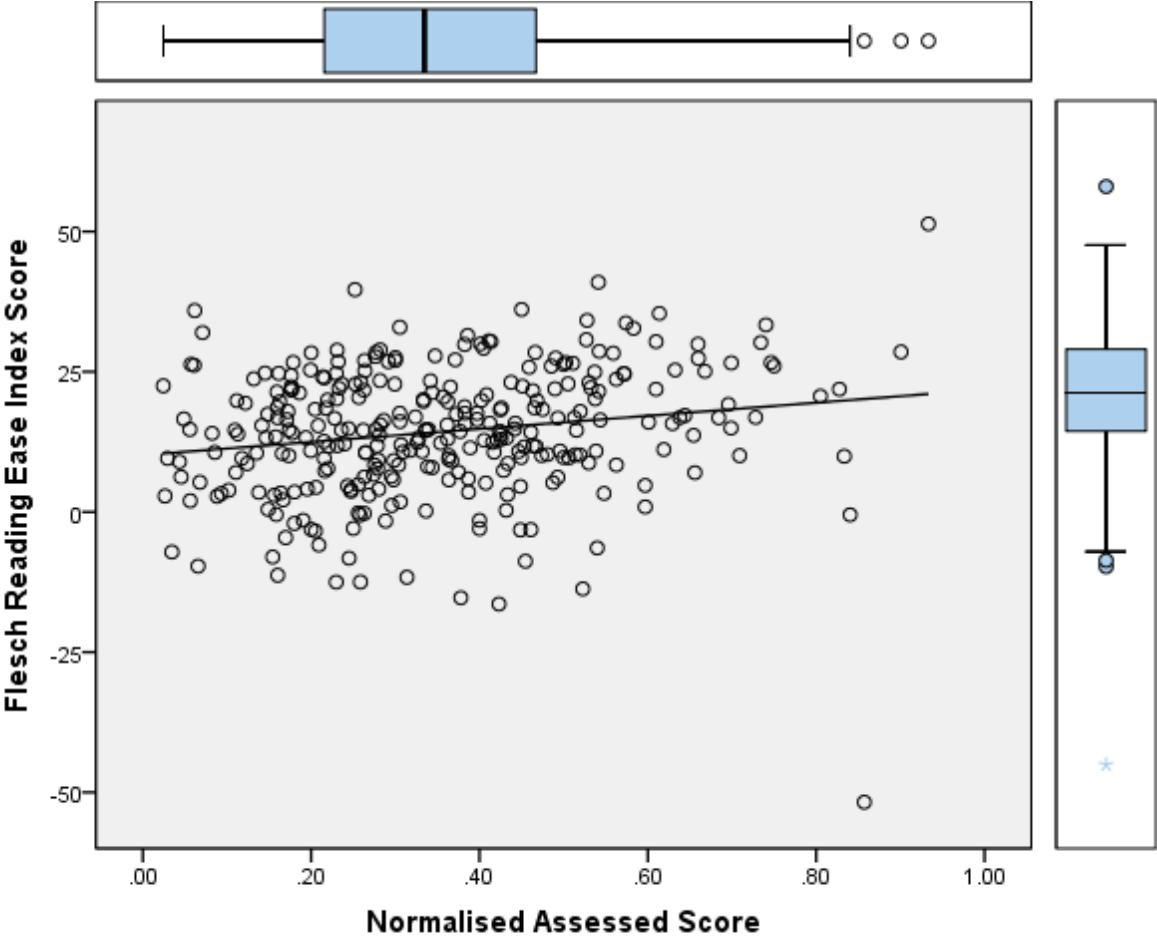
Machine learning will aim to produce the lowest possible error relative to a gold standard (the human-assigned scores) by manipulating the weights of the various features we can make available to it (see section 2.6.3 for a high-level overview of how machine learning works). We needed to take a number of steps before it became possible to apply these techniques to readability prediction. We first needed to concatenate the various scores and process the excerpts in order to provide the machine learner access to the features it needs to approximate human assessment. In order to achieve the latter, we used CoreNLP to annotate the excerpts and De Clercq & Hoste’s (2016) Python-based feature extraction

framework exactly as we did for the full corpus (see Chapter 3). The method used to concatenate the various scores also matched De Clercq & Hoste's: the first step was normalising assigned scores over the full 0-100 range to which they had access. Few annotators used (almost) the entire range of 0-100, nor did they receive instructions to do so. Consequently, averaging their scores was impossible without normalising towards a shared range. To do so, we calculated a weighted, normalised score for every text, slightly favouring scores assigned in larger batches of text. This followed the approach set out in De Clercq et al. (2014). After normalisation, we averaged scores in order to have a single score for every excerpt.

Comparison to Flesch Score

To gain additional insight into the relationship between these 'gold standard' scores assigned by human readers and the traditional readability formulae, we calculated the correlations between the excerpts normalised score and its Flesch score and, in spite of a significance of 0.001, found an altogether modest Pearson correlation coefficient of 0.185. Similarly, attempting to build a linear model with the Flesch score as the dependent and normalised score as the independent led to the same high significance ($p = 0.001$) but very small percentage of variance explained (at an adjusted R^2 of 0.031). The figure below contains the scatterplot and regression line for that analysis.

Figure 12 Scatterplot and regression line for normalised averaged score and Flesch score.



This outcome is likely the strongest piece of evidence in this study that readability formulae are a useful yardstick of how readable a text is, but that they can at best approximate the complex set of factors that make a text readable, let alone understandable. That is not to say they have no value or use in a study such as this – on the contrary, they can offer an immensely valuable, computationally efficient at-a-glance impression of a text’s readability. Instead, this outcome illustrates the absurdity of ‘writing to the formulae’ rather than using one’s own (immeasurably more complex) grasp of the language, as an author, to ensure a readable text.

6.6 Machine Learning⁷

6.6.1 Measures Used

As a genre-specific re-implementation of De Clercq & Hoste's (2016) general corpus readability prediction approach, this study's machine learning implementation was able to draw on many features (i.e. quantifications) that participants had flagged as important in their comments. However, the need to re-train the system also presented the opportunity to add a number of features based on (quantifiable) aspects specific to the genre that participants had identified, and to test the value those features would add to a re-trained version of the original system. This implementation of the learner sought to minimise the root mean squared error for the regression function based on the following features, many of which we previously described as potential predictors for readability (see section 2.4.2). We compare this implementation with previous approaches, which include both implementations trained on the generic corpus and on this subcorpus.

The machine learner used the following set of features also present in De Clercq & Hoste (2016):

- Features used in the full-corpus analysis (Chapter 3):
 - Average word length
 - Average sentence length
 - Ratio of long words (three or more syllables)⁸
 - Percentage of polysyllable words
 - Percentage of words in Dale and Chall (1995) list
- Type-token ratio (i.e. ratio of unique words to all words)
- Lexical features (calculated based on written part of BNC corpus (Aston & Burnard 1998))
 - Perplexity score (how well does the language model predict the text?), calculated using SRILM toolkit (Stolcke 2002) based on BNC (see section 6.6.2)
 - Normalised perplexity score based on document length (also using SRILM)
 - Term Frequency – Inverse Document Frequency (TF-IDF; see section 2.4.2) calculated using BNC as background corpus

⁷ This section especially merits an acknowledgement of Orphée De Clercq's efforts in implementing the new genre-specific features and training the machine learner on the new dataset.

⁸The syllabification process relied on a classification-based syllabifier described in van Oosten, Tanghe & Hoste (2010).

- Mean log-Likelihood of words in the text (Rayson & Garside 2000) calculated using BNC as background corpus; this captures the relative frequency of words relative to the background corpus
 - Syntactic features
 - Based on Part of Speech tags:
 - In text:
 - Absolute frequency
 - Relative frequency
 - Per sentence
 - Absolute frequency
 - Relative frequency
 - Average type, i.e. average number of unique words per sentence
 - For:
 - Nouns
 - Adjectives
 - Verbs
 - Adverbs
 - Prepositions
 - Average number of function words
 - Average number of content words⁹
 - Deep syntactic features
 - Parse tree depth (Schwarm and Oostendorf 2005)
 - Number of subordinators (Schwarm and Oostendorf 2005)
 - Ratio of noun, verb and prepositional phrases (Schwarm and Oostendorf 2005)
 - Number of passives (De Clercq & Hoste 2016)
 - Semantic information
 - Average number of connectives (sentence and document level)
 - Causal
 - Temporal
 - Additive
 - Contrastive
 - Concessive
 - Named entity (e.g. recognising the company's name as referring to a distinct entity)

⁹ These two variables combine into lexical density as section 2.4.2 described it.

- Predicted entities (recognised by CoreNLP’s Named Entity Recognition system)
 - Number of entities (sentence and document level)
 - Number of unique entities
- Shallow entities (based on part of speech)
 - Number of entities (sentence and document level)
 - Number of unique entities
- Coreferential information (see section 2.4.2)
 - Number of coreferential chains
 - Average length of chains
 - Average number of coreferring expressions and unique mentions
 - Number of chains spanning across more than half of text

The following features joined the aforementioned ones in an attempt to better accommodate the idiosyncrasies of the genre (in as much as they were quantifiable):

- Use of rare words based on SUBTLEX-US lexicon (Brysbaert & New 2009)
- The same four lexical features as mentioned above, but modelled on different corpora in order to reduce potential language variety bias (see sections 3.2.2 and 6.6.2):
 - Written part of Corpus of Contemporary American English (COCA; Davies 2008-)
 - Combination of (written only) British National Corpus and COCA corpora
 - Non-blog components of Corpus of Global Web-Based English (GloWbE; Davies 2013-) from British, Irish, Australian and Indian sources.
- Concreteness score (as respondents experienced vagueness as an obstacle) based on Brysbaer, Warriner & Kuperman’s (2014) per-lemma concreteness scores.
- Presence of numerical information (based on part-of-speech tags)
- Presence of acronyms and/or abbreviations as detected by regular expressions
- Use of personal pronouns based on part-of-speech tags
 - First-person
 - Any
- Exemplification based on common phrases detected by regular expression (we captured ‘for instance’, ‘for example’, ‘i.e.’, ‘e.g.’, ‘such as’, ‘illustrated’, ‘illustration’, and ‘in particular’).

6.6.2 Language Modelling

Of specific note as an addition to the existing system was our expansion of language modelling that aimed to account for the different varieties of English present in the corpus. Based on respondents' input, we sought to implement a system of determining how rare (or common) the words in a report are, which would be more accurate than simply measuring word length. While, on average, longer words may be more uncommon (this is why readability formulae calculate word length as a difficulty heuristic), it would be fallacious to assume that any given word is less common than any words shorter than it is. Consequently, we might gain considerable information from determining word rarity based on the word itself, rather than the number of characters required to write it.

In order to shore up the weaknesses of word length-based difficulty prediction, we synthesised several language models that indicate how words and sequences of words occur in natural language. We did so with the intent of quantifying a text's adherence to or divergence from one or more language models as a means of predicting its readability. While language models based on the British National Corpus were present in the original system, we perceived that this study's multi-varietal corpus might benefit from expanding beyond that single variety. As was the case in De Clercq & Hoste (2016), SRILM, the SRI Language Modeling Toolkit, compiled these models (see Stolcke 2002 for technical details).

To create useful language models, we needed reference corpora large enough to distil accurate patterns of word use. We drew on (combinations of) three corpora:

- The British National Corpus (BNC), which represents approximately 90 million words of written British English (in addition to approximately 10 million words of spoken British English) from the 1980s up to 1993.
- The Corpus of Contemporary American English (COCA), which contains approximately 420 million words of written American English (and approximately 100 million words of spoken language) dating between 1990 and 2015 (Davies 2008-).
- The Corpus of Global Web-Based English (GloWbE), which contains 1.9 billion English-language words from 20 different varieties of English. Of particular interest to this study are the American, British, Irish, Indian and Australian general (as opposed to blog) sites, which add up to approximately 760 million words (Davies 2013-).

SRILM requires plaintext input of one sentence per line. Although we had a compatible version of BNC available from previous projects (De Clercq & Hoste 2016), preparing the latter two corpora took some additional processing. This was due to the format in which both corpora were available: running text with one source per line and <p> tags indicating paragraph breaks, with 5% of the corpus replaced by @ symbol sequences in order to

ensure fair use compliance. In order to transform these corpora into a one-sentence-per-line format, we applied a sequence of regular expressions (with '=' between tabs denoting a replacement):

1. `##[0-9]+ ? =>`
Delete every number preceded by double # symbols optionally followed by a space. As these numbers denote source documents ID for the corpus and are not linguistically useful information, we discard them.
2. `<(p|h)> =>`
Delete <p> and occasional <h> tags denoting markup, as they are not linguistically useful information.
3. `([.?!] ["'"] ,]? ?) => \1\n`
Insert a line break after every sentence boundary punctuation character (period, question mark or exclamation mark) followed by a space, optionally followed by closing quotation marks, optionally followed by a space. As tokens in the corpora are already separated by space, every sentence boundary character followed by a space does delineate a sentence boundary; such characters as part of a token (such as 'No.' in 'No. 1') would not have spaces on both sides.
4. `^.*?(@){2,}.*?[.?!]? ?["'"] ,]? ?$ =>`
Delete every line that contains a sequence of two or more @ symbols separated by spaces and ends in sentence boundary punctuation optionally followed by closing quotation marks. As the previous step separates sentences with line breaks, this deletes every sentence modified for fair use purposes, as they are unsuitable for building language models.
5. `^ =>`
Remove leading spaces at the start of every line.
6. `^\n =>`
Delete empty lines.

6.6.3 Prediction

While de Clercq & Hoste's (2016) readability prediction system partially drew on Genetic Algorithm (GA)-based machine learning using the Gallop toolkit (Desmet & Hoste 2013), the present implementation, as a genre adaptation task rather than an attempt to ensure technical advances, employs support vector machine (SVM)-based machine learning. It performed vector regressions using LibSVM (Chang & Lin 2011) and tests using 10-fold cross-validation, attempting to optimise (i.e. minimise) the root mean squared error (RMSE) between the predicted score and the gold standard (i.e. human-assigned) score across the subcorpus. The machine learner calculated RMSE using the following formula (from De Clercq & Hoste 2016, p. 468):

$$\text{RMSE} = \sqrt{\frac{1}{m} \sum_{i=1}^m (X_i - x_i)^2}$$

in which X_i is the prediction and x_i is the response value, that is, the correct value, for the regression task at hand, and m is the number of texts for which a prediction is made. The lower the RMSE value, the better.

The system saw five implementations: one as trained on De Clercq & Hoste’s (2016) generic corpus (described in section 6.2.3), one trained on this subcorpus without the addition of genre-specific features, and one with the addition of those features. The latter two also saw an implementation after grid search-based hyperparameter optimisation (which attempts to optimise the learner’s parameters by exhaustive search through all options to minimise the loss function)¹⁰, for a total of five. Table 37 describes the outcomes:

Table 37 RMSE scores for different machine learning training scenarios. Best performance in **bold**.

Training Set	Includes genre-specific features?	Grid Search Optimisation	Root Mean Squared Error (lower is more accurate)
Generic	No	No	0.1855
Sustainability Reporting	No	No	0.0813
Sustainability Reporting	Yes	No	0.0879
Sustainability Reporting	No	Yes	0.0038
Sustainability Reporting	Yes	Yes	0.0051

As the above table will illustrate, the best-performing scenario is the one in which an optimised system trains on the sustainability report excerpt subcorpus and, strikingly, does not include the additional genre-specific features that participants indicated as relevant, such as the number of acronyms or exemplification structures. While this was not the most intuitive outcome given commenters’ indications that these aspects detracted from readability, it is not an implausible one, either.

Two factors likely contributed to this outcome: on the one hand, given the optimised scenarios’ high performance overall, it is likely the generic model already contained all

¹⁰ For more theoretical insight into grid search, see e.g. Bergstra & Bengio (2012).

the necessary features – i.e. predictors – for high accuracy. In that case, while the additional genre-specific features may be informative, the predictive power of those additional features was redundant to the point of mostly adding noise instead of predictive power. On the other hand, as we have previously explored, the actual motivations behind assessors assigning a score might diverge from those quantifiable aspects of the text that best predict the score they assigned from a regression standpoint. This is further compounded by the previously mentioned possible divergence between what annotators *perceive* as influencing their decision on which score they assign, and what might influence that score subconsciously. To further explore these factors, we calculated correlations between the normalised, averaged scores that the learner had to predict and the features we explored during the full-corpus analysis (Chapter 3) as well as the new features we added on to the learner based on participants' comments.

Before we proceed, however, we must stress the most important two implications of these RMSE results for the learner: first, that it is very possible to train a learning system that can accurately approximate human assessment on a full 0-100 band of perceived reading difficulty. Second, that genre adaptation of existing readability prediction systems is a difficult, but viable and necessary task for optimum accuracy – one that does not appear to demand additional genre-specific features. Taking into account results for other readability measures, which systematically place corporate reporting amongst the most difficulty categories, this implies that the answer to the questions whether we can and should use genre-specific readability metrics for the most accurate possible insight into reports' accessibility is a resounding 'yes'. As out of all the different component studies this one draws closest to understandability compared to readability, these outcomes only reinforce the notion that this avenue of research can help both authors and audiences improve how they deal with the genre. The greatest caveat here is the 'black box' nature of SVM-based machine learning: while the reader can achieve a high accuracy, it is (practically) impossible to reverse engineer which features best contributed to that accuracy. Accordingly, we need to approximate and understanding of which features were most important through less technically advanced means.

6.6.4 Correlations & Discussion

This section examines how the gold standard score and features we have explored in previous sections of the study – or added based on assessors' comments in this one – correlate with the normalised scores and the readability measures we examined in Chapter 3. As the machine learning process itself is a 'black box', this is one of the better means of providing further insight into the determinants of (perceived) report readability. Knowing what enhances or detracts from (perceived) readability is crucial when dealing with a wider audience of stakeholders in the company's operations rather

than an audience with the greater expertise that tends to accompany shareholdership. To do so, we present two tables. The first describes correlations and significances of those correlations between the normalised average score for an excerpt and the number of components of readability formulae and features used in deeper-level lexicosyntactic analysis we discussed in Chapter 2 and Chapter 3, as well as those that quantified aspects that assessors for these excerpts flagged as important contributors or detractors from readability. The second describes the correlations between the first two categories and the latter, as well as the significances of those correlations.

Table 38 Correlations between readability predictors and the normalised readability score obtained during this experiment. Significant correlations (at $p \leq 0.05$) in **bold**. Values apply to excerpts only (n = 804).

Correlation with Normalised Readability Score			
Type	Variable	Pearson Correlation Coefficient	Significance
Readability formula component (shallow)	Average Word Length	-0.191	0.001
	Average Sentence Length	-0.153	0.007
	Ratio of Long Words	-0.183	0.001
	Percentage of Polysyllable Words	-0.168	0.003
Lexicosyntactic	Subordination	0.123	0.031
	Passivisation	-0.109	0.056
	Lexical Density	-0.138	0.015
	Parse Tree Depth	0.016	0.777
Based on assessor comments (genre adaptation)	Acronym Score	-0.223	< .001
	Example Score	0.103	0.07
	Average Number of First-Person Pronouns	0.305	< .001
	Perplexity (Coca)	-0.337	< .001
	Normalised Perplexity (Coca)	0.09	0.113
	Perplexity (Coca + BNC)	-0.316	< .001
	Normalised Perplexity (Coca + BNC)	0.1	0.08
	Perplexity (GloWBE)	-0.254	0.036
	Normalised Perplexity (GloWBE)	0.119	0.036

Table 39 Correlations between genre-specific readability predictors introduced based on comments and previously used readability measures. Significant correlations (at $p \leq 0.05$) in **bold**. Values apply to excerpts only (n = 804).

Feature	Average Word Length		Average Sentence Length		Ratio of Long Words		Percentage of Polysyllable Words		Subordination		Passivisation		Lexical Density		Parse Tree Depth	
	Coeff.	p	Coeff.	p	Coeff.	p	Coeff.	p	Coeff.	p	Coeff.	p	Coeff.	p	Coeff.	p
Acronym Score	-0.007	0.902	-0.033	0.567	-0.09	0.115	-0.004	0.946	-0.143	0.011	-0.021	0.297	0.297	< .001	-0.143	0.011
Example Score	-0.042	0.46	-0.037	0.511	-0.064	0.258	-0.083	0.147	-0.044	0.441	-0.089	0.118	0.105	0.064	-0.61	0.281
Average Number of First-Person Pronouns	-0.274	< .001	-0.191	0.001	-0.288	< .001	-0.26	< .001	0.14	0.013	-0.31	< .001	-0.403	< .001	.056	.324
Perplexity (COCA)	0.121	0.034	0.075	0.19	0.114	0.015	0.014	0.809	-0.184	0.001	0.073	0.201	0.595	< .001	-0.123	0.03
Perplexity (BNC + COCA)	0.155	0.006	0.018	0.751	0.139	0.014	0.007	0.898	-0.198	< .001	0.05	0.385	0.626	< .001	-0.155	0.006
Perplexity (GloWBE)	0.093	0.103	-0.055	0.333	0.104	0.066	-0.086	0.132	-0.213	< .001	0.096	0.092	0.595	< .001	-0.162	0.004

As we can discern from Table 38, most predictors' correlation with the normalised, averaged score is significant enough that we can expect them to be informative when building a regression model to predict the latter; this is certainly the case for each of the variables we introduced based on assessors' comments. The presence of acronyms and first-person pronouns, for instance, exhibit highly significant correlations with the score (at $p \leq 0.001$). While their correlation coefficients would fall closest to the 'weak' classification of correlations (at -0.223 and 0.305, respectively), we can still observe that they are relatively high – arguably notably so given the relative complexity and unpredictability of scoring decisions. This outcome also lends credence to the notion that participants were able to adequately reflect on the scores they assign.

Furthermore, we also find that the various ways of measuring difficulty in the more traditional readability formulae – word length, sentence length, and the two ratios of longer words – also correlate significantly with the normalised score. While we have previously (in sections 2.3 and 6.5) illustrated how the readability formulae are less nuanced and complex than e.g. the Flesch score, these correlations do support the formulae's validity in spite of their reductiveness. They affirm Flesch's (1979, p. 21) assertion that while the readability formula approach “seems like a very crude way of dealing with writing [...] it is based on some very complicated facts of human psychology.”

Within the genre of corporate reporting, this study finds at least some evidence for both these 'shallow' features' crudeness and their sound basis in human psychology. As the second table demonstrates, we can see some significant correlations between these most basic surface features and the genre-specific determinants of readability (such as exemplification, use of first-person pronouns, and perplexity) even when there is no direct causal connection (such as there is in the case of more first-person pronouns, which are generally short, also shortening average word length). These correlations are also evidence for the aforementioned notion that the learner may perform marginally better without the addition of the genre-specific features because, as we posited, the learner already captures these effects through the more generic features, and adding the genre-specific features creates more noise than it enhances accuracy.

Two potential predictors, however, are remarkable in not meeting a .05 level of significance: extent of passivisation and parse tree depth. Given how close the passivisation variable is to the threshold of significance, in addition to the considerable evidence we have discussed for its role as an impediment to readability, we see very little reason to doubt its value as such. Parse tree depth, however, out of all the variables we examine, is the only one not to display any meaningful or significant correlation with the normalised score. While it is a common variable in readability prediction (e.g. Dell'Orletta et al. 2014 or De Clercq & Hoste 2016) and Beaman (1984), *inter alia*, considers it a component of syntactic complexity, these outcomes do not support a link between syntactic depth and readability – at least within the genre of sustainability reporting.

One possible explanation for this phenomenon lies in the genre adaptation effort. Given the subcorpus' fairly uniform high complexity, parse tree depth may become a relatively poorer differentiator between texts. This is unlikely to be the only variable at play because the full-corpus analysis (see section 3.4.1) found at least modest variability for parse tree depth.

From a more theoretical angle, we might also assert that a correlation of virtually zero does not support the 'chunking' (Pearson 1974) perspective of greater syntactic depth reducing the need for inference on the reader's part either, as that would imply a stronger positive correlation. It may be possible, however, that a greater syntactic depth shifts difficulty to different areas of processing rather than unambiguously increasing or decreasing it. We might expect a similar scenario for extent of subordination, where the sign of the correlation suggests that more subordination very weakly combines with greater readability. Nevertheless, as both manifested differently in different language varieties, they remain valuable to take into account when characterising a genre of text potentially sensitive to differences in language variety between author and audience. If anything, these outcomes indicate that syntactic depth and use of subordination deserve more scholarly attention in how they interact with reading ease.

If we consider how the previous examined features correlate with the ones we introduced based on assessors' comments (see Table 39), results for lexical density are likely the most salient. We see that greater lexical density correlates with a lower use of first-person pronouns – and again, while conventions might label this a moderate correlation, it is a large one relative to those we have previously examined. Higher lexical density appears to similarly correlate with both a greater use of acronyms and greater perplexity (i.e. a greater difficulty of predicting elements in a sentence based on the language model). Again, given the size of these correlations, adding the genre-specific features to the learner may create more noise than it enhances accuracy, as the generic features already included lexical density. Additionally, while lexical density and perplexity are inherently likely to correlate (as function words tend to be a more closed class, greater diversity of content words and less use of common words go hand in hand) this does again help re-emphasise the importance of writing reports with the most everyday possible vocabulary.

To sum up the chief findings of this chapter, we can conclude that, while human assessment of reading ease takes into account exponentially more aspects of the text than formulae do, the logic underlying these formulae is nevertheless sound. However, that does not necessarily vindicate their use as a sole means of readability estimation, especially for genres potentially as sensitive to poor readability as the non-financial aspects of business communication can be. However, we also found that it is viable, though both more challenging to implement and more computationally demanding, to approximate the 'gold standard' of the (ideal) human annotation scenario. Although an automatic system trying to predict which difficulty a human reader would assign to a text

might place different emphases than that human would, it is nevertheless likely to arrive at a similar score. While this does not replace an editor's job, it can nevertheless facilitate it; for one, the system could indicate to an editor how complex a text is compared to other sustainability reports, rather than compared to children's stories, which form the other extreme of conventional readability formulae's scale. That is, a system that builds on this proof of concept can give far more nuanced insight into a sustainability report's readability. In a slightly more elaborate form still, such a system might be able to indicate why it considers a text difficult or warn an author against their frequent use of uncommon words.

Enabling the system to explain *why* it perceives difficulty would likely be a necessary step before making such a system widely available, as this chapter's results also provide ample evidence for the risk that 'writing to formulae' entails (see e.g. Klare & Buck 1954). While writing to optimise the score assigned by a more holistic system, any author – and especially those writing to as broad an audience as these reports may have – must remain mindful that no matter how advanced the system advising them, those systems are still just trying to approximate a 'gold standard' to which the author already has access: that of human judgment.

Part 3: Towards Better Reporting: Discussion and Conclusions

Chapter 7

Discussion

One of the key conclusions of the preceding chapters is that, in spite of a number of crucial differences, from a presentation perspective sustainability reporting resembles financial reporting in more ways than it differs. This is most notable in a number of cases where that resemblance risks actively impeding the genre's stated goals. Sustainability reporting's self-asserted wider stakeholder audience (as also discussed in Townsend, Bartels & Renaut 2010, Bouten 2011 and Jenkins & Yakovleva 2014) has been a central theme throughout these chapters because of the mismatch between this audience's requirements and the reports' textual features. Every chapter has encountered ways in which this relatively new genre fails to accommodate or at least suboptimally accommodates the far more limited number of assumptions the company issuing the reports can make about their readership. The members of this wider-stakeholder audience are less likely to be experts in the company's operations, less likely to have their interests align with the company's, and may even be less likely to share a language (variety) with the company issuing the report. Furthermore, as motivation and engagement can also influence the ease with which a reader processes a text, we must also note that a shareholder may be far more intrinsically motivated to strive for a full, thorough comprehension of a report. That is, the threshold of difficulty they are willing to tolerate before deciding not to engage with the text may often be higher (see section 2.4.3.4).

All of these factors have a substantial impact on the plausible 'worst case' scenarios a company ought to prepare for if they are aiming to address their stakeholders rather than their shareholders, especially if they want them to be able to fully decode their reports. Whether they indeed want to do so is a key question, and we have encountered considerable evidence that reporting practices may align more with the notion that engaging in CSR communications efforts is more important than ensuring that those efforts are efficient. We might attribute this form of 'greenwashing' (e.g. Hrasky 2012, Boiral 2013) to the power disparity between the various groups of stakeholders (see e.g. Bouten 2011); as stakeholders grow more liminal to the company's operations, their leverage over the company – and thus the consequences for the company not

accommodating their requirements – tend to decrease. That is, a shareholder – the strictest-sense stakeholder – has the option of divesting themselves of those shares if they no longer believe the company is upholding their fiduciary duty. A (member of the) community local to where a company operates typically has some recourse if they feel the company is not adding value (or in fact detracting from it); such mechanics tend to fall under the umbrella of legitimacy as an operational resource. However, they are less direct than a shareholder’s fairly immediate ability to no longer be a shareholder. A local community’s ability to no longer be a stakeholder is, by comparison, a longer-term process, if not impossible. In other words, they have less immediate sway over the company, and especially in the case of a firmly entrenched one may also rely on it for, for example, employment.

Nevertheless, in spite of this discrepancy, it has proven a valuable exercise to hold the corpus up to the linguistic standards that a stakeholder-inclusive, transparent CSR strategy would require. While it is naïve to assume a sincere intent to reach all stakeholders for every company in the corpus, if the genre’s asserted audience is one broad enough to include a wide set of stakeholders, the most liminal members’ requirements are the ones by which the genre can and should be held accountable. Furthermore, while a greater accessibility might be more important to sustainability reporting in order to achieve its stated goals, the financial reporting genre might similarly benefit from greater reading ease. As it stands, corporate reports’ reputation of often impenetrable and simultaneously overly positive business jargon is unlikely to attract readers beyond the core audience of shareholders and analysts. While companies may benefit little from such an expanded audience, initiatives towards greater accessibility of disclosures led by regulatory agencies, such as the SEC’s Plain English Handbook (1998, p. 3), underline the wider benefits more accessible reporting may have in general:

“Investors need to read and understand disclosure documents to benefit fully from the protections offered by [...] federal securities laws. Because many investors are neither lawyers, accountants, nor investment bankers, we need to start writing disclosure documents in a language investors can understand: plain English.”

As this plea was two decades old at the time of writing, it seems clear that there is work yet to be done – for both financial and sustainability reporting. As this study worked with only plain text, however – stripping away paratext such as graphs and figures – we should note that readers struggling with textual comprehension may be able to fall back on visual aids to a greater extent than this study was able to investigate. However, as Cho, Michelon & Patten (2012a, b) found, these visual aids are often subject to their own extent of distortion and manipulation.

We find one plausible reason for sustainability reports’ higher difficulty in the annual (financial) reports’. The process of isomorphism (DiMaggio & Powell 1983), may well drive

both companies' CSR reporting practices and the design decisions behind sustainability reporting. Many companies began issuing CSR reports because it was becoming the de facto standard to report (KPMG 2013) and not reporting was becoming a competitive disadvantage, rather than out of a vision of what that reporting process should look like. As a consequence, driven by the same principles, the younger genre may well have evolved or been conceived of to resemble the older because emulating it might also have evoked its positive characteristics of being a well-established, legitimate and credible genre.

The full-corpus analysis found, based on a set of readability formulae and syntactic features that we used as proxies for deeper-level accessibility, that sustainability reporting is as difficult or more difficult than financial content. This conclusion stems from a direct comparison between the two content types in LtSs. Based on the above 'mimetic isomorphism' interpretation, this need not be an unexpected or implausible outcome, especially given that every company present in the corpus will already have a firmly entrenched linguistic toolkit they approach reporting with. Unsurprisingly, that toolkit confirms many of the less favourable aspects associated with the genre: it makes for difficult reading, is fairly lexically dense and notably more indirect (in its use of passives) than more general-purpose text tends to be. These are stylistic aspects typical of the genre that, if removed, are liable to negatively affect the reading experience of those – often higher-powered – stakeholders that interact with the genre most frequently, just as the removal of the consistently positive tone might. This deserves some nuance in that findings suggest that the negative impact of simplifying reports would likely be rather modest, if not entirely negligible. However, the more active voice that the SEC recommends might also cause impression management concerns when a company attributes more unfavourable outcomes to themselves, rather than creating distance through passive structures. In short, although many of these stylistic characteristics are undesirable from the perspective of a less experienced stakeholder reader, their presence is organic to the genre's aims, as is their transference to sustainability reporting given the conditions in which the latter genre arose.

We must also note that, in contrast with the strata of legal enforcement that Leuz, Nanda & Wysocki (2003) report as relevant to impression management tactics¹, we find that the variety of English that companies report in has the most significant impact on reports' FRE scores, lexical density and extents of subordination and passivisation. This effect occurs even within the same cluster of legal enforcement, for instance with US reports containing significantly fewer passives than British ones, and the latter exhibiting significantly lower lexical density. There are numerous reasons why we might see these differences (and why they merit further exploration), but for instance in the case of

¹ Cho et al. 2012a find limited evidence for the influence thereof.

passivisation, a likely one is that American English is generally more direct (Precht 2003a, b).

Although we can expect the composition process of these reports to mitigate the effects of individual linguistic preferences on one author's part, that process need not diminish the impact of effects shared amongst the language variety or varieties of those composing the reports. Given that we detected the presence of defensive attribution behaviour on a sentence level, these different preferences between varieties might have far-reaching impacts. For instance, although an elevated number of passive structures appears to be the norm throughout the genre of corporate reporting, an awareness of the association between positive information (or a lack thereof) and agency framing might cause an American reader to interpret a British report as evasive or overly cautious and potentially negative. Similarly, a British reader might interpret the American report as overly positive or lacking nuance. As the impact of language variety on the language of corporate reporting remains highly underexamined, these are certainly outcomes that invite further inquiry on the matter.

While this study was not able to examine the effects of such a dissonance between regions, results for the manipulation component at least indicate that there will likely be some effect. We found that the complex sentences and high extent of passivisation that are iconic to the genre did affect non-experts' perceptions of the company – remarkably, for the better. This effectiveness is likely one of the main reasons behind the high linguistic complexity of the genre: as an impression management strategy, it appears to work – at least for a part of the audience.

As studies such as Leavy, Li & Merkley (2011) indicate that investors need to rely increasingly on analysts as disclosures' readability goes down, we can assume that those analysts and potentially other experts will be less - or not at all - affected by textual impression management tactics. Optimising the impression their text leaves on (relative) laypersons is then likely a best-case scenario for companies issuing disclosures. Moderately or exhaustively applying the Plain English guidelines to the document did not appear to detract from expert judgments, but it did normalise laypersons' perception with that of their more experienced counterparts. This is likely because they are able to achieve a similar level of understanding as the non-laypersons but this effect may also be due to the unaltered LtS better conforming with genre conventions, i.e. better adhering to what laypersons feel a LtS 'should' look like, and thereby attributing more positive aspects to it. Regardless of which is the case, we can note that while a more accessible text erases some of the benefits of complex disclosures, it does not detract from the audience's perception of the company, regardless of their degree of expertise. Overall, readers appeared unlikely to think less of a company's professionalism or credibility if companies communicate their results in simpler language.

The same question of how adherence to genre conventions influences the reader's impression is also important to sentiment analysis. Although we found high overall

positivity that aligned with the so-called ‘Pollyanna Effect’, this pilot study leaves a number of questions unanswered. Further sentiment analysis might benefit from a manipulation study similar to the one on readability; this might reveal how readers react to a greater or lesser extent of positivity in the text. While excessive positivity limits the amount of nuance the genre can display, we might equally expect that no company would want to be the first to publish a factually focused report that attempts to avoid words with positive or negative associations. Such words are a tightly interwoven aspect of how the company forms a narrative around the numbers on which it is reporting. Due to this inflation of positivity, a company minimising the positivity with which it interprets its results may well face disastrously worse perceptions of its performance than would be accurate.

Finally, we note that assessors responded favourably to more narratively framed disclosure content that attempted to engage with its readers. This may well be an organic direction for sustainability reporting to take as it further evolves into its own niche as a genre, or attempts to form a harmonic whole with financial reporting in the shape of an integrated report. If the core of expert readers is unlikely to think less of a report or the company issuing it because it uses less complex language, companies may have less reason to fear adverse effect from its more powerful stakeholders than they might expect. While there is certainly a point where language might become too simple to express the complexity it is attempting to portray, assessors’ analyses suggest a wide margin left before reports reach that point, and most authors and editors will have the insight to stop reducing (the complexity of) a text before it reaches the other end of the spectrum.

Chapter 8

Implications

In spite of the potential tension between the goals that those issuing sustainability reports claim to pursue and the goals they might actually have for the genre, we can still identify considerable demand for more readable reports. This demand comes from a wide range of sources, in addition to companies' internal signalling of intent to communicate in a stakeholder-inclusive fashion. Approaching the topic from an academic angle, Farewell, Fisher & Daily (2014) see room for companies to use less complex language. From the auditors' perspective, in 2013 (the date closest to most of the corpus' publication) KPMG's *Survey of Corporate Responsibility Reporting* emphasised that "[yes], CR reports are often not an easy read and companies should seek to communicate information in more digestible and engaging ways" (p. 10). They add that this difficulty does not excuse companies from issuing such reports, and, as a corollary, we might assert that in terms of transparency a difficult-to-read report is still better than no report at all.

We have already highlighted the regulatory impetus towards more accessible reporting in the SEC's Plain English Handbook (1998), but the sustainability-oriented GRI framework (2013) also highlights the principle of 'Clarity', which stipulates that

The report should present information in a way that is understandable, accessible, and usable by the organization's range of stakeholders (whether in print form or through other channels). A stakeholder should be able to find desired information without unreasonable effort. Information should be presented in a manner that is comprehensible to stakeholders who have a reasonable understanding of the organization and its activities.

In other words, even when it does not necessarily align with the authors' interests, more powerful stakeholders and regulatory agents also have an interest in companies issuing comprehensible disclosures, of which readability is an important component. We expect this demand to continue to grow into a requirement as sustainability reporting continues to evolve from the voluntary to the mandatory (KPMG 2017). However, with the exception of the SEC Plain English Handbook, this demand for more readable reports is seldom accompanied by explanations on how to achieve it. Accordingly, this section aims to

formulate a number of recommendations based on findings throughout the chapters to help companies achieve more readable reports.

We must note in that respect that while there is certainly room to make corporate reporting a more readable genre, some of the complexity of the topics it reports on will be irreducible – or, at the very least, impossible to reduce through choice of words and linguistic structure alone. Reports from older (and possibly larger) or more complex organisations may hint at this phenomenon (see e.g. Rutherford 2003, Li 2008): organisational complexity may be a determinant of textual complexity, but the latter is substantially easier to reduce than the former (up to a point). Chapter 4 found that editing an LtS from a sustainability report to a Plain English level of readability was effectively infeasible without compromising the message. Furthermore, based on the same chapter’s findings that deeper sentence structures may enhance cohesion in spite of increased cognitive load, this ‘easiest’ version according to formulae already began to see diminishing returns or even adverse effects in terms of readers’ actual experiences with the text (Van Hoecke 2018).

This highlights another issue: while it has been a frequent caveat throughout these chapters, any set of recommendations for more readable report writing must caution against a ‘writing to formulae’ approach to increasing reading ease. As the comparison between the scores participants assigned to excerpts and those excerpts’ FRE scores made clear, while the two certainly have some relationship (and readability formulae are a useful yardstick, to borrow a term from Flesch himself), the association between the two is too weak to argue that readability formulae can substitute for human judgment. By contrast, the author(s) of these documents invariably have access to human judgment – their own - as a readability metric. Readability formulae’s value lies in being an objective standard that can help authors discover where their judgment of a text’s readability may have failed,¹ but tweaking a text towards better formula-based readability is likely to backfire, as, for instance, using words that the audience are likely to be familiar with is more important than using the shortest possible ones. While the latter would optimise the text to show the best possible readability score, using ‘lest’ instead of ‘because we are worried that otherwise...’ will not bring a text closer to universal readability in spite of its greater brevity.

The same is likely true of sentence structure. While sentence length is a component of the readability formulae and, for instance, the SEC Plain English Handbook recommends using shorter sentences where possible, this addition of ‘where possible’ is an absolutely

¹ These misjudgements may occur due to the so-called ‘Curse of knowledge’ (see e.g. Camerer, Loewenstein & Weber 1989 or Kennedy 1995), the phenomenon where an expert fails to accurately judge the difficulty of a specific task or knowledge required to complete it due to the difficulty of thinking as they would without this knowledge. For another iconic example, consider the stereotype of an academic unable to translate their expertise into generally understandable terms.

crucial condition, although ‘where possible’ is perhaps better phrased as ‘where likely to improve readability’. We can illustrate this. Short sentences can emphasise important ideas. They can also be grating. At worst, they detract from cohesion. Authors should use them carefully. They should also use longer sentences. One crucial advantage of longer sentences is that they are better able to show what is otherwise an implicit connection between ideas. Instead of the staccato rhythm of those previous sentences, phrasing them as ‘Long sentences can emphasise important ideas, but they can also detract from cohesion, so authors should use them carefully’ would have much better conveyed the underlying logic of how the sentences interrelate. Nevertheless, a sentence the length of the preceding one can also cause excessive cognitive load. This is because of a phenomenon that readability formulae do correctly account for: sentence boundaries are ‘rest points’ for the short-term memory. In that respect, it is likely better advice to terminate sentences wherever doing so is organic, rather than where it is possible. Again, by the simple virtue of being human, authors are more capable of performing what we might term ‘empathic’ natural language processing than any Natural Language Processing can be. They will likely outstrip computers in that respect for the foreseeable future – at least until NLP technologies make another several generational leaps. Rather than writing to formulae, authors should use readability metrics as a ‘second opinion’ to supplement their own.

In terms of deeper-level sentence structure – specifically, parse tree depth and extent of subordination – we found little evidence for any obfuscation patterns based on corporate performance. As parse tree depth and extent of subordination are also the least intuitively interpretable of the readability metrics (which also see use in e.g. Dell’Orletta et al. 2014) we can only make cautious recommendations about the extent of syntactic depth and subordination authors should use. In the case of parse tree depth, this is compounded by findings based on human annotation suggesting no meaningful correlation between the normalised score and parse tree depth as a variable. We would argue that, because it is unlikely that syntactic depth does not affect readability or understandability, this outcome is due to tension between the ‘deep structure’ interpretation of syntactic depth that implies complexity and increased cognitive load, and the ‘chunking’ interpretation that the preceding paragraph also illustrated, which asserts benefits for the reader in processing semantic wholes (Pearson 1974).

That is, deeper sentences are likely to have both beneficial and harmful effects on a text’s ease of understanding; parse tree depth may not have a linear relationship with readability, and for most audiences, either extreme is likely to detract more from readability than contribute to it. Low-proficiency readers might be an exception, as deeper, more complex sentences are more likely to use structures those readers are less equipped to deal with, or cause a higher cognitive load than they are capable of processing. That implies that a text that aims for universal understanding may require suboptimal choices for general-audience readability, as those audiences will be capable

of dealing with the more complex structures and increased cognitive load, and will simultaneously benefit from the advantages that more elaborate ‘chunking’ offers.

Given the many interlocking factors involved, syntactic depth may be one of the most crucial areas for authors to favour their own ability to process and evaluate language over that of automatic systems. While they must primarily rely on their own insights into the language, their expertise can also make it difficult for them to accurately gauge how complex a text might be for those with less expertise (analogous to e.g. Camerer, Loewenstein & Weber 1989 or Kennedy 1995). In that respect, readability metrics can still provide an invaluable ‘second opinion’ to mitigate this ‘curse of knowledge’.

Use of subordination is similarly problematic. It exhibits a significant, but extremely weak positive correlation with the normalised score. While the correlation coefficient is too low to assert that more subordination leads to more readable disclosures, we similarly find evidence that more subordinate clauses need not invariably lead to less understandable writing. Again, we note that Dell’Orletta et al. (2014) and Collins-Thompson (2014) integrate it as a variable in readability prediction, as do De Clercq & Hoste (2016), but Pearson (1974)’s assertion that there is more to how subordination and complexity interact than a linear relationship seems to be as applicable here as we found it to be for syntactic depth. For instance, while a subordinate clause adds structural complexity, it can also assist in making the underlying information more coherent and cohesive. As an example, while (143) ‘s use of subordination creates a longer and more structurally complex sentence, compared to (144) it more explicitly highlights the deliberate causal relationship between the two ideas the sentence conveys:

(143) To ensure the optimal structure for a growth-oriented and geographically diverse business, PanAust is structured into three business units: Asia, South America, and Project Development. (PanAust 2013)

(144) PanAust is structured into three business units: Asia, South America, and Project Development. This ensures the optimal structure for a growth-oriented and geographically diverse business.

Again, we can expect either extremely frequent or infrequent use of subordination to have its disadvantages. We find that, while it would be difficult for automatic systems to make such judgments, authors writing for a general audience are likely best served neither omitting subordinate clauses nor inserting them where doing so would be inorganic, as neither would be unambiguously likely to improve understandability.

The impact of passivisation should be more straightforward, but we can observe, similar to subordination, that it only has a very slight (and not quite significant) correlation with the normalised score. Use of passives, however, sees significantly more theoretical support as a detractor from readability, most notably in the SEC’s Plain English Handbook (1998). For syntactic complexity, Dell’Orletta et al. (2014) and Collins-Thompson (2014) use parse tree depth and amount of subordination as predictors for readability. However, Pearson (1974) sees a more complex relationship between the two. By contrast, a considerable number of sources (see section 2.4.2.2) argue with greater

consensus that passivisation lowers reading ease. This assertion remains largely uncontroversial; all else being equal, an active-voice structure conveying the same information will almost certainly be easier for readers to process than a passive-voice one. Again, there are of course sentence structures in which the passive is the preferred voice, and, as with any alteration, authors are unlikely to improve readability by warping their sentences towards inorganic structures.

There is, of course, also the issue of defensive attribution structures that appear an intrinsic genre convention, similar to how a high degree of positivity is. This convention might make active-voice structures appear less organic to authors within this genre than they might be to readers; Chapter 4 found that increased use of passives does not appear to detract from credibility or professionalism. Based on the results of Chapter 6, we see further reason to recommend that authors strive to minimise their use of the passive. While Chapter 3 found no evidence for companies using obfuscating agency patterns on a document level based on overall performance, Chapter 5 did reveal strong evidence for such tendencies on a sentence level.

Although this use of the passive is entrenched in the genre conventions (corporate reporting uses multiple times the number of passive structures that most text does), reports would benefit from more active-voice constructions, not just in terms of readability but also, crucially, in terms of transparency. Clear, explicit agency patterning is one of the key linguistic choices to ensure greater accountability in reporting. Finally, we emphasise that, based on the significant differences in passivisation patterning between regions, disclosures from a region inclined towards more passivisation might register as evasive to readers from one used to less. All of these combine to make corporate reporting with as few passive-voice constructions as organically possible highly desirable. Fortunately, detecting such constructions is a fairly trivial task for NLP applications and other automatic editing assistance systems, including the ubiquitous Microsoft Word.

As a drawback to potentially any of these proposed alterations, we should note that responses to a low-readability LTS were more positive amongst a layperson audience than one with greater expertise. This difference did not exist for the two simplified versions of the text, and is thus best attributed to a low readability target enabling or impression management techniques that are capable of influencing laypersons, but not those with more experience. That is, the genre conventions are likely what they are because they are effective at optimising the impression the company makes on a non-expert audience. However, as reporting regulations continue to shift from the voluntary to the mandatory (KPMG 2017), we can only expect attention to (linguistically) transparent reporting to increase. Especially in areas or industries where GRI-compliant reporting becomes mandatory, these findings inform recommendations that can help companies better pursue the tenet of clarity, and can at the same time help keep readers aware of where companies might make impression management decisions that detract from readability.

Regarding the Pollyanna Effect, we can identify risks both in its presence in the documents and its potential removal. The genre's de facto conventions of excessive positivity have the chief downsides of potentially warping the impressions of a reader less familiar with the genre, and of limiting the amount of nuance available to express what went well and what went poorly. Reports that have access to the much wider range of positivity and negativity markers that other genres do will inevitably be more balanced and transparent. A rapid change towards more nuanced use of sentiment would likely have a significant disadvantage, however: it might translate poorly to the perceptions of those more familiar with the genre conventions, i.e. those who are capable, perhaps even subconsciously, of discounting the genre's typical positivity back down to what reports are actually communicating. As applying that compensation to a more nuanced report without inflated positivity might yield an equally or more warped impression, this scenario is likely less desirable still than an excessively positive one. In short, although the genre is likely to gain nuance and balance in shedding its excess positivity, it should do so gradually, lest it change more rapidly than users' perceptions of it can accommodate.

The assessment phase preceding machine learning experiments likely provided the most genre-tailored advice for authors intent on ensuring the greatest possible understanding and engagement amongst their audience, although many of these recommendations are difficult to quantify and will rely heavily on the authors' own judgment and efforts. Perceived inhibitors of readability include poor flow (e.g. through use of product names and acronyms), vagueness, excessive complexity, numerical information as well as inadequately framed concepts that assume background knowledge on the reader's part. In addition to issues to do with readers' knowledge and capacity to deal with complexity, the most notable problem appears to occur when readers perceive attempts to create engagement with the reader as inadequate. One major - and likely cost-efficient - recommendation is for authors to include a lexicon of technical terms and abbreviations, so non-experts can better fill the gaps in their knowledge without the terms' definitions detracting from experts' reading flow.

The chief factors that assessors indicated as enhancing readability were efficiency and engagement. Annotators perceived an active voice, variations in sentence length and a narrative, sufficiently informal style that uses personal pronouns as enhancing their motivation to continue reading. Furthermore, a text that does not make demands on the readers' knowledge outside of the text, as well as providing some redundancy in the form of examples and explanations appears easier to process, especially when counterbalanced by a good structure that avoids unnecessarily complicated language. Equally notably, readability appeared to improve when authors used a fairly low lexical density, allowing room for function words that highlight the aforementioned structure (i.e. cohesion markers). In spite of how sweeping these recommendations may appear, assessors' ability to give positive examples of readability-enhancing features indicates that there is room

for such changes within sustainability reporting, and authors may be well served creating more narrative, perhaps even conversational disclosures where possible to be less exclusionary towards non-expert readers. In that respect, the trend towards question-and-answer style conversations taking over the role of Letters to Stakeholders seems a fine compromise between experts' and laypersons' requirements.

Chapter 9

Future Research

As the growing adoption of mandatory reporting guidelines will illustrate, sustainability reporting is a rapidly evolving genre, admittedly likely more so than linguistic research into it.¹ Given the time expenditure of data collection, experimentation and processing, the metaphorical goal posts of reporting have already changed by the time research is able to address the genre's current (or, at that point, former) state. That inevitably makes this study a scaffold for further research. This section will, based on its findings throughout the previous chapters, formulate a number of directions in which future studies might explore.

A first one, due to the rapidly evolving nature of the genre this study examined, is a diachronic study that examines how a set of companies' reporting practices evolve over time, and what influences them to do so. While increasingly mandatory reporting is likely to influence the report's textual aspects, there is more than one way in which it might do so. An increased, mandatory adoption of guidelines and reporting principles such as GRI could incent greater linguistic transparency, but may also lead to more obfuscation as unfavourable outcomes shift from omittable in a voluntary reporting climate to a necessary inclusion in a mandatory reporting climate. Such a shift may cause a greater occurrence of obfuscation as companies' opportunities to practice impression management shift to later in the reporting and editing process, i.e. *how* they phrase or include it rather than whether they do at all.

Such a diachronic study would come with the critical caveat that corpus collection and processing can be an extremely time-intensive process, especially if the researcher aims to apply more sophisticated NLP techniques to it, which require a high standard of 'cleaning' (i.e. removing potential noise) if they are to function reliably. While a stray additional character would affect formula-based results, formulae are fairly robust to occasional noise. NLP techniques, by contrast, tend to require well-formed sentences, and

¹ By nature of the academic process, published results tend to describe data sets at least a few years old, such as is the case for this study.

suffer more severely from smaller amounts of noise, which can affect not just a word, but also its context, and thereby the processing of the rest of the sentence. Moreover, in very noisy text NLP stops yielding any meaningful output whatsoever. While studies such as Van Hee (2017) demonstrate that it is possible to apply NLP to other textual data than well-formed sentences (in this case tweets), such adaptation is notably difficult and labour-intensive. Alternatives would be to ensure a high availability of labour available for PDF-to-plaintext conversion, to further automate the process, or to carry out such diachronic research on a smaller corpus of companies.

Rather than the aforementioned lack of evidence for performance-based obfuscation on a text-level (which, again, merits continuous evaluation under shifting regulation), our primary finding based on the full-corpus analysis was that the language variety a company writes a report in has notable effects on that report's syntax. We might hypothesise that this can have considerable impacts on audiences approaching such documents from a different linguistic background, and may well change these cross-varietal audiences' perception of the results. Testing such a hypothesis could lend additional insight into how influential this discrepancy between varieties actually is. One fairly straightforward testing method could be to investigate how British and American readers – both expert and layperson – respond to the (increased) presence or absence of passive structures. Another might be a more qualitative rhetorical analysis of reports or LtSs issued by similarly performing companies headquartered or listed in different regions. While we found variation between reports that aligned with language variety much more than with broad-sense clusters of regulatory enforcement, some of that variation may also be attributable to how guidelines specific to particular regions have shaped the language of their reporting (e.g. the SEC's advice against passive structures). This, too, might merit a closer, more qualitative inquiry, especially with respect to how non-native users of a report's language respond to linguistic variation. The breadth of attested differences between varieties, however, also merits investigation into why we find the results we do, and to what extent these differences are limited to the genre of corporate reporting. Comparing the differences between reports in different varieties with text present in multi-varietal reference corpora (6.6.2) might provide further insight.

However, this study certainly does not provide conclusive evidence that there is no obfuscation in sustainability reporting; rather, it fails to detect it on a text level. Due to the multiple aspects of performance present in sustainability reporting, a valuable avenue for future research into obfuscation might be a more fine-grained, content-based analysis. That is, a study that determines readability separately for passages on the different aspects (for instance generating a readability score for environmental or social-themed paragraphs) and investigates these scores' association with the relevant performance. Farewell, Fisher & Daily (2014) used a lexicon-based approach to clustering sentences by content and found some differences in readability between the different

topics present in sustainability reports, but did not investigate how these passages' readability correlated with performance. The greatest potential of this approach lies in its potential to prevent the different perspectives present in sustainability reporting from 'averaging out' the full document's readability.

While the manipulation study yielded valuable results, especially involving how linguistic impression management techniques affect laypersons and non-laypersons differently, it also allowed for considerable room for expansion, especially given the outcomes of our pilot study into the sentiment present in the LtSs. For one, our questionnaire implementation was a fairly rigid format that allowed little room for open comments, which might have provided further insight into respondents' perceptions, much like the comments for the machine learning gold standard data did.

One such approach – likely better implemented in a fairly controlled setting rather than a wide survey – could be to present respondents with the three variants and the hypothetical problem that the company is debating which of the three to publish. The question would then be which the respondents prefer and why, and possibly how they can improve them. This would yield more qualitative insight but likely fewer data suitable to quantitative analysis. Somewhere between the two extremes, offering respondents the ability to comment on why they rated a given aspect of the text what they did is also likely to offer additional insight (but was unfortunately not a viable addition within our surveying logistics). Responses may also differ with accompanying visual information and paratext available, which this study was unable to include. Finally, comparing manipulated LtSs at a given readability score with unmanipulated ones very close to that score might be able to detect whether the manipulation itself had any adverse effects. However, that approach would no longer be able to keep the letters' content constant as this experiment did.

The pilot study into sentiment use also merits further expansion. The inter-annotator agreement scores that annotators obtained implied both the viability of the inquiry and, especially with regards to more difficult areas of annotation, such as subjectivity and governance performance, the need to further refine and iterate on annotation instructions. However, given the relatively high quality of the outcomes, it may be possible to refine annotations to the point of being high-quality gold standard data for a machine learning system that could automatically provide sentiment analyses of LtSs or even full reports. While such a system would inevitably yield lower-quality results than human annotation, it may well be possible to optimise it to the point of providing useful good news/bad news summaries and, if capable of taking into account agency framing, could even automatically (help) identify defensive attribution techniques.

Similarly, the results for human assessment-based readability scoring are promising in terms of making a fully automatic readability assessment tool for these genres. We provide proof of concept for a machine learning-based system with a far finer resolution. Specifically, there was as much resolution available to the machine learner within the

genre as readability formulae have to accommodate any genre of writing. While this study offered proof of concept, a major technical hurdle for widespread deployment would lie in accommodating the PDF format most typical to the genre. Although authors are likely to have access to the source text and thus be able to provide the plain text document format that the NLP underlying the system would need, potential readers will chiefly have access to the finished PDF. Under ideal circumstances, a finished, genre-tailored readability prediction system would have access to an automatic means of extracting only the running text from a PDF document. However, such a project would likely require technical and academic expertise from a wide number of disciplines. Nevertheless, it would be one of the better means of differentiating between reports that attempt to accommodate a wide audience (with various degrees of success), and those that do the genre's reputation of impenetrability justice.

Ideally, such a system would also be able to indicate *why* it assigns a given score, i.e. what the document is doing poorly or well in terms of readability, lest the system simply become a different, but equally opaque formula to write to. Without that ability, the system might again disincite authors from using their own judgment. A system that is also capable of indicating why it assigns a score would offer a far more valuable second opinion. Such a recommendation would require thorough review, however – as we previously emphasised, the best predictors in machine learning are not necessarily the most influential factors to the actual experience for humans. That said, if authors are to write to a formula, a machine learning-based formula trained on human assessment is likely the far better one to write to, given that it is capable of accounting for far more factors than readability formulae are – unsurprisingly so, given the vast gulf in technology now available compared to when readability formulae were initially designed.

Conclusion

On Writing Corporate Reports

This study perceived and attempted to address a number of gaps in research into corporate (sustainability) reporting. As this final part has attempted to synthesise the outcomes of the different chapters and studies contained therein, all that remains is to summarise the answers we have found to the gaps we perceived and the questions they, in turn, posed.

First, we found that, to a very large extent, financial reporting's often complex language transfers to sustainability reporting. Sustainability content is no more readable, and perhaps even less readable than financial content. Based on what we know about the worst-case demands that the requirement of general readability can place on a text, we can confidently assert that this case of mimetic isomorphism affects the genre's utility to a wider stakeholder audience. While the two may differ in terms of content, formally, sustainability reporting reads much like financial reporting does. In other words, because many companies may claim to want to engage with indirect stakeholders such as local communities, and because the sustainability report is the prime means for companies to communicate their CSR efforts, these (alleged) efforts are quite likely to be unsuccessful – or at worst, insincere. While it may appear somewhat naïve to hold a specialised genre to a standard of general readability, both researchers' findings on the potential width of the audience (e.g. Townsend et al. 2010, Bouten 2011, Farewell, Fisher & Daily 2014) as well as companies' own assertions of stakeholder inclusivity (e.g. Adidas 2013) justify examining the genre in these terms. We found, however, that it may well be true of this 2012 corpus that claiming to engage with all stakeholders had greater utility for the company than actually achieving that communication. However, that utility may well change as regulators' demands shift towards increasingly mandatory reporting on non-financial issues.

In spite of the above, we must also acknowledge that universal understandability as the formulae understand it – true 'Plain English' at a FRE score of 70 or above – is likely unattainable for the genre of nonfinancial reporting. As financial reporting entails

covering highly complex themes, and sustainability reporting adds a number of other performance aspects – including social and environmental performance – the genre, in terms of content alone, will be faced with irreducible complexity that no amount of linguistic simplification can compensate for while still conveying the same content. As we saw that organisational complexity can correlate with linguistic complexity, there must inevitably come a point where reading ease comes at the expense of transparency – *Ulysses*, by analogy, could be written using the words and structures of *The Cat in the Hat*. Even if such a reduction were possible, it would likely frustrate more expert readers. The latter may be willing to tolerate greater linguistic complexity if it yields more depth and nuance. This, more than anything, underscores the dangers of writing to formulae: authors must first rely on their own judgment, and rely on readability tools and metrics as an aid and a tool – a yardstick – not as a higher authority than their own.

We also investigated how the readability of corporate reports interacts with the performance underlying them. On a document level, we found very little evidence for obfuscation as previous studies into financial reporting have reported it, not even when looking into the LtSs from financial reports. For the sustainability-themed documents, at least, we might argue that linking readability with performance hinges on the crucial question of *which* performance. As we identified financial, social, environmental and governance-related performance aspects as relevant to these reports, we might also expect that the influence of any one performance aspect to readability will weaken. This is not to assert that there are no signs whatsoever of obfuscation within the corpus; there is simply little evidence of the most straightforwardly identifiable type thereof, i.e. a document-level positive linear relationship between performance and readability. In this respect, obfuscation analysis on, for instance, a per-paragraph basis that links the readability of passages on one performance aspect to that aspect could be a very valuable avenue for future research.

The corpus did, notably, exhibit signs of defensive attribution behaviour – a closely related type of impression management – on a sentence level. We found sentences containing positive news to use more direct agency framing, i.e. to be more inclined to use the first person when reporting on favourable outcomes. Sentences not reporting on favourable outcomes were more inclined to refer to the company indirectly, e.g. through metonymy, or attribute the outcomes to non-company agents. At the same time, however, rhetorical considerations appeared to take precedence: when reporting unfavourable outcomes in a high-engagement construction that emphasised the relationship between the author and the reader (such as an apology), authors were also significantly more inclined to use agency patterning closer to the first person. While this is evidence in favour of defensive attribution, those cases may not register as defensive attribution to all readers. As we found a high tendency towards passivisation overall, many familiar with the genre may simply consider this part of its conventions, regardless of their awareness of the impression management patterns behind agency framing.

We also noted that these low-readability, high-passivisation genre conventions may be as they are because of their effectiveness. Optimising disclosures for general-audience readability affected some aspects of the layperson audience's perception. However, it did not do so in a way companies might want: results indicated that, in a number of key aspects of perception, a layperson group of readers had a better impression of the company after reading the original, most difficult and passive version of a disclosure than a group with at least some experience did. Both a somewhat simplified and a greatly simplified version seemed to erase that effect. While these outcomes show little direct reason for companies to create more universally understandable disclosures, evolving regulations may increasingly force them to. As with many aspects of corporate sustainability, such a change might be a detriment to the company's short-term goals, but is likely to benefit them and their stakeholders in the longer term, as it would be demonstrably more transparent. No aspect of the more readable LtSs changed non-laypersons' opinions for the negative; it only brought laypersons' opinions in line with theirs. Again, these findings were potential evidence of obfuscation not detectable through a formula-based document-level analysis.

Just as we found evidence of high passivisation typical of the genre's reputation, it exhibited high positivity consistent with the 'Pollyanna Hypothesis'. In spite of the overall good performance for the companies analysed, a ratio of eight positive elements to every negative (likely) still violates the tenet of balance inherent to many reporting philosophies – if not balance in content, then balance in tone. Especially as the presence of positive or negative elements regarding a performance aspect did not exhibit a meaningful association with performance for that aspect, these results imply high positivity overall. Based on previous research, we can assert that this uniformly high degree of positivity independent of performance evidences the Pollyanna Effect. As was the case with the inflated use of passivisation, a gradual change in positivity would likely be better than a sudden one, as both are deeply ingrained genre conventions. However, implementing such a change would likely enable the genre to be more balanced, transparent and overall better than it is now.

We also found that there is certainly room to optimise the genre's readability to better cater to a wider audience, with the caveat that, if taken to an extreme, such changes might make it less appealing to a core audience with greater expertise. That is, to recall the definition of readability this study used, especially in terms of syntactic complexity there may well come a point where optimising for universal readability inhibits expert readers' ability to extract information from the text efficiently.

Automating genre-tailored readability prediction capable of far more resolution and nuance is certainly one way research can assist practice in ensuring higher-readability reports. Here, too, we must reiterate that the aim of such technology cannot be to remove decision-making for authors and enable them to 'write to the formulae'. Such systems must assist authors in judging and refining their own work, rather than setting the goal

posts for the author. Nevertheless, implementing genre adaptation to readability prediction appears to be quite viable from a technological perspective. While human judgment on a number of excerpts was able to formulate potential genre-specific readability predictors, a modern machine learning system did not benefit from those predictors' inclusion. Although optimal accuracy required retraining on the scores assessors provided, the generic, genre-agnostic system already considered a wide enough range of features to achieve excellent results.

To sum up, we can reiterate that we conceived of almost every aspect of this study with two primary aims. The first was helping readers become better readers by offering them a more critical insight into the linguistic dynamics of corporate (sustainability) reporting. The second, and perhaps the most crucial, was to enable authors to better communicate with a widening audience as well as understand where and why that communication might break down. For all the technical interpretations of readability and understandability we can offer, however, applying those insights falls on the authors and editors themselves: their own understanding of language can, if properly applied, outstrip that of any machine. While this study has, hopefully, enabled greater understanding of how the genre can improve, the onus of making those improvements falls on those writing it. We hope that they will continue to evolve non-financial reporting towards its full potential.

Bibliography

- Abrahamson, E., & Amir, E. (1996). The Information Content of the President's Letter to Shareholders. *Journal of Business Finance & Accounting*, 23(8), 1157–1182. <https://doi.org/10.1111/j.1468-5957.1996.tb01163.x>
- Abramahson, E., & Park, C. (1994). Concealment of Negative Organizational Outcomes: An Agency Theory Perspective. *Academy of Management Journal*, 37(5), 1302–1334. <https://doi.org/10.2307/256674>
- Abu Bakar, A. S., & Ameer, R. (2011). Readability of Corporate Social Responsibility communication in Malaysia. *Corporate Social Responsibility and Environmental Management*, 18(1), 50–60. <https://doi.org/10.1002/csr.240>
- Accounting Standards Board. (1993). *Operating and financial review*. London.
- Adidas Group. (2013). Sustainability Progress Report 2012: Performance Counts. Retrieved from http://www.adidas-group.com/media/filer_public/2013/08/13/adidas_spr2012_full.pdf
- Aerts, W. (1994). On the use of accounting logic as an explanatory category in narrative accounting disclosures. *Accounting, Organizations and Society*. [https://doi.org/10.1016/0361-3682\(94\)90001-9](https://doi.org/10.1016/0361-3682(94)90001-9)
- Aerts, W., & Cheng, P. (2011). Causal disclosures on earnings and earnings management in an IPO setting. *Journal of Accounting and Public Policy*, 30(5), 431–459. <https://doi.org/10.1016/j.jaccpubpol.2011.03.006>
- Aerts, W., Cormier, D., & Magnan, M. (2008). Corporate environmental disclosure, financial markets and the media: An international perspective. *Ecological Economics*. <https://doi.org/10.1016/j.ecolecon.2007.04.012>
- Aerts, W., & Yan, B. (2017). Rhetorical impression management in the letter to shareholders and institutional setting: A metadiscourse perspective. *Accounting, Auditing & Accountability Journal*, 30(2), 404–432. <https://doi.org/10.1108/AAAJ-01-2015-1916>
- Alexopoulos, P., & Pavlopoulos, J. (2014). A Vague Sense Classifier for Detecting Vague Definitions in Ontologies. <https://doi.org/10.3115/v1/E14-4007>
- Anglo American. (2013). *Creating Value with the Future in Mind. Sustainable Development Report 2012*.
- Arrium Limited. (2013). *Growth Through Sustainability. Arrium Limited Sustainability Report 2012*.
- Bafilemba, F., Mueller, T., & Lezhnev, S. (2014). The Impact of Dodd-Frank and Conflict Minerals Reforms on Eastern Congo's Conflict.
- Bailin, A., & Grafstein, A. (2016). *Readability: Text and Context*. Palgrave Macmillan UK. Retrieved from <https://books.google.be/books?id=epZ7CwAAQBAJ>
- Bayerlein, L. (2010). Positive Versus Negative News: Readability and Obfuscation in Financial Reports.
- BBC. (2013). Bangladesh factory collapse toll passes 1,000. Retrieved from bbc.com/news/world-asia-22476774

- Beaman, K. (1984). Coordination and subordination revisited: Syntactic complexity in spoken and written narrative discourse. *Coherence in Spoken and Written Discourse*, 12, 45–80.
- Bean, J. C., & Weimer, M. (2011). *Engaging Ideas: The Professor's Guide to Integrating Writing, Critical Thinking, and Active Learning in the Classroom*. Wiley. Retrieved from <https://books.google.be/books?id=Xbgs9MvcsjsC>
- Beattie, V. A., & Jones, M. J. (2000). Changing Graph Use in Corporate Annual Reports: A Time-Series Analysis. *Contemporary Accounting Research*, 17(2), 213–226. <https://doi.org/10.1506/AAT8-3CGL-3J94-PH4F>
- Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13(Feb), 281–305.
- Berliner, D., & Prakash, A. (2015). “Bluewashing” the firm? Voluntary regulations, program design, and member compliance with the united nations global compact. *Policy Studies Journal*. <https://doi.org/10.1111/psj.12085>
- Boiral, O. (2013). Sustainability reports as simulacra? A counter-account of A and A+ GRI reports. *Accounting, Auditing & Accountability Journal*, 26(7), 1036–1071. <https://doi.org/10.1108/AAAJ-04-2012-00998>
- Bormuth, J. R. (1969). Development of Readability Analysis.
- Boucher, J., & Osgood, C. E. (1969). The pollyanna hypothesis. *Journal of Verbal Learning and Verbal Behavior*, 8(1), 1–8. [https://doi.org/https://doi.org/10.1016/S0022-5371\(69\)80002-2](https://doi.org/https://doi.org/10.1016/S0022-5371(69)80002-2)
- Bouten, L. (2011). On the determinants of social and environmental reporting and its role as an accountability mechanism. Ghent University.
- Brown, N., & Deegan, C. (1998). The public disclosure of environmental performance information—a dual test of media agenda setting theory and legitimacy theory. *Accounting and Business Research*, 29(1), 21–41. <https://doi.org/10.1080/00014788.1998.9729564>
- Brundtland, G. H. (1987). Our Common Future. *Oxford Paperbacks*. <https://doi.org/10.2307/633499>
- Brussels Airlines. (2018). CO2 Offsetting. Retrieved from <https://www.brusselsairlines.com/en-be/corporate/corporate-social-responsibility/co2-offsetting.aspx>
- Brysbart, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, 41(4), 977–990. <https://doi.org/10.3758/BRM.41.4.977>
- Brysbart, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*. <https://doi.org/10.3758/s13428-013-0403-5>
- Burgoon, M., & Miller, G. R. (1985). An expectancy interpretation of language and persuasion. In H. Gilles & R. Clair (Eds.), *The social and psychological contexts of language* (pp. 199–229). London: Lawrence Erlbaum Associates.
- Camerer, C., Loewenstein, G., & Weber, M. (1989). The Curse of Knowledge in Economic Settings: An Experimental Analysis. *Journal of Political Economy*, 97(5), 1232–1254. Retrieved from <http://www.jstor.org/stable/1831894>
- Carbon Disclosure Project. (2018). About Us. Retrieved from <https://www.cdp.net/en/info/about-us>
- Castello, E. (2008). *Text Complexity and Reading Comprehension Tests*. Peter Lang. Retrieved from <https://books.google.be/books?id=rYzvuQ5mHUcC>
- Chall, J. S., & Dale, E. (1995). *Readability revisited: the new Dale-Chall readability formula*. Brookline Books. Retrieved from <https://books.google.be/books?id=2nbuAAAAMAAJ>
- Chall, J. (1996). Varying Approaches to Readability Measurement. *Revue Québécoise de Linguistique*, 25(1), 23–40.
- Chang, A. X., & Manning, C. D. (n.d.). Suntime: A library for recognizing and normalizing time expressions.

- Chartprasert, D. (1993). How bureaucratic writing style affects source credibility. *Journalism Quarterly*, 70(1), 150–159.
- Cho, C. H., Michelon, G., & Patten, D. M. (2012). Impression management in sustainability reports: An empirical investigation of the use of graphs. *Accounting and the Public Interest*, 12(1), 16–37. <https://doi.org/10.2308/apin-10249>
- Cho, C. H., Michelon, G., & Patten, D. M. (2012). Enhancement and obfuscation through the use of graphs in sustainability reports: An international comparison. *Sustainability Accounting, Management and Policy Journal*.
- Cho, C. H., & Patten, D. M. (2007). The role of environmental disclosures as tools of legitimacy: A research note. *Accounting, Organizations and Society*, 32(7–8), 639–647. <https://doi.org/10.1016/j.aos.2006.09.009>
- Cho, C. H., Roberts, R. W., & Patten, D. M. (2010). The language of US corporate environmental disclosure. *Accounting, Organizations and Society*. <https://doi.org/10.1016/j.aos.2009.10.002>
- Church, K. W., & Gale, W. A. (1991). Probability scoring for spelling correction. *Statistics and Computing*, 1(2), 93–103. <https://doi.org/10.1007/BF01889984>
- Clark, K., & Manning, C. D. (2016). Improving coreference resolution by learning entity-level distributed representations. *ArXiv Preprint ArXiv:1606.01323*.
- Clatworthy, M. A., & Jones, M. J. (2003). Financial reporting of good and bad news: evidence from accounting narratives., (April 2015), 37–41. <https://doi.org/10.1080/00014788.2003.9729645>
- Collins-Thompson, K. (2014). Computational assessment of text readability: A survey of current and future research. *ITL - International Journal of Applied Linguistics*, 165(2), 97–135. <https://doi.org/http://dx.doi.org/10.1075/itl.165.2.01col>
- Costa, R., & Menichini, T. (2013). A multidimensional approach for CSR assessment: The importance of the stakeholder perception. *Expert Systems with Applications*. <https://doi.org/10.1016/j.eswa.2012.07.028>
- Courtis, J. K. (1995). Readability of annual reports: Western versus Asian evidence. *Accounting, Auditing & Accountability Journal*, 8(2), 4–17. <https://doi.org/10.1108/09513579510086795>
- Courtis, J. K. (1998). Annual report readability variability: tests of the obfuscation hypothesis. *Accounting, Auditing & Accountability Journal*. <https://doi.org/10.1108/09513579810231457>
- Crilly, D., Hansen, M., & Zollo, M. (2016). The grammar of decoupling: A cognitive-linguistic perspective on firms' sustainability claims and stakeholders' interpretation. *Academy of Management Journal*. <https://doi.org/10.5465/amj.2015.0171>
- Dale, E., & Chall, J. S. (1948). A Formula for Predicting Readability: Instructions. *Educational Research Bulletin*, 27(2), 37–54 CR–Copyright © 1948 Taylor & Francis. <https://doi.org/10.2307/1473669>
- Dash, M. (2017). Active and passive voice. Retrieved from <https://plainlanguage.gov/resources/articles/dash-writing-tips/>
- Davies, G., Chun, R., Da Silva, R., & Roper, S. (2005). *Corporate Reputation and Competitiveness*. Taylor & Francis. Retrieved from <https://books.google.be/books?id=OnscBgAAQBAJ>
- Davies, M. (2013). Corpus of Global Web-Based English: 1.9 billion words from speakers in 20 countries (GloWbE). Retrieved from <https://corpus.byu.edu/glowbe/>
- Davies, M. (n.d.). The Corpus of Contemporary American English (COCA): 560 million words, 1990–present. Retrieved from <https://corpus.byu.edu/coca/>
- Davison, A. (1985). *Readability--the situation today*. University of Illinois at Urbana-Champaign. Retrieved from <https://books.google.be/books?id=eEpxysRu5uYC>
- De Clercq, O. (2015). Tipping the scales: exploring the added value of deep semantic processing on readability prediction and sentiment analysis. Ghent University.
- De Clercq, O., & Hoste, V. (2014). Hoe meetbaar is leesbaarheid? In *Beschouwingen uit een talenhuis: opstellen over onderwijs en onderzoek in de vakgroep Vertalen, Tolken en Communicatie aangeboden aan Rita Godyns* (pp. 147–155). Academia Press.

- De Clercq, O., & Hoste, V. (2016). All mixed up?: Finding the optimal feature set for general readability prediction and its application to English and Dutch. *Computational Linguistics*. https://doi.org/10.1162/COLI_a_00255
- De Clercq, O., Hoste, V., Desmet, B., Van Oosten, P., De Cock, M., & Macken, L. (2012). Using the crowd for readability prediction. *Natural Language Engineering*, 20(03), 293–325. <https://doi.org/10.1017/S1351324912000344>
- De Clercq, O., Schulz, S., Desmet, B., & Hoste, V. (2014). Towards Shared Datasets for Normalization Research. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*.
- de Villiers, C., Rinaldi, L., & Unerman, J. (2014). Integrated reporting: Insights, gaps and an agenda for future research. *Accounting, Auditing and Accountability Journal*. <https://doi.org/10.1108/AAAJ-06-2014-1736>
- Deegan, C., Rankin, M., & Tobin, J. (2002). An examination of the corporate social and environmental disclosures of BHP from 1983-1997: A test of legitimacy theory. *Accounting, Auditing & Accountability Journal*. <https://doi.org/10.1108/09513570210435861>
- Dell', Orletta, F., Wieling, M., Cimino, A., Venturi, G., Dell'Orletta, F., & Montemagni, S. (2014). Assessing the Readability of Sentences: Which Corpora and Features? *Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications*, 163–173. Retrieved from <http://www.aclweb.org/anthology/W/W14/W14-1820>
- Desmet, B., Hoste, V., Verstraeten, D., & Verhasselt, J. (n.d.). Gallop documentation.
- Dillard, J. P., & Pfau, M. (2002). *The Persuasion Handbook: Developments in Theory and Practice*. SAGE Publications. Retrieved from https://books.google.be/books?id=I_ByAAQBAJ
- DiMaggio, P., & Powell, W. (1983). The Iron Cage Revisited: Institutional Isomorphism and Collective Rationality in Organizational Fields. *American Sociological Review*, 48(2), 147–160. Retrieved from <http://www.jstor.org/stable/2095101>
- Doak, C. C., Doak, L. G., & Root, J. H. (1996). Teaching Patients with Low Literacy Skills. *AJN The American Journal of Nursing*, 96(12). Retrieved from https://journals.lww.com/ajnonline/Fulltext/1996/12000/Teaching_Patients_with_Low_Literacy_Skills.22.aspx
- Dr. Seuss. (1957). *The Cat in the Hat (Wonderful)*. Harper-Collins.
- DuBay, W. (2004). *The principles of readability*. Costa Mesa: Impact Information. <https://doi.org/10.1.1.91.4042>
- Elkington, J. (1997). Cannibals with forks. *Cannibals with Forks: The Triple Bottom Line of 21st Century*. *The Triple Bottom Line of 21st Century*. <https://doi.org/http://doi.wiley.com/10.1002/tqem.3310080106>
- Eurasian Natural Resources Corporation. (2013). *Unlocking value for sustainable growth. 2012 Sustainable Development Report*.
- ExxonMobil. (2013). *2012 Corporate Citizenship Report*.
- Farewell, S., Fisher, I., & Daily, C. (2014). The Lexical Footprint of Sustainability Reports: A Pilot Study of Readability. In *American Accounting Association Annual Meeting and Conference on Teaching and Learning in Accounting*.
- Flak, A. (2012). South Africa's Amplats fires 12,000 strikers, union leader shot. Retrieved from <https://www.reuters.com/article/us-safrica-strikes/south-africas-amplats-fires-12000-strikers-union-leader-shot-idUSBRE8930W320121005>
- Flesch, R. (1962). *The Art of Readable Writing*. Wiley. Retrieved from <https://books.google.be/books?id=4JMB1WybUvYC>
- Flesch, R. (1962). *The Art of Plain Talk*. Collier Books. Retrieved from <https://books.google.be/books?id=Ku8eAAAAMAAJ>
- Flesch, R. (1981). *How to Write Plain English: A Book for Lawyers and Consumers*. Barnes & Noble. Retrieved from <https://books.google.be/books?id=oKtOAgAACAAJ>
- Flowerdew, J. (2012). *Discourse in English Language Education*. Routledge. Retrieved from <https://books.google.be/books?id=xfjDn-Xi4-oC>

- Freeport-McMoRan. (2013). *Expanding Resources. 2012 Working Towards Sustainable Development Report*.
- Gem Diamonds. (2012). *Our commitment is total. Sustainable Development Report 2012*.
- Gibson, T. R. (1993). *Towards a discourse theory of abstracts and abstracting*. Retrieved from <http://eprints.nottingham.ac.uk/13205/>
- Global Reporting Initiative. (2018). About GRI. Retrieved from <https://www.globalreporting.org/information/about-gri/Pages/default.aspx>
- Global Reporting Initiative. (2013). Principles for Defining Report Quality. Retrieved from <https://g4.globalreporting.org/how-you-should-report/reporting-principles/principles-for-defining-report-quality/Pages/default.aspx>
- Global Reporting Initiative, & International Finance Corporation. (2010). *Getting More Value Out of Sustainability Reporting*. Retrieved from https://www.ifc.org/wps/wcm/connect/8fe80400488658abb6bef66a6515bb18/WB_IFC_GettingMoreValue.pdf?MOD=AJPERES&CACHEID=8fe80400488658abb6bef66a6515bb18
- Grootes, S. (2014). End of South Africa's platinum mine strike signals end of ANC domination. Retrieved from <https://www.theguardian.com/world/2014/jun/25/south-africa-platinum-miners-strike-anc>
- Grupa Lotos S.A. (2013). *The Culture of Values. Integrated Annual Report 2012*.
- Gunning, R. (1968). *The technique of clear writing*. McGraw-Hill. Retrieved from <https://books.google.be/books?id=vJZpAAAAMAAJ>
- Hąbek, P., & Wolniak, R. (2016). Assessing the quality of corporate social responsibility reports: the case of reporting practices in selected European Union member states. *Quality & Quantity: International Journal of Methodology*. <https://doi.org/10.1007/s11135-014-0155-z>
- Haley, A. (2017). It's About Legibility. Retrieved from <https://www.fonts.com/content/learning/fontology/level-4/fine-typography/legibility>
- Halliday, M. A. K. (1989). *Spoken and written language. Language education*. Oxford: Oxford University Press.
- Halliday, M. A. K., & Hasan, R. (2014). *Cohesion in English*. Taylor & Francis. Retrieved from <https://books.google.be/books?id=rAOtAgAAQBAJ>
- Harrison, S., & Bakker, P. (1998). Two new readability predictors for the professional writer: pilot trials. *Journal of Research in Reading*, 21(2), 121–138. <https://doi.org/10.1111/1467-9817.00049>
- Higgins, K. F. (2014). Statement on the Effect of the Recent Court of Appeals Decision on the Conflict Minerals Rule. Retrieved from <https://www.sec.gov/news/public-statement/2014-spch042914kfh>
- Hildebrandt, H. W., & Snyder, R. D. (1981). The Pollyanna Hypothesis in Business Writing: Initial Results, Suggestions for Research. *Journal of Business Communication*. <https://doi.org/10.1177/002194368101800102>
- Hochschild Mining. (2013). *Exploring for Growth. Annual Report & Accounts 2012*.
- Hrasky, S. (2012). Visual disclosure strategies adopted by more and less sustainability-driven companies. *Accounting Forum*, 36(3), 154–165. <https://doi.org/10.1016/j.accfor.2012.02.001>
- Iivonen, K., & Moisander, J. (2014). Rhetorical Construction of Narcissistic CSR Orientation. *Journal of Business Ethics*. <https://doi.org/10.1007/s10551-014-2298-1>
- Infineon. (2013). *The Determining Factor. Infineon Technologies AG Annual Report 2012*. Retrieved from <http://www.infineon.com/dgdl/Infineon-GB2012-E.pdf?fileId=db3a30433b92f0e8013b989bf5cd15f3>
- Institute of Directors in Southern Africa. (2009). *King Code of Governance for South Africa 2009. King Report on Governance for South Africa 2009*. <https://doi.org/10.1177/1524839909332800>

- International Integrated Reporting Council. (2015). The International <IR> Framework. Retrieved from <http://integratedreporting.org/wp-content/uploads/2015/03/13-12-08-THE-INTERNATIONAL-IR-FRAMEWORK-2-1.pdf>
- Jacobson, L. A., Ryan, M., Martin, R. B., Ewen, J., Mostofsky, S. H., Denckla, M. B., & Mahone, E. M. (2011). Working memory influences processing speed and reading fluency in ADHD. *Child Neuropsychology: A Journal on Normal and Abnormal Development in Childhood and Adolescence*. <https://doi.org/10.1080/09297049.2010.532204>
- Jenkins, H., & Yakovleva, N. (2006). Corporate social responsibility in the mining industry: Exploring trends in social and environmental disclosure. *Journal of Cleaner Production*, 14(3), 271–284. <https://doi.org/https://doi.org/10.1016/j.jclepro.2004.10.004>
- Johansson, V. (2008). *Lexical diversity and lexical density in speech and writing: A developmental perspective*. Lund Working Papers in Linguistics (Vol. 53).
- Joyce, J. (1922). *Ulysses*. (D. Kiberd, Ed.) (Annotated). London: Penguin Books, 2000.
- Jurafsky, D., & Martin, J. H. (2014). *Speech and Language Processing*. Pearson. Retrieved from <https://books.google.be/books?id=km-kngEACAAJ>
- Kazakhmys PLC. (2013). *Investment. Development. Growth. Annual Report and Accounts 2012*.
- Kennedy, J. (1995). Debiasing the Curse of Knowledge in Audit Judgment. *The Accounting Review*. <https://doi.org/10.1017/CBO9781107415324.004>
- Kernighan, M. D., Church, K. W., & Gale, W. A. (1990). A Spelling Correction Program Based on a Noisy Channel Model. In *Proceedings of the 13th Conference on Computational Linguistics - Volume 2* (pp. 205–210). Stroudsburg, PA, USA: Association for Computational Linguistics. <https://doi.org/10.3115/997939.997975>
- Kincaid, J. P., Fishburne Jr, R. P., Rogers, R. L., & Chissom, B. S. (1975). *Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel*. Naval Technical Training Command Millington TN Research Branch.
- Kirsch, I. S., Jungeblut, A., Jenkins, L., & Kolstad, A. (1993). *Adult Literacy in America: A First Look at the Results of the National Adult Literacy Survey*. Office of Educational Research and Improvement, U.S. Department of Education. Retrieved from <https://books.google.be/books?id=3P6eAAAAMAAJ>
- Klare, G. R., & Buck, B. (1954). *Know Your Reader: The Scientific Approach to Readability*. Hermitage House. Retrieved from https://books.google.be/books?id=_YEWAAAIAAJ
- Klare, G. R. (1963). *The Measurement of Readability*. Annes, Iowa: Iowa State University Press.
- Kleijn, S. (2018). *Clozing in on readability: How linguistic features affect and predict text comprehension and on-line processing*.
- KPMG. (2017). *The road ahead. The KPMG Survey of Corporate Responsibility Reporting 2017*. Retrieved from <https://assets.kpmg.com/content/dam/kpmg/be/pdf/2017/kpmg-survey-of-corporate-responsibility-reporting-2017.pdf>
- KPMG, CCGA, GRI, & UNEP. (2013). Carrots and Sticks: Sustainability Reporting Policies Worldwide—Today’s Best Practice, Tomorrow’s Trends. KPMG Advisory NV, *Global Reporting Initiative, Centre for Corporate Governance in Africa, United Nations Environment Programme*, 96.
- Kumar, G. (2014). Determinants of Readability of Financial Reports of U.S.-Listed Asian Companies. *Asian Journal of Finance & Accounting* (Vol. 6). <https://doi.org/10.5296/ajfa.v6i2.5695>
- Laufer, W. S. (2003). Social Accountability and Corporate Greenwashing. *Journal of Business Ethics*, 43(3), 253–261. <https://doi.org/10.1023/A:1022962719299>
- Lehavy, R., Li, F., & Merkley, K. (2011). The Effect of Annual Report Readability on Analyst Following and the Properties of Their Earnings Forecasts. *The Accounting Review*, 86(3), 1087–1115. <https://doi.org/10.2308/accr.00000043>
- Leuz, C., Nanda, D., & Wysocki, P. D. (2003). Earnings management and investor protection: An international comparison. *Journal of Financial Economics*, 69(3), 505–527. [https://doi.org/10.1016/S0304-405X\(03\)00121-1](https://doi.org/10.1016/S0304-405X(03)00121-1)

- Li, F. (2008). Annual report readability, current earnings, and earnings persistence. *Journal of Accounting and Economics*, 45(2-3), 221-247. <https://doi.org/10.1016/j.jacceco.2008.02.003>
- Lodhia, S., & Hess, N. (2014). Sustainability accounting and reporting in the mining industry: current literature and directions for future research. *Journal of Cleaner Production*. <https://doi.org/10.1016/j.jclepro.2014.08.094>
- Loh, L., Thomas, T., & Wang, Y. (2017). Sustainability reporting and firm value: Evidence from Singapore-listed companies. *Sustainability (Switzerland)*. <https://doi.org/10.3390/su9112112>
- Lonmin. (2013). *Sustainable Development Report for the Year Ended 30 September 2012*. Retrieved from <http://sd-report.lonmin.com/2012/downloads/lonmin-online-sustainable-development-report-2012.pdf>
- Lu, Y., & Abeysekera, I. (2014). Stakeholders' power, corporate characteristics, and social and environmental disclosure: Evidence from China. *Journal of Cleaner Production*. <https://doi.org/10.1016/j.jclepro.2013.10.005>
- Lybaert, C. (2016). *Moet tussentaal een (grotere) plaats krijgen in lessen Nederlands voor nieuwkomers? Over Taal* (Vol. 55).
- MacMahon. (2013). *HSEQ Report 2012*.
- Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press. Retrieved from <https://books.google.be/books?id=GNvtngEACAAJ>
- Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., & McClosky, D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (pp. 55-60). Retrieved from <http://nlp.stanford.edu/pubs/StanfordCoreNlp2014.pdf>
- Matten, D., & Moon, J. (2008). "Implicit" and "Explicit" CSR: A Conceptual Framework for a Comparative Understanding of Corporate Social Responsibility. *The Academy of Management Review*, 33(2), 404-424. <https://doi.org/10.2307/20159405>
- Mays, E., Damerau, F. J., & Mercer, R. L. (1991). Context Based Spelling Correction. *Inf. Process. Manage.*, 27(5), 517-522. [https://doi.org/10.1016/0306-4573\(91\)90066-U](https://doi.org/10.1016/0306-4573(91)90066-U)
- McLaughlin, G. H. (1969). SMOG Grading - a New Readability Formula. *Journal of Reading*, 12(8), 639-646. Retrieved from http://harrymclaughlin.com/SMOG_Readability_Formula_G._Harry_McLaughlin_%281969%29.pdf
- Microsoft. (2018). Frequently asked questions about grammar proofing in Word. Retrieved from <https://support.microsoft.com/en-us/help/290943/frequently-asked-questions-about-grammar-proofing-in-word>
- MSDN Archive. (2006). Contextual spelling in the 2007 Microsoft Office system. Retrieved from <https://blogs.msdn.microsoft.com/correcteurorthographiqueoffice/2006/06/05/contextual-spelling-in-the-2007-microsoft-office-system/>
- National Assessment of Adult Literacy. (2016). Average prose, document and quantitative literacy scores of adults: 1992 and 2003. Retrieved from https://nces.ed.gov/naal/kf_demographics.asp
- National Association of Insurance Commissioners. (1995). Life and Health Insurance Policy Language Simplification Model Act. Retrieved from <https://www.naic.org/store/free/MDL-575.pdf>
- Navigli, R. (2009). Word Sense Disambiguation: A Survey. *ACM Comput. Surv.*, 41(2), 10:1-10:69. <https://doi.org/10.1145/1459352.1459355>
- Nazari, J. A., Hrazdil, K., & Mahmoudian, F. (2017). Assessing social and environmental performance through narrative complexity in CSR reports. *Journal of Contemporary Accounting and Economics*. <https://doi.org/10.1016/j.jcae.2017.05.002>

- Neu, D., Warsame, H., & Pedwell, K. (1998). Managing Public Impressions: Environmental Disclosures in Annual Reports. *Accounting, Organizations and Society*, 23(3), 265–282. [https://doi.org/10.1016/S0361-3682\(97\)00008-1](https://doi.org/10.1016/S0361-3682(97)00008-1)
- Newcrest Mining. (2013). *Newcrest Mining Limited Sustainability Report 2012*.
- Nishiyama, K., & Johnson, J. V. (1997). Karoshi —Death from Overwork: Occupational Health Consequences of Japanese Production Management. *International Journal of Health Services*, 27(4), 625–641. <https://doi.org/10.2190/1JPC-679V-DYNT-HJ6G>
- Oil & Natural Gas Corporation. (2013). *Leadership Means Responsibility. The Conscience Report*.
- O'Toole, G. (2014). Easy Reading is Hard Writing. Retrieved June 26, 2018, from <https://quoteinvestigator.com/2014/11/05/hard-writing/>
- Ownby, R. L. (2005). Influence of Vocabulary and Sentence Complexity and Passive Voice on the Readability of Consumer-Oriented Mental Health Information on the Internet. *AMIA Annual Symposium Proceedings, 2005*, 585–588. Retrieved from <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1560876/>
- Oz Minerals. (2013). *A Modern Mining Company. Sustainability Report 2012*.
- Paladin Energy Ltd. (2013). *Sustainability Report 2012*.
- PanAust. (2013). *Our People, Community and Landscape. Sustainability Report 2012*.
- Parker, G. (2014). S. Africa Platinu Strike Ends, But Not Its Impact. Retrieved from <https://www.voanews.com/a/platinum-strike-ends-but-not-its-impact/1944791.html>
- Parsons, R., & McKenna, B. J. (2005). Constructing Social Responsibility in Mining Company Reports. In T. Lê & M. Short (Eds.), *Proceedings of the International Conference on Critical Discourse Analysis Theory into Research* (pp. 595–608). Retrieved from [http://195.130.87.21:8080/dspace/bitstream/123456789/262/1/Parsons & McKenna-constructing social responsibility.pdf](http://195.130.87.21:8080/dspace/bitstream/123456789/262/1/Parsons%20&%20McKenna-constructing%20social%20responsibility.pdf)
- Pearson, P. D. (1974). The effects of grammatical complexity on children's comprehension, recall, and conception of certain semantic relations. *Reading Research Quarterly*, 10(2), 155–192. <https://doi.org/10.2307/747180>
- Peterson, C. L., Caverly, D. C., Nicholson, S. A., O'Neal, S., & Cusenbary, S. (2000). Building Reading Proficiency at the Secondary Level: A Guide to Resources. *Independent School*, 152. Retrieved from <http://www.sedl.org/pubs/reading16/8.html>
- Pitler, E., & Nenkova, A. (2008). Revisiting readability: A unified framework for predicting text quality. *Proceedings of the Conference on Empirical ...* <https://doi.org/anthology/D08-1020>
- Plain English Campaign. (2013). How to write in plain English. Retrieved from <http://www.plainenglish.co.uk/files/howto.pdf>
- Platinum, A. (2013). *Sustainable Development Report 2012*.
- Prado-Lorenzo, J.-M., Gallego-Alvarez, I., & Garcia-Sanchez, I. M. (2009). Stakeholder engagement and corporate social responsibility reporting: the ownership structure effect. *Corporate Social Responsibility and Environmental Management*. <https://doi.org/10.1002/csr.189>
- Precht, K. (2003). Stance moods in spoken English: Evidentiality and affect in British and American conversation. *Text - Interdisciplinary Journal for the Study of Discourse*. <https://doi.org/10.1515/text.2003.010>
- Precht, K. (2003). Great vs. lovely: Grammatical and lexical stance differences in American and British English. In C. Meyer & P. Leistyina (Eds.), *Corpus Analysis: Language Structure and Language Use* (pp. 133–151). Amsterdam: Rodopi.
- Premier Oil. (2013). *Growing Value. Corporate Responsibility Report Year to 31 December 2012*.
- Qantas. (2018). Offsetting emissions together. Retrieved from <https://www.qantas.com/travel/airlines/offsetting-emissions-together/global/en>
- Rayson, P., & Garside, R. (2000). Comparing corpora using frequency profiling. In *Proceedings of the workshop on Comparing Corpora* (pp. 1–6). Association for Computational Linguistics.
- Readable.io. (2017). Readable.io. Retrieved from <https://readable.io/>

- Repsol. (2013). 2012 Corporate Responsibility Report. Retrieved from https://www.repsol.com/imagenes/global/en/Corporate_Responsibility_Report_2012_tcm14-22484.pdf
- ROC Oil. (2013). *Sustainability Report 2012*.
- Roland, D., Dick, F., & Elman, J. L. (2007). Frequency of basic English grammatical structures: A corpus analysis. *Journal of Memory and Language*. <https://doi.org/10.1016/j.jml.2007.03.002>
- Rosa, C. (2005). Corporate reputation: Meaning and measurement. *International Journal of Management Reviews*, 7(2), 91–109. <https://doi.org/10.1111/j.1468-2370.2005.00109.x>
- Royal Dutch Shell PLC. (2013). *Royal Dutch Shell PLC Sustainability Report 2012*.
- Rutherford, B. a. (2005). Genre Analysis of Corporate Annual Report Narratives: A Corpus Linguistics-Based Approach. *Journal of Business Communication*. <https://doi.org/10.1177/0021943605279244>
- Rutherford, B. a. (2003). Obfuscation, textual complexity and the role of regulated narrative accounting disclosure in corporate governance. *Journal of Management and Governance*, 7(2), 187–210. <https://doi.org/10.1023/A:1023647615279>
- Sailaja, P. (2012). Indian English: Features and Sociolinguistic Aspects. *Linguistics and Language Compass*. <https://doi.org/10.1002/lnc3.342>
- Salton, G. (1989). *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc.
- Santos. (2013). *Delivering sustainable. Sustainability Report 2012*.
- Saras. (2013). *Environmental, Health and Safety Report 2012*.
- Scandinavian Airlines System. (2018). CO2 Offsets. Retrieved from <https://www.flysas.com/us-en/fly-with-us/travel-extras/co2-offsets/>
- Securities and Exchange Commission. (1998). *A Plain English Handbook. How to create clear SEC disclosure documents*. Retrieved from <https://www.sec.gov/pdf/handbook.pdf>
- Shanahan, T., Fisher, D., & Frey, N. (2012). The Challenge of Challenging Text. *Educational Leadership*.
- Smith, M., & Taffler, R. (1992). The Chairman's Statement and Corporate Financial Performance. *Accounting & Finance*, 32(2), 75–90. <https://doi.org/10.1111/j.1467-629X.1992.tb00187.x>
- Smith, M., & Taffler, R. (1992). Readability and Understandability: Different Measures of the Textual Complexity of Accounting Narrative. *Accounting, Auditing & Accountability Journal*, 5(4), 84. <https://doi.org/10.1108/09513579210019549>
- Smith, M. S., Ogilvia, D. M., Stone, P. J., Dunphy, D. C., & Hartman, J. J. (1967). The General Inquirer: A Computer Approach to Content Analysis. *American Sociological Review*. <https://doi.org/10.2307/2092070>
- Snyder, T. D., & Hoffman, C. M. (1993). *Digest of Educational Statistics, 1993*. Retrieved from <https://nces.ed.gov/pubs93/93292.pdf>
- Socher, R., Perelygin, A., & Wu, J. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. *Proceedings of the* <https://doi.org/10.1371/journal.pone.0073791>
- Stanford NLP Group. (2018). Stanford CoreNLP - Natural Language Software. Retrieved from <https://stanfordnlp.github.io/CoreNLP/#about>
- Stanford NLP Group. (2015). Stanford Parser. Retrieved from <http://nlp.stanford.edu:8080/parser/>
- Stanton, P., & Stanton, J. (2002). Corporate annual reports: research perspectives used. *Accounting, Auditing & Accountability Journal*. <https://doi.org/10.1108/09513570210440568>
- Stolcke, A. (2002). *Srilm --- An Extensible Language Modeling Toolkit. Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP 2002) (Vol. 2)*.
- Strizver, I. (2017). Serif vs. Sans for Text in Print. Retrieved from <https://www.fonts.com/content/learning/fontology/level-1/type-anatomy/serif-vs-sans-for-text-in-print>

- Suchman, M. C. (1995). Managing Legitimacy: Strategic and Institutional Approaches. *Academy of Management Review*, 20(3), 571–610. <https://doi.org/10.5465/AMR.1995.9508080331>
- Taylor, L. (1953). Cloze procedure: A new tool for measuring readability. *Journalism Quarterly*.
- Temouri, Y., & Jones, C. (2014). Introduction: International business and institutions after the financial crisis. *Academy of International Business (Uki)*, 21, 1–6. <https://doi.org/10.1057/9781137367204>
- Tezcan, A. (2018). *Informative Quality Estimation of Machine Translation Output*.
- Thaler, R. H. (1993). *Advances in Behavioral Finance*. Russell Sage Foundation. Retrieved from <https://books.google.be/books?id=kAtba1WxkKkC>
- Thomson Reuters. (2012). ASSET4 Environmental, Social and Corporate Governance Data. Data Collection and Rating Methodology.
- Thomson Reuters. (2013). Thomson Reuters Corporate Responsibility Ratings (TRCRR). Rating and Ranking. Rules and Methodologies. Retrieved from <https://financial.thomsonreuters.com/content/dam/openweb/documents/pdf/tr-com-financial/methodology/corporate-responsibility-ratings.pdf>
- Thomson Reuters. (2013). ASSET4 ESG Data Glossary. Retrieved from http://extranet.datastream.com/data/ASSET4ESG/documents/ASSET4_ESG_DATA_GLOSSARY_april2013.xlsx
- Thomson Reuters. (2018). Global ESG Research. ESG Research Data. Retrieved from <https://financial.thomsonreuters.com/en/products/data-analytics/company-data/esg-research-data.html>
- Tilling, M. V. (2004). Refinements to Legitimacy Theory in Social and Environmental Accounting. *Commerce Research Paper Series*, 06(04), 1–11.
- Total. (2013). *Working Together for Responsible Energy*. Retrieved from <http://www.total.com/sites/default/files/atoms/file/total-society-and-environment-report-2012>
- Toutanova, K., Klein, D., Manning, C. D., & Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1* (pp. 173–180). Association for Computational Linguistics.
- Townsend, S., Bartels, W., & Renaut, J.-P. (2010). Reporting Change. *Change*, 1–33. Retrieved from <http://www.sustainability.com/library>
- Tullow Oil PLC. (2013). *Creating Shared Prosperity. 2012 Corporate Responsibility Report*.
- Ullmann, A. A. (1985). Data in search of a theory: A critical examination of the relationships among social performance, social disclosure, and economic performance of US firms. *Academy of Management Review*, 10(3), 540–557.
- Van de Kauter, M., Breesch, D., & Hoste, V. (2015). Fine-grained analysis of explicit and implicit sentiment in financial news articles. *Expert Systems with Applications*, 42(11), 4999–5010.
- Van de Kauter, M., Desmet, B., & Hoste, V. (2015). The good, the bad and the implicit: a comprehensive approach to annotating explicit and implicit sentiment. *Language Resources and Evaluation*, 49(3), 685–720.
- Van den Bogaerd, M., & Aerts, W. (2011). Applying machine learning in accounting research. *Expert Systems with Applications*, 38(10), 13414–13424. <https://doi.org/https://doi.org/10.1016/j.eswa.2011.04.172>
- Van Hee, C. (2017). *Can machines sense irony?: exploring automatic irony detection on social media*. Ghent University.
- Van Hoecke, S. (2018). *Does Readability of CEO Letters affect the perception of the company?* Ghent University.
- vor der Brück, T., & Hartrumpf, S. (2007). A Semantically Oriented Readability Checker for German. *Proceedings of the 3rd Language & Technology Conference*, (October), 270–274. Retrieved from http://pi7.fernuni-hagen.de/papers/brueck_hartrumpf07_online.pdf

- Vu, T., Aw, A. T., & Zhang, M. (2008). Term Extraction Through Unithood And Termhood Unification. *Proceedings of the Third International Joint Conference on Natural Language Processing (IJCNLP 2008)*.
- Wee, M., Tarca, A., Krug, L., Aerts, W., Pink, P., & Tilling, M. V. (2015). *Factors Affecting Preparers' and Auditors' Judgements about Materiality and Conciseness in Integrated Reporting*. ACCA.
- Wegner, M. V., & Girasek, D. C. (2003). How readable are child safety seat installation instructions? *Pediatrics*. <https://doi.org/http://dx.doi.org/10.1542/peds.111.3.588>
- Wink, K. A. (2016). *Rhetorical Strategies for Composition: Cracking an Academic Code*. Rowman & Littlefield Publishers. Retrieved from <https://books.google.be/books?id=7TpgCwAAQBAJ>
- Wyk, D. (2012). *Policy Gap 6: A Review of Platinum Mining in the Bojanala District of the North West Province: A Participatory Action Research (PAR) Approach*.
- Yamaguchi, M. (2016). Japan overwork deaths among young show lessons unlearned. Retrieved from <https://apnews.com/b972d8ccc27a4060a8a3385f7b11ed41/japan-overwork-deaths-among-young-show-lessons-unlearned>

Appendix

Appendix 1: Corpus Description & Additional Analyses

Companies in Corpus

Company	Annual report LtS	Sustain- ability LtS	Sustain- ability Report	Total Files	Tokens after Cleaning
Australia	64	13	131	116	379523
ADITYA BIRLA	1	0	0	1	820
ALUMINA	1	0	1	2	3022
AMPELLA	1	0	0	1	778
ANTARES ENERGY	1	0	0	1	643
ARAFURA	1	0	0	1	1119
ARRIUM	1	1	1	3	18143
ATLAS IRON	1	0	1	2	3163
AURORA OIL And GAS	1	0	0	1	1241
AUSDRILL	1	0	0	1	1021
AWE	1	1	1	3	8647
BEACH ENERGY	1	0	1	2	5577
BERKELEY	0	0	1	1	676
BLUESCOPE STEEL	1	0	0	1	1577
CALTEX AUSTRALIA	1	0	1	2	3133
CAPE	1	0	0	1	573
CARNARVON	1	0	0	1	1293
CENTAMIN	1	0	1	2	6378
COOPER ENERGY	1	0	0	1	911
DART ENERGY	1	0	0	1	1571
DISCOVERY METALS	1	0	1	2	7406
ENERGY RES	1	0	1	2	7577
EVOLUTION	1	0	1	2	2226
FLINDERS	1	0	0	1	1787
FORTECUE	1	0	1	2	12329
GALAXY RESOURCES	1	0	0	1	598
GINDALBIE METALS	1	0	1	2	6116
GRANGE RESOURCES	1	0	0	1	1430
GRYPHON MINERALS	1	0	1	2	2960
ILUKA RESOURCES	1	0	1	2	9167
INDEPENDENCE	1	0	0	1	1112
INDOPHIL	1	0	1	2	2740
KAROON GAS	1	0	0	1	1061
LYNAS	1	0	1	2	3154
MACMAHON	1	1	1	3	9361
MEDUSA MINING	1	0	0	1	928
MEO AUSTRALIA	1	0	0	1	2343
MINCOR RESOURCES	1	0	1	2	1995
MINERAL	1	0	0	1	998

MOLOPO	1	0	1	2	2194
MOUNT GIBSON	1	0	1	1	3451
MOUNT GIBSON	1	0	1	1	1602
NEWCREST MINING	1	1	1	3	29021
NEXUS	1	0	1	2	1917
NIDO	1	0	1	2	1309
NORTHERN IRON	1	0	0	1	811
OIL SEARCH	1	1	1	3	16907
OZ MINERALS	1	1	1	3	18063
PALADIN ENERGY	1	1	1	3	16968
PANAUST	1	1	1	3	37657
PANORAMIC	1	1	1	3	9703
PERILYA	1	0	1	2	2265
PERSEUS MINING	1	0	1	2	7926
REGIS	1	0	0	1	584
RESOLUTE MINING	1	0	1	2	4856
ROC OIL COMPANY	1	1	1	3	7360
SANDFIRE	1	0	1	2	2480
SANTOS	1	1	1	3	19113
SILEX	1	0	0	1	982
SILVER LAKE	1	0	0	1	1050
SIMS	1	0	1	2	13637
ST BARBARA	1	0	1	2	2098
SUNDANCE	1	1	1	3	11504
TAP OIL	1	0	1	2	1638
WESTERN AREAS	1	0	0	1	1865
WOODSIDE	1	1	1	3	26988
Europe	45	20	45	108	814445
ACERINOX	1	1	1	3	21719
ADIDAS	1	1	1	3	31730
AIXTRON	1	0	0	1	1812
ALLIANCE OIL	1	0	1	2	4542
ARCELORMITTAL	1	1	1	3	28518
ASML HOLDING	0	1	1	2	24351
AURUBIS	1	0	1	2	2519
BENETTON	0	0	1	1	3493
BOLIDEN	1	0	1	2	15283
DRAGON	1	0	1	1	3585
DRAGON OIL	1	0	1	1	2846
ENI	1	0	1	2	18110
ERAMET	1	0	1	2	3912
GALP	1	1	1	3	33296
GEMALTO	1	1	1	3	25468
HELLENIC	1	1	1	3	22350

HERMES	1	0	1	2	8324
HOGANAS	1	0	1	2	2692
INDITEX	1	0	1	2	22686
INFINEON	1	0	1	2	8753
KGHM	1	1	1	3	44542
KLOECKNER And CO	1	0	1	2	3982
LUNDIN PETROLEUM	1	0	1	2	5227
LUXOTTICA	1	0	0	1	600
LVMH	1	0	1	2	10610
MAUREL	1	0	1	2	8960
MICRONAS	1	0	0	1	2343
MOL	1	1	1	3	23763
MOTOR OIL	0	1	1	2	38325
NESTE OIL	1	0	1	2	37250
OMV	1	1	1	3	26704
OUTOKUMPU	1	1	1	3	55829
OUTOTEC	1	1	1	3	23546
POLISH	1	0	1	2	3203
PUMA	1	0	1	2	21994
RAUTARUUKKI	1	1	1	3	22284
REPSOL	1	0	1	2	1407
RICHEMONT	1	0	1	2	23910
SALZGITTER	1	0	1	2	5365
SARAS	1	1	1	3	30513
SSAB	1	1	1	3	20573
STMICROELECTRONICS	1	1	1	3	31809
TALVIVAARA	1	1	1	3	23718
THE SWATCH	1	0	1	2	12531
TOTAL	1	1	1	3	14752
UMICORE	1	0	1	2	11382
VALLOUREC	1	1	1	3	21206
VOESTALPINE	1	0	1	2	2128
India	8	4	8	20	101727
CAIRN INDIA	1	0	1	2	3837
JINDAL	1	1	1	3	4907
JSW	1	1	1	3	14788
OIL And NATURAL GAS	1	1	1	3	26290
RELIANCE INDUSTRIES	1	0	1	2	5739
STEEL AUTHORITY	1	1	1	3	26772
STERLITE	1	0	1	2	7681
TATA	1	0	1	2	11713
UK	30	22	30	82	655990
ANGLO AMERICAN	1	1	1	3	42517
ANTOFAGASTA	1	1	1	3	23976

AQUARIUS	1	1	1	3	14456
ARM HOLDINGS	1	1	1	3	10429
BG GROUP	1	1	1	3	41482
BHP BILLITON	1	1	1	3	22835
BODYCOTE	1	0	1	2	3911
BP	1	1	1	3	32947
CAIRN ENERGY	1	1	1	3	28294
CSR	1	0	1	2	9016
ESSAR ENERGY	1	1	1	3	22880
EURASIAN	1	1	1	3	18117
FERREXPO	1	0	1	2	11904
FRESNILLO	1	0	1	2	5324
GEM	1	1	1	3	23530
HOCHSCHILD	1	1	1	3	6126
JKX OIL And GAS	1	0	1	2	6453
JOHNSON MATTHEY	1	0	1	2	17234
KAZ MINERALS	1	1	1	3	4252
LONMIN	1	1	1	3	70983
PETROPAVLOVSK	1	1	1	3	10079
PREMIER OIL	1	1	1	3	20250
RANDGOLD	1	1	1	3	30579
RIO TINTO	1	1	1	3	38676
ROYAL DUTCH	1	1	1	3	33129
SALAMANDER	1	0	1	2	4779
SOCO	1	0	1	2	6952
TULLOW OIL	1	1	1	3	33276
VEDANTA	1	1	1	3	34963
XSTRATA	1	1	1	3	26641
USA	71	28	39	138	696043
ADVANCED	1	1	1	3	33481
AK STEEL	1	0	1	2	4944
ALCOA	1	0	1	2	8297
ALLEGHENY	1	0	1	2	4670
ALTERA	1	0	0	1	1394
ANADARKO	1	0	1	2	4625
ANALOG DEVICES	1	0	0	1	1171
APACHE	1	1	1	3	11761
APPLIED MAT	1	0	0	1	1520
BROADCOM	1	1	1	3	3171
CABOT OIL And GAS	1	1	1	3	5368
CHESAPEAKE	1	1	1	3	11456
CHEVRON	1	1	1	3	22294
CIMAREX	1	0	1	2	1238
CLIFFS	0	1	1	2	21804

COEUR MINING	1	0	0	1	1922
COMMERCIAL	1	0	1	2	1542
COMSTOCK	1	0	0	1	1985
CONCHO	1	0	0	1	626
CONOCO	1	1	1	3	50646
CONTINENTAL	1	0	0	1	1916
CREE	1	0	0	1	796
CYPRESS	1	0	0	1	6384
DENBURY	1	0	1	2	2122
DEVON ENERGY	1	1	1	3	12132
ENERGEN	1	0	0	1	146
EOG	1	0	0	1	1215
EQT	0	1	1	2	14169
EXCO RESOURCES	1	0	0	1	987
EXXON MOBIL	1	1	1	3	38188
FLEXTRONICS	1	0	0	1	1319
FOREST OIL	1	0	0	1	869
FREEPOR-T-MCMORAN	1	1	1	3	16555
HESS	1	1	1	3	29453
HOLLYFRONTIER	1	0	0	1	1026
INTEGRATED	1	0	0	1	1822
INTEL	1	1	1	3	58886
INTERSIL	1	0	0	1	278
JABIL CIRCUIT	1	0	1	2	2230
JONES GROUP	1	0	0	1	1368
KLA	1	0	0	1	465
LAM RESEARCH	1	0	1	2	3407
LINEAR	1	0	0	1	1789
LSI	0	1	1	2	12644
MARATHON OIL	1	1	1	3	26564
MARVELL	1	0	0	1	773
MICROCHIP	0	1	1	2	8332
MURPHY OIL	1	0	0	1	1755
NEWFIELD	1	0	0	1	1629
NEWMONT MINING	1	1	1	3	80844
NIKE	1	1	1	3	37813
NOBLE ENERGY	1	1	1	3	17287
NVIDIA	1	1	1	3	17192
OCCIDENTAL	1	1	1	3	13181
ON	1	0	0	1	1008
PIONEER	1	0	0	1	2478
PVH	1	1	1	3	19903
QEP RESOURCES	1	0	0	1	1515
QUESTAR	1	1	1	3	9216

QUICKSILVER	1	0	0	1	1158
RANGE	1	0	0	1	484
RELIANCE STEEL	1	0	0	1	1248
ROYAL GOLD	1	0	1	2	1847
SCHNITZER	1	0	1	2	2338
SKYWORKS	1	0	0	1	1215
SM ENERGY	1	0	0	1	1744
SOUTHERN COPPER	0	1	0	1	897
SOUTHWESTERN	1	0	0	1	917
SUNEDISON	0	1	1	2	12769
TERADYNE	1	0	0	1	787
TESORO	0	1	1	2	8767
TEXAS	1	1	1	3	6063
ULTRA	1	0	0	1	2324
UNITED STATES	1	0	0	1	1537
V F	1	0	0	1	1459
VALERO ENERGY	1	0	1	2	5044
WHITING	1	0	0	1	562
WILLIAMS	1	0	0	1	1312
Grand Total	0	0	0	464	2647728

CSR Reporting Requirements per Country (year 2012)

Synthesised from KPMG, et al. 2013

Region	Requirement	How mandatory?
European Union	EU Modernisation Directive, 2003 requires inclusion of "non-financial information in [...] annual and consolidated reports", provided "it is necessary for an understanding of the company's development, performance or position".	Room for interpretation in what is 'necessary'; Member States can exempt small and medium-sized companies
Australia	Corporations Act - Sect. 229, 2001 demands that those companies that issue an Annual Directors' Report indicate environmental regulations to which they are subject, and their performance relative to those regulations. This later evolved to include the company's "financial position [...] and prospects".	Mandatory if subject to regulations, which depends on state and national laws.
	The ASX Listing Rules regarding Corporate Governance Council Principles and Recommendations, 2010 oblige listed companies' annual reports to "disclos[e] the extent to which they have followed the Corporate Governance Council's Principles and Recommendations".	Companies are only obliged to disclose their deviations from these Principles and Recommendations.
Belgium	The Social Balance Sheet requires companies that employ staff to report on "the nature and the evolution of employment, e.g. training" as part of their annual accounts.	Mandatory.
Finland	The Finnish Accounting Act, 1997 requires the directors' report section of the financial or annual report to "defin[e] the key ratios necessary to understand operations and financial position [as well as those] on personnel and environmental factors, and other potentially significant matters impacting on the operations of the reporting entity."	Room for interpretation in what is 'necessary', similar to EU Modernisation Directive.
France	The Grenelle Act II, 2010 obliges a company's annual report describe it and its subsidiaries' environmental and social performance from 2012 onwards, and have this information verified by a third party, "effectively turning it into the foundation for a full integrated report".	Not yet mandatory in 2012 for firms with fewer than 500 employees and "total assets or net annual sales [below] €100 million".
Germany	German Accounting Standard No. 15 stipulates "clarity and transparency" and "focus on sustainable value creation" as two of five principles for management reports.	Subject to interpretation.

India	The 2012 requirement to submit Business Responsibility Reports obliges the companies with the 100 companies with the greatest market capitalisations to report along "the key principles enunciated in the 'National Voluntary Guidelines on Social, Environmental and Economic Responsibilities of Business'".	Subject to 'comply or explain'.
Netherlands	In accordance with the EU Modernisation Directive, the Dutch Civil Code stipulates that where necessary an understanding of an organisation's "development, performance or position", provide financial and non-financial performance information (the later including environmental and social issues as well as risks) in their annual reports.	Mandatory for all listed companies and large non-listed companies; 'necessary' however subject to interpretation.
Spain	The Spanish Sustainable Economy Law, 2011 mandates that listed companies publish annual corporate governance reports according to an official template, in addition to reporting on remuneration policy. The law also encourages disclosure of CSR policy, performance and assurance.	This is not "a strong obligation"; the CSR reporting policy was unenforceable at time of writing due to a lack of guidelines on how or where to submit these reports.
Sweden	The Annual Accounts Act, 1999 obliges certain companies to disclose "environmental and social information in the Board of Directors' Report section of the annual report", with increased requirements based on since 2005 based on the EU Accounting Modernisation Directive.	Mandatory.
UK	Similar to the EU Accounting Modernisation Directive, the Companies Act, 2006 mandates that quote companies include in their annual review "information on environmental, employee, social and community matters to the extent necessary for an understanding of the development, performance or position of the company".	Room for interpretation in terms of what is 'necessary'.

Post Hoc Analyses for Full Corpus

Sustainability Reports – Readability Formulae

Estimates

Dependent Variable: Flesch Reading Ease Index

region	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Australia	23.431 ^a	1.396	20.671	26.192
Europe	29.850 ^a	1.260	27.360	32.340
India	27.746 ^a	2.663	22.481	33.011
UK	29.182 ^a	1.410	26.393	31.970
USA	26.433 ^a	1.181	24.098	28.768

a. Covariates appearing in the model are evaluated at the following values: env_perf = 65.21, soc_perf = 68.53, gov_perf = 69.65, ecn_perf = 60.60, total_assets = 24198506.95.

Pairwise Comparisons

Dependent Variable: Flesch Reading Ease Index

(I) region		Mean Difference (I-J)	Std. Error	Sig. ^b	95% Confidence Interval for Difference ^b	
					Lower Bound	Upper Bound
Australia	Europe	-6.419*	1.889	0.009	-11.808	-1.030
	India	-4.315	2.960	1.000	-12.758	4.128
	UK	-5.751*	1.680	0.008	-10.541	-0.960
	USA	-3.002	1.721	0.833	-7.910	1.906
Europe	Australia	6.419*	1.889	0.009	1.030	11.808
	India	2.104	2.497	1.000	-5.018	9.227
	UK	0.668	1.847	1.000	-4.599	5.936
	USA	3.417	1.728	0.499	-1.511	8.345
India	Australia	4.315	2.960	1.000	-4.128	12.758
	Europe	-2.104	2.497	1.000	-9.227	5.018
	UK	-1.436	2.929	1.000	-9.789	6.917
	USA	1.313	2.907	1.000	-6.979	9.604
UK	Australia	5.751*	1.680	0.008	0.960	10.541
	Europe	-0.668	1.847	1.000	-5.936	4.599
	India	1.436	2.929	1.000	-6.917	9.789
	USA	2.749	1.680	1.000	-2.044	7.542
USA	Australia	3.002	1.721	0.833	-1.906	7.910
	Europe	-3.417	1.728	0.499	-8.345	1.511
	India	-1.313	2.907	1.000	-9.604	6.979
	UK	-2.749	1.680	1.000	-7.542	2.044

Based on estimated marginal means

*. The mean difference is significant at the .05 level.

b. Adjustment for multiple comparisons: Bonferroni.

Estimates

Dependent Variable: Flesch Reading Ease Index

industry	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Apparel	30.267 ^a	2.296	25.727	34.806
Mining	26.926 ^a	0.821	25.303	28.550
Oil	27.086 ^a	0.967	25.175	28.998
Semiconductors	25.034 ^a	1.808	21.461	28.608

a. Covariates appearing in the model are evaluated at the following values: env_perf = 65.21, soc_perf = 68.53, gov_perf = 69.65, ecn_perf = 60.60, total_assets = 24198506.95.

Pairwise Comparisons

Dependent Variable: Flesch Reading Ease Index

(I) industry		Mean Difference (I-J)	Std. Error	Sig. ^a	95% Confidence Interval for Difference ^a	
					Lower Bound	Upper Bound
Apparel	Mining	3.340	2.363	0.959	-2.985	9.665
	Oil	3.181	2.395	1.000	-3.229	9.590
	Semiconductors	5.233	2.750	0.355	-2.127	12.592
Mining	Apparel	-3.340	2.363	0.959	-9.665	2.985
	Oil	-0.160	1.219	1.000	-3.422	3.103
	Semiconductors	1.892	1.946	1.000	-3.315	7.100
Oil	Apparel	-3.181	2.395	1.000	-9.590	3.229
	Mining	0.160	1.219	1.000	-3.103	3.422
	Semiconductors	2.052	1.995	1.000	-3.288	7.392
Semiconductors	Apparel	-5.233	2.750	0.355	-12.592	2.127
	Mining	-1.892	1.946	1.000	-7.100	3.315
	Oil	-2.052	1.995	1.000	-7.392	3.288

Based on estimated marginal means

a. Adjustment for multiple comparisons: Bonferroni.

Estimates

Dependent Variable: Flesch-Kincaid Grade Level

region	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Australia	15.901 ^a	0.319	15.272	16.531
Europe	14.922 ^a	0.287	14.354	15.490
India	14.811 ^a	0.607	13.610	16.012
UK	14.960 ^a	0.322	14.324	15.597
USA	15.143 ^a	0.269	14.610	15.675

a. Covariates appearing in the model are evaluated at the following values: env_perf = 65.21, soc_perf = 68.53, gov_perf = 69.65, ecn_perf = 60.60, total_assets = 24198506.95.

Pairwise Comparisons

Dependent Variable: Flesch-Kincaid Grade Level

(I) region		Mean Difference (I-J)	Std. Error	Sig. ^a	95% Confidence Interval for Difference ^a	
					Lower Bound	Upper Bound
Australia	Europe	0.979	0.431	0.247	-0.250	2.208
	India	1.090	0.675	1.000	-0.836	3.016
	UK	0.941	0.383	0.153	-0.152	2.034
	USA	0.759	0.393	0.553	-0.361	1.878
Europe	Australia	-0.979	0.431	0.247	-2.208	0.250
	India	0.111	0.570	1.000	-1.513	1.736
	UK	-0.038	0.421	1.000	-1.240	1.164
	USA	-0.220	0.394	1.000	-1.344	0.904
India	Australia	-1.090	0.675	1.000	-3.016	0.836
	Europe	-0.111	0.570	1.000	-1.736	1.513
	UK	-0.149	0.668	1.000	-2.055	1.756
	USA	-0.332	0.663	1.000	-2.223	1.560
UK	Australia	-0.941	0.383	0.153	-2.034	0.152
	Europe	0.038	0.421	1.000	-1.164	1.240
	India	0.149	0.668	1.000	-1.756	2.055
	USA	-0.182	0.383	1.000	-1.276	0.911
USA	Australia	-0.759	0.393	0.553	-1.878	0.361
	Europe	0.220	0.394	1.000	-0.904	1.344
	India	0.332	0.663	1.000	-1.560	2.223
	UK	0.182	0.383	1.000	-0.911	1.276

Based on estimated marginal means

a. Adjustment for multiple comparisons: Bonferroni.

Estimates

Dependent Variable: Flesch-Kincaid Grade Level

industry	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Apparel	14.858 ^a	0.524	13.822	15.893
Mining	15.113 ^a	0.187	14.743	15.483
Oil	15.281 ^a	0.221	14.845	15.717
Semiconductors	15.338 ^a	0.412	14.523	16.153

a. Covariates appearing in the model are evaluated at the following values: env_perf = 65.21, soc_perf = 68.53, gov_perf = 69.65, ecn_perf = 60.60, total_assets = 24198506.95.

Pairwise Comparisons

Dependent Variable: Flesch-Kincaid Grade Level

(I) industry		Mean Difference (I-J)	Std. Error	Sig. ^a	95% Confidence Interval for Difference ^a	
					Lower Bound	Upper Bound
Apparel	Mining	-0.255	0.539	1.000	-1.698	1.188
	Oil	-0.423	0.546	1.000	-1.885	1.039
	Semiconductors	-0.480	0.627	1.000	-2.159	1.198
Mining	Apparel	0.255	0.539	1.000	-1.188	1.698
	Oil	-0.168	0.278	1.000	-0.913	0.576
	Semiconductors	-0.225	0.444	1.000	-1.413	0.962
Oil	Apparel	0.423	0.546	1.000	-1.039	1.885
	Mining	0.168	0.278	1.000	-0.576	0.913
	Semiconductors	-0.057	0.455	1.000	-1.275	1.161
Semiconductors	Apparel	0.480	0.627	1.000	-1.198	2.159
	Mining	0.225	0.444	1.000	-0.962	1.413
	Oil	0.057	0.455	1.000	-1.161	1.275

Based on estimated marginal means

a. Adjustment for multiple comparisons: Bonferroni.

Estimates

Dependent Variable: Gunning Fog Index

region	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Australia	19.783 ^a	0.361	19.069	20.497
Europe	18.806 ^a	0.326	18.162	19.450
India	18.560 ^a	0.688	17.199	19.921
UK	18.731 ^a	0.365	18.011	19.452
USA	18.802 ^a	0.305	18.198	19.406

a. Covariates appearing in the model are evaluated at the following values: env_perf = 65.21, soc_perf = 68.53, gov_perf = 69.65, ecn_perf = 60.60, total_assets = 24198506.95.

Pairwise Comparisons

Dependent Variable: Gunning Fog Index

(I) region		Mean Difference (I-J)	Std. Error	Sig. ^a	95% Confidence Interval for Difference ^a	
					Lower Bound	Upper Bound
Australia	Europe	0.977	0.488	0.474	-0.416	2.370
	India	1.223	0.765	1.000	-0.960	3.406
	UK	1.052	0.434	0.167	-0.187	2.290
	USA	0.981	0.445	0.291	-0.288	2.250
Europe	Australia	-0.977	0.488	0.474	-2.370	0.416
	India	0.246	0.646	1.000	-1.595	2.087
	UK	0.075	0.477	1.000	-1.287	1.437
	USA	0.004	0.447	1.000	-1.270	1.278
India	Australia	-1.223	0.765	1.000	-3.406	0.960
	Europe	-0.246	0.646	1.000	-2.087	1.595
	UK	-0.171	0.757	1.000	-2.331	1.988
	USA	-0.242	0.752	1.000	-2.386	1.902
UK	Australia	-1.052	0.434	0.167	-2.290	0.187
	Europe	-0.075	0.477	1.000	-1.437	1.287
	India	0.171	0.757	1.000	-1.988	2.331
	USA	-0.071	0.434	1.000	-1.310	1.169
USA	Australia	-0.981	0.445	0.291	-2.250	0.288
	Europe	-0.004	0.447	1.000	-1.278	1.270
	India	0.242	0.752	1.000	-1.902	2.386
	UK	0.071	0.434	1.000	-1.169	1.310

Based on estimated marginal means

a. Adjustment for multiple comparisons: Bonferroni.

Estimates

Dependent Variable: Gunning Fog Index

industry	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Apparel	18.449 ^a	0.594	17.276	19.623
Mining	18.992 ^a	0.212	18.572	19.411
Oil	19.171 ^a	0.250	18.676	19.665
Semiconductors	19.135 ^a	0.467	18.211	20.059

a. Covariates appearing in the model are evaluated at the following values: env_perf = 65.21, soc_perf = 68.53, gov_perf = 69.65, ecn_perf = 60.60, total_assets = 24198506.95.

Pairwise Comparisons

Dependent Variable: Gunning Fog Index

(I) industry		Mean Difference (I-J)	Std. Error	Sig. ^a	95% Confidence Interval for Difference ^a	
					Lower Bound	Upper Bound
Apparel	Mining	-0.542	0.611	1.000	-2.177	1.093
	Oil	-0.721	0.619	1.000	-2.378	0.936
	Semiconductors	-0.686	0.711	1.000	-2.588	1.217
Mining	Apparel	0.542	0.611	1.000	-1.093	2.177
	Oil	-0.179	0.315	1.000	-1.022	0.664
	Semiconductors	-0.143	0.503	1.000	-1.490	1.203
Oil	Apparel	0.721	0.619	1.000	-0.936	2.378
	Mining	0.179	0.315	1.000	-0.664	1.022
	Semiconductors	0.036	0.516	1.000	-1.345	1.416
Semiconductors	Apparel	0.686	0.711	1.000	-1.217	2.588
	Mining	0.143	0.503	1.000	-1.203	1.490
	Oil	-0.036	0.516	1.000	-1.416	1.345

Based on estimated marginal means

a. Adjustment for multiple comparisons: Bonferroni.

Sustainability Reports – Lexicosyntactic Features

Estimates

Dependent Variable:

Lexical Density

region	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Australia	.646 ^a	0.004	0.638	0.654
Europe	.633 ^a	0.004	0.625	0.640
India	.660 ^a	0.008	0.645	0.676
UK	.627 ^a	0.004	0.619	0.635
USA	.652 ^a	0.003	0.645	0.659

a. Covariates appearing in the model are evaluated at the following values: env_perf = 65.21, soc_perf = 68.53, gov_perf = 69.65, ecn_perf = 60.60, total_assets = 24198506.95.

Pairwise Comparisons

Dependent Variable:

Lexical Density

(I) region		Mean Difference (I-J)	Std. Error	Sig. ^b	95% Confidence Interval for Difference ^b	
					Lower Bound	Upper Bound
Australia	Europe	0.013	0.006	0.193	-0.003	0.029
	India	-0.015	0.009	0.964	-0.039	0.010
	UK	.019*	0.005	0.002	0.005	0.033
	USA	-0.006	0.005	1.000	-0.021	0.008
Europe	Australia	-0.013	0.006	0.193	-0.029	0.003
	India	-.028*	0.007	0.002	-0.049	-0.007
	UK	0.006	0.005	1.000	-0.010	0.021
	USA	-.019*	0.005	0.002	-0.034	-0.005
India	Australia	0.015	0.009	0.964	-0.010	0.039
	Europe	.028*	0.007	0.002	0.007	0.049
	UK	.033*	0.009	0.002	0.009	0.058
	USA	0.008	0.009	1.000	-0.016	0.033
UK	Australia	-.019*	0.005	0.002	-0.033	-0.005
	Europe	-0.006	0.005	1.000	-0.021	0.010
	India	-.033*	0.009	0.002	-0.058	-0.009
	USA	-.025*	0.005	0.000	-0.039	-0.011
USA	Australia	0.006	0.005	1.000	-0.008	0.021
	Europe	.019*	0.005	0.002	0.005	0.034
	India	-0.008	0.009	1.000	-0.033	0.016
	UK	.025*	0.005	0.000	0.011	0.039

Based on estimated marginal means

*. The mean difference is significant at the .05 level.

b. Adjustment for multiple comparisons: Bonferroni.

Estimates

Dependent Variable:

Lexical Density

industry	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Apparel	.636 ^a	0.007	0.623	0.649
Mining	.648 ^a	0.002	0.644	0.653
Oil	.646 ^a	0.003	0.640	0.652
Semiconductors	.644 ^a	0.005	0.634	0.655

a. Covariates appearing in the model are evaluated at the following values: env_perf = 65.21, soc_perf = 68.53, gov_perf = 69.65, ecn_perf = 60.60, total_assets = 24198506.95.

Pairwise Comparisons

Dependent Variable:

Lexical Density

(I) industry		Mean Difference (I-J)	Std. Error	Sig. ^a	95% Confidence Interval for Difference ^a	
					Lower Bound	Upper Bound
Apparel	Mining	-0.013	0.007	0.431	-0.031	0.006
	Oil	-0.010	0.007	0.912	-0.029	0.009
	Semiconductors	-0.008	0.008	1.000	-0.030	0.013
Mining	Apparel	0.013	0.007	0.431	-0.006	0.031
	Oil	0.002	0.004	1.000	-0.007	0.012
	Semiconductors	0.004	0.006	1.000	-0.011	0.020
Oil	Apparel	0.010	0.007	0.912	-0.009	0.029
	Mining	-0.002	0.004	1.000	-0.012	0.007
	Semiconductors	0.002	0.006	1.000	-0.014	0.018
Semiconductors	Apparel	0.008	0.008	1.000	-0.013	0.030
	Mining	-0.004	0.006	1.000	-0.020	0.011
	Oil	-0.002	0.006	1.000	-0.018	0.014

Based on estimated marginal means

a. Adjustment for multiple comparisons: Bonferroni.

Estimates

Dependent Variable:

Parse Tree Depth

region	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Australia	10.699 ^a	0.176	10.350	11.047
Europe	10.397 ^a	0.159	10.083	10.711
India	10.016 ^a	0.336	9.351	10.681
UK	10.647 ^a	0.178	10.295	10.999
USA	10.172 ^a	0.149	9.877	10.466

a. Covariates appearing in the model are evaluated at the following values: env_perf = 65.21, soc_perf = 68.53, gov_perf = 69.65, ecn_perf = 60.60, total_assets = 24198506.95.

Pairwise Comparisons

Dependent Variable:

Parse Tree Depth

(I) region		Mean Difference (I-J)	Std. Error	Sig. ^a	95% Confidence Interval for Difference ^a	
					Lower Bound	Upper Bound
Australia	Europe	0.302	0.239	1.000	-0.378	0.982
	India	0.683	0.374	0.697	-0.383	1.749
	UK	0.052	0.212	1.000	-0.553	0.657
	USA	0.527	0.217	0.165	-0.092	1.147
Europe	Australia	-0.302	0.239	1.000	-0.982	0.378
	India	0.381	0.315	1.000	-0.518	1.280
	UK	-0.250	0.233	1.000	-0.915	0.415
	USA	0.226	0.218	1.000	-0.397	0.848
India	Australia	-0.683	0.374	0.697	-1.749	0.383
	Europe	-0.381	0.315	1.000	-1.280	0.518
	UK	-0.631	0.370	0.901	-1.685	0.424
	USA	-0.156	0.367	1.000	-1.202	0.891
UK	Australia	-0.052	0.212	1.000	-0.657	0.553
	Europe	0.250	0.233	1.000	-0.415	0.915
	India	0.631	0.370	0.901	-0.424	1.685
	USA	0.475	0.212	0.266	-0.130	1.080
USA	Australia	-0.527	0.217	0.165	-1.147	0.092
	Europe	-0.226	0.218	1.000	-0.848	0.397
	India	0.156	0.367	1.000	-0.891	1.202
	UK	-0.475	0.212	0.266	-1.080	0.130

Based on estimated marginal means

a. Adjustment for multiple comparisons: Bonferroni.

Estimates

Dependent Variable: Parse Tree Depth

industry	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Apparel	10.521 ^a	0.290	9.948	11.094
Mining	10.206 ^a	0.104	10.001	10.410
Oil	10.427 ^a	0.122	10.185	10.668
Semiconductors	10.391 ^a	0.228	9.940	10.842

a. Covariates appearing in the model are evaluated at the following values: env_perf = 65.21, soc_perf = 68.53, gov_perf = 69.65, ecn_perf = 60.60, total_assets = 24198506.95.

Pairwise Comparisons

Dependent Variable: Parse Tree Depth

(I) industry		Mean Difference (I-J)	Std. Error	Sig. ^a	95% Confidence Interval for Difference ^a	
					Lower Bound	Upper Bound
Apparel	Mining	0.316	0.298	1.000	-0.482	1.114
	Oil	0.095	0.302	1.000	-0.714	0.904
	Semiconductors	0.131	0.347	1.000	-0.798	1.060
Mining	Apparel	-0.316	0.298	1.000	-1.114	0.482
	Oil	-0.221	0.154	0.917	-0.633	0.191
	Semiconductors	-0.185	0.246	1.000	-0.843	0.472
Oil	Apparel	-0.095	0.302	1.000	-0.904	0.714
	Mining	0.221	0.154	0.917	-0.191	0.633
	Semiconductors	0.036	0.252	1.000	-0.638	0.710
Semiconductors	Apparel	-0.131	0.347	1.000	-1.060	0.798
	Mining	0.185	0.246	1.000	-0.472	0.843
	Oil	-0.036	0.252	1.000	-0.710	0.638

Based on estimated marginal means

a. Adjustment for multiple comparisons: Bonferroni.

Estimates

Dependent Variable: Subordination

region	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Australia	.445 ^a	0.028	0.390	0.499
Europe	.430 ^a	0.025	0.381	0.479
India	.289 ^a	0.053	0.186	0.393
UK	.494 ^a	0.028	0.439	0.549
USA	.450 ^a	0.023	0.404	0.496

a. Covariates appearing in the model are evaluated at the following values: env_perf = 65.21, soc_perf = 68.53, gov_perf = 69.65, ecn_perf = 60.60, total_assets = 24198506.95.

Pairwise Comparisons

Dependent Variable: Subordination

(I) region		Mean Difference (I-J)	Std. Error	Sig. ^b	95% Confidence Interval for Difference ^b	
					Lower Bound	Upper Bound
Australia	Europe	0.015	0.037	1.000	-0.091	0.121
	India	0.156	0.058	0.086	-0.011	0.322
	UK	-0.049	0.033	1.000	-0.144	0.045
	USA	-0.005	0.034	1.000	-0.102	0.092
Europe	Australia	-0.015	0.037	1.000	-0.121	0.091
	India	0.140	0.049	0.050	0.000	0.281
	UK	-0.065	0.036	0.782	-0.169	0.039
	USA	-0.020	0.034	1.000	-0.118	0.077
India	Australia	-0.156	0.058	0.086	-0.322	0.011
	Europe	-0.140	0.049	0.050	-0.281	0.000
	UK	-.205*	0.058	0.005	-0.370	-0.040
	USA	-0.161	0.057	0.057	-0.324	0.003
UK	Australia	0.049	0.033	1.000	-0.045	0.144
	Europe	0.065	0.036	0.782	-0.039	0.169
	India	.205*	0.058	0.005	0.040	0.370
	USA	0.044	0.033	1.000	-0.050	0.139
USA	Australia	0.005	0.034	1.000	-0.092	0.102
	Europe	0.020	0.034	1.000	-0.077	0.118
	India	0.161	0.057	0.057	-0.003	0.324
	UK	-0.044	0.033	1.000	-0.139	0.050

Based on estimated marginal means

*. The mean difference is significant at the .05 level.

b. Adjustment for multiple comparisons: Bonferroni.

Estimates

Dependent Variable: Subordination

industry	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Apparel	.438 ^a	0.045	0.349	0.528
Mining	.410 ^a	0.016	0.378	0.442
Oil	.440 ^a	0.019	0.402	0.477
Semiconductors	.399 ^a	0.036	0.328	0.469

a. Covariates appearing in the model are evaluated at the following values: env_perf = 65.21, soc_perf = 68.53, gov_perf = 69.65, ecn_perf = 60.60, total_assets = 24198506.95.

Pairwise Comparisons

Dependent Variable: Subordination

(I) industry		Mean Difference (I-J)	Std. Error	Sig. ^a	95% Confidence Interval for Difference ^a	
					Lower Bound	Upper Bound
Apparel	Mining	0.028	0.047	1.000	-0.096	0.153
	Oil	-0.001	0.047	1.000	-0.128	0.125
	Semiconductors	0.040	0.054	1.000	-0.105	0.185
Mining	Apparel	-0.028	0.047	1.000	-0.153	0.096
	Oil	-0.030	0.024	1.000	-0.094	0.035
	Semiconductors	0.011	0.038	1.000	-0.091	0.114
Oil	Apparel	0.001	0.047	1.000	-0.125	0.128
	Mining	0.030	0.024	1.000	-0.035	0.094
	Semiconductors	0.041	0.039	1.000	-0.064	0.146
Semiconductors	Apparel	-0.040	0.054	1.000	-0.185	0.105
	Mining	-0.011	0.038	1.000	-0.114	0.091
	Oil	-0.041	0.039	1.000	-0.146	0.064

Based on estimated marginal means

a. Adjustment for multiple comparisons: Bonferroni.

Estimates

Dependent Variable:

Passivisation

region	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Australia	.315 ^a	0.018	0.279	0.351
Europe	.300 ^a	0.016	0.267	0.332
India	.302 ^a	0.035	0.233	0.370
UK	.283 ^a	0.018	0.246	0.319
USA	.202 ^a	0.015	0.172	0.233

a. Covariates appearing in the model are evaluated at the following values: env_perf = 65.21, soc_perf = 68.53, gov_perf = 69.65, ecn_perf = 60.60, total_assets = 24198506.95.

Pairwise Comparisons

Dependent Variable:

Passivisation

(I) region		Mean Difference (I-J)	Std. Error	Sig. ^b	95% Confidence Interval for Difference ^b	
					Lower Bound	Upper Bound
Australia	Europe	0.016	0.025	1.000	-0.054	0.086
	India	0.014	0.039	1.000	-0.096	0.123
	UK	0.033	0.022	1.000	-0.030	0.095
	USA	.113 [*]	0.022	0.000	0.049	0.177
Europe	Australia	-0.016	0.025	1.000	-0.086	0.054
	India	-0.002	0.032	1.000	-0.095	0.090
	UK	0.017	0.024	1.000	-0.052	0.085
	USA	.097 [*]	0.022	0.000	0.033	0.161
India	Australia	-0.014	0.039	1.000	-0.123	0.096
	Europe	0.002	0.032	1.000	-0.090	0.095
	UK	0.019	0.038	1.000	-0.090	0.128
	USA	0.099	0.038	0.096	-0.009	0.207
UK	Australia	-0.033	0.022	1.000	-0.095	0.030
	Europe	-0.017	0.024	1.000	-0.085	0.052
	India	-0.019	0.038	1.000	-0.128	0.090
	USA	.080 [*]	0.022	0.003	0.018	0.143
USA	Australia	-.113 [*]	0.022	0.000	-0.177	-0.049
	Europe	-.097 [*]	0.022	0.000	-0.161	-0.033
	India	-0.099	0.038	0.096	-0.207	0.009
	UK	-.080 [*]	0.022	0.003	-0.143	-0.018

Based on estimated marginal means

*. The mean difference is significant at the .05 level.

b. Adjustment for multiple comparisons: Bonferroni.

Estimates

Dependent Variable:

Passivisation

industry	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Apparel	.258 ^a	0.030	0.199	0.317
Mining	.288 ^a	0.011	0.267	0.310
Oil	.295 ^a	0.013	0.271	0.320
Semiconductors	.279 ^a	0.024	0.233	0.326

a. Covariates appearing in the model are evaluated at the following values: env_perf = 65.21, soc_perf = 68.53, gov_perf = 69.65, ecn_perf = 60.60, total_assets = 24198506.95.

Pairwise Comparisons

Dependent Variable:

Passivisation

(I) industry		Mean Difference (I-J)	Std. Error	Sig. ^a	95% Confidence Interval for Difference ^a	
					Lower Bound	Upper Bound
Apparel	Mining	-0.031	0.031	1.000	-0.113	0.052
	Oil	-0.038	0.031	1.000	-0.121	0.046
	Semiconductors	-0.022	0.036	1.000	-0.117	0.074
Mining	Apparel	0.031	0.031	1.000	-0.052	0.113
	Oil	-0.007	0.016	1.000	-0.049	0.035
	Semiconductors	0.009	0.025	1.000	-0.059	0.077
Oil	Apparel	0.038	0.031	1.000	-0.046	0.121
	Mining	0.007	0.016	1.000	-0.035	0.049
	Semiconductors	0.016	0.026	1.000	-0.053	0.086
Semiconductors	Apparel	0.022	0.036	1.000	-0.074	0.117
	Mining	-0.009	0.025	1.000	-0.077	0.059
	Oil	-0.016	0.026	1.000	-0.086	0.053

Based on estimated marginal means

a. Adjustment for multiple comparisons: Bonferroni.

Financial (Annual Report) Letters to Shareholders

Estimates

Dependent Variable: Flesch Reading Ease Index

region	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Australia	33.174 ^a	1.433	30.348	36.001
Europe	35.943 ^a	1.637	32.715	39.170
India	39.897 ^a	3.466	33.062	46.731
UK	37.352 ^a	1.815	33.773	40.932
USA	34.127 ^a	1.173	31.813	36.441

a. Covariates appearing in the model are evaluated at the following values: env_perf = 54.80, soc_perf = 55.99, gov_perf = 67.94, ecn_perf = 53.27, total_assets = 19003738.45.

Pairwise Comparisons

Dependent Variable: Flesch Reading Ease Index

(I) region		Mean Difference (I-J)	Std. Error	Sig. ^a	95% Confidence Interval for Difference ^a	
					Lower Bound	Upper Bound
Australia	Europe	-2.769	2.195	1.000	-9.000	3.463
	India	-6.723	3.672	0.686	-17.146	3.701
	UK	-4.178	2.074	0.453	-10.066	1.709
	USA	-0.953	1.703	1.000	-5.787	3.882
Europe	Australia	2.769	2.195	1.000	-3.463	9.000
	India	-3.954	3.266	1.000	-13.226	5.318
	UK	-1.410	2.349	1.000	-8.077	5.258
	USA	1.816	2.082	1.000	-4.095	7.727
India	Australia	6.723	3.672	0.686	-3.701	17.146
	Europe	3.954	3.266	1.000	-5.318	13.226
	UK	2.544	3.764	1.000	-8.140	13.228
	USA	5.770	3.694	1.000	-4.715	16.256
UK	Australia	4.178	2.074	0.453	-1.709	10.066
	Europe	1.410	2.349	1.000	-5.258	8.077
	India	-2.544	3.764	1.000	-13.228	8.140
	USA	3.226	2.049	1.000	-2.590	9.041
USA	Australia	0.953	1.703	1.000	-3.882	5.787
	Europe	-1.816	2.082	1.000	-7.727	4.095
	India	-5.770	3.694	1.000	-16.256	4.715
	UK	-3.226	2.049	1.000	-9.041	2.590

Based on estimated marginal means

a. Adjustment for multiple comparisons: Bonferroni.

Estimates

Dependent Variable: Flesch Reading Ease Index

industry	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Apparel	39.386 ^a	2.657	34.146	44.626
Mining	35.768 ^a	1.050	33.698	37.838
Oil	36.811 ^a	1.127	34.589	39.034
Semiconductors	32.429 ^a	1.880	28.722	36.136

a. Covariates appearing in the model are evaluated at the following values: env_perf = 54.80, soc_perf = 55.99, gov_perf = 67.94, ecn_perf = 53.27, total_assets = 19003738.45.

Pairwise Comparisons

Dependent Variable: Flesch Reading Ease Index

(I) industry		Mean Difference (I-J)	Std. Error	Sig. ^a	95% Confidence Interval for Difference ^a	
					Lower Bound	Upper Bound
Apparel	Mining	3.618	2.713	1.000	-3.613	10.850
	Oil	2.575	2.732	1.000	-4.706	9.855
	Semiconductors	6.957	2.932	0.112	-0.858	14.772
Mining	Apparel	-3.618	2.713	1.000	-10.850	3.613
	Oil	-1.044	1.388	1.000	-4.743	2.656
	Semiconductors	3.339	1.995	0.575	-1.979	8.656
Oil	Apparel	-2.575	2.732	1.000	-9.855	4.706
	Mining	1.044	1.388	1.000	-2.656	4.743
	Semiconductors	4.382	1.983	0.169	-0.902	9.666
Semiconductors	Apparel	-6.957	2.932	0.112	-14.772	0.858
	Mining	-3.339	1.995	0.575	-8.656	1.979
	Oil	-4.382	1.983	0.169	-9.666	0.902

Based on estimated marginal means

a. Adjustment for multiple comparisons: Bonferroni.

Estimates

Dependent Variable: Flesch-Kincaid Grade Level

region	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Australia	15.087 ^a	0.344	14.408	15.765
Europe	14.406 ^a	0.393	13.632	15.181
India	13.192 ^a	0.832	11.552	14.832
UK	14.325 ^a	0.436	13.466	15.184
USA	14.246 ^a	0.282	13.691	14.801

a. Covariates appearing in the model are evaluated at the following values: env_perf = 54.80, soc_perf = 55.99, gov_perf = 67.94, ecn_perf = 53.27, total_assets = 19003738.45.

Pairwise Comparisons

Dependent Variable: Flesch-Kincaid Grade Level

(I) region		Mean Difference (I-J)	Std. Error	Sig. ^a	95% Confidence Interval for Difference ^a	
					Lower Bound	Upper Bound
Australia	Europe	0.680	0.527	1.000	-0.815	2.176
	India	1.895	0.881	0.328	-0.607	4.396
	UK	0.762	0.498	1.000	-0.651	2.175
	USA	0.841	0.409	0.410	-0.319	2.001
Europe	Australia	-0.680	0.527	1.000	-2.176	0.815
	India	1.214	0.784	1.000	-1.011	3.439
	UK	0.081	0.564	1.000	-1.519	1.681
	USA	0.160	0.500	1.000	-1.258	1.579
India	Australia	-1.895	0.881	0.328	-4.396	0.607
	Europe	-1.214	0.784	1.000	-3.439	1.011
	UK	-1.133	0.903	1.000	-3.697	1.431
	USA	-1.054	0.886	1.000	-3.570	1.462
UK	Australia	-0.762	0.498	1.000	-2.175	0.651
	Europe	-0.081	0.564	1.000	-1.681	1.519
	India	1.133	0.903	1.000	-1.431	3.697
	USA	0.079	0.492	1.000	-1.317	1.474
USA	Australia	-0.841	0.409	0.410	-2.001	0.319
	Europe	-0.160	0.500	1.000	-1.579	1.258
	India	1.054	0.886	1.000	-1.462	3.570
	UK	-0.079	0.492	1.000	-1.474	1.317

Based on estimated marginal means

a. Adjustment for multiple comparisons: Bonferroni.

Estimates

Dependent Variable: Flesch-Kincaid Grade Level

industry	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Apparel	13.756 ^a	0.638	12.498	15.013
Mining	14.325 ^a	0.252	13.828	14.821
Oil	14.233 ^a	0.270	13.699	14.766
Semiconductors	14.691 ^a	0.451	13.802	15.581

a. Covariates appearing in the model are evaluated at the following values: env_perf = 54.80, soc_perf = 55.99, gov_perf = 67.94, ecn_perf = 53.27, total_assets = 19003738.45.

Pairwise Comparisons

Dependent Variable: Flesch-Kincaid Grade Level

(I) industry		Mean Difference (I-J)	Std. Error	Sig. ^a	95% Confidence Interval for Difference ^a	
					Lower Bound	Upper Bound
Apparel	Mining	-0.569	0.651	1.000	-2.304	1.166
	Oil	-0.477	0.656	1.000	-2.224	1.270
	Semiconductors	-0.936	0.704	1.000	-2.811	0.940
Mining	Apparel	0.569	0.651	1.000	-1.166	2.304
	Oil	0.092	0.333	1.000	-0.796	0.980
	Semiconductors	-0.367	0.479	1.000	-1.643	0.909
Oil	Apparel	0.477	0.656	1.000	-1.270	2.224
	Mining	-0.092	0.333	1.000	-0.980	0.796
	Semiconductors	-0.459	0.476	1.000	-1.727	0.809
Semiconductors	Apparel	0.936	0.704	1.000	-0.940	2.811
	Mining	0.367	0.479	1.000	-0.909	1.643
	Oil	0.459	0.476	1.000	-0.809	1.727

Based on estimated marginal means

a. Adjustment for multiple comparisons: Bonferroni.

Estimates

Dependent Variable: Gunning Fog Index

region	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Australia	18.961 ^a	0.379	18.213	19.708
Europe	17.991 ^a	0.433	17.137	18.844
India	16.864 ^a	0.917	15.056	18.673
UK	18.101 ^a	0.480	17.154	19.048
USA	17.711 ^a	0.310	17.099	18.323

a. Covariates appearing in the model are evaluated at the following values: env_perf = 54.80, soc_perf = 55.99, gov_perf = 67.94, ecn_perf = 53.27, total_assets = 19003738.45.

Pairwise Comparisons

Dependent Variable: Gunning Fog Index

(I) region		Mean Difference (I-J)	Std. Error	Sig. ^a	95% Confidence Interval for Difference ^a	
					Lower Bound	Upper Bound
Australia	Europe	0.970	0.581	0.964	-0.679	2.619
	India	2.096	0.971	0.321	-0.661	4.854
	UK	0.859	0.549	1.000	-0.698	2.417
	USA	1.250	0.450	0.061	-0.029	2.529
Europe	Australia	-0.970	0.581	0.964	-2.619	0.679
	India	1.126	0.864	1.000	-1.327	3.579
	UK	-0.111	0.621	1.000	-1.875	1.653
	USA	0.280	0.551	1.000	-1.284	1.844
India	Australia	-2.096	0.971	0.321	-4.854	0.661
	Europe	-1.126	0.864	1.000	-3.579	1.327
	UK	-1.237	0.996	1.000	-4.063	1.590
	USA	-0.847	0.977	1.000	-3.621	1.928
UK	Australia	-0.859	0.549	1.000	-2.417	0.698
	Europe	0.111	0.621	1.000	-1.653	1.875
	India	1.237	0.996	1.000	-1.590	4.063
	USA	0.390	0.542	1.000	-1.148	1.929
USA	Australia	-1.250	0.450	0.061	-2.529	0.029
	Europe	-0.280	0.551	1.000	-1.844	1.284
	India	0.847	0.977	1.000	-1.928	3.621
	UK	-0.390	0.542	1.000	-1.929	1.148

Based on estimated marginal means

a. Adjustment for multiple comparisons: Bonferroni.

Estimates

Dependent Variable: Gunning Fog Index

industry	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Apparel	17.275 ^a	0.703	15.888	18.661
Mining	17.979 ^a	0.278	17.431	18.526
Oil	17.906 ^a	0.298	17.318	18.494
Semiconductors	18.542 ^a	0.497	17.562	19.523

a. Covariates appearing in the model are evaluated at the following values: env_perf = 54.80, soc_perf = 55.99, gov_perf = 67.94, ecn_perf = 53.27, total_assets = 19003738.45.

Pairwise Comparisons

Dependent Variable: Gunning Fog Index

(I) industry		Mean Difference (I-J)	Std. Error	Sig. ^a	95% Confidence Interval for Difference ^a	
					Lower Bound	Upper Bound
Apparel	Mining	-0.704	0.718	1.000	-2.617	1.209
	Oil	-0.632	0.723	1.000	-2.558	1.295
	Semiconductors	-1.268	0.776	0.623	-3.335	0.800
Mining	Apparel	0.704	0.718	1.000	-1.209	2.617
	Oil	0.073	0.367	1.000	-0.906	1.051
	Semiconductors	-0.564	0.528	1.000	-1.970	0.843
Oil	Apparel	0.632	0.723	1.000	-1.295	2.558
	Mining	-0.073	0.367	1.000	-1.051	0.906
	Semiconductors	-0.636	0.525	1.000	-2.034	0.762
Semiconductors	Apparel	1.268	0.776	0.623	-0.800	3.335
	Mining	0.564	0.528	1.000	-0.843	1.970
	Oil	0.636	0.525	1.000	-0.762	2.034

Based on estimated marginal means

a. Adjustment for multiple comparisons: Bonferroni.

Financial (Annual Report) Letters to Shareholders – Lexicosyntactic Features

Estimates

Dependent Variable: Lexical Density

region	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Australia	.625 ^a	0.004	0.617	0.633
Europe	.610 ^a	0.005	0.601	0.620
India	.641 ^a	0.010	0.621	0.660
UK	.616 ^a	0.005	0.606	0.626
USA	.638 ^a	0.003	0.631	0.644

a. Covariates appearing in the model are evaluated at the following values: env_perf = 54.80, soc_perf = 55.99, gov_perf = 67.94, ecn_perf = 53.27, total_assets = 19003738.45.

Pairwise Comparisons

Dependent Variable: Lexical Density

(I) region	Mean Difference (I-J)	Std. Error	Sig. ^b	95% Confidence Interval for Difference ^b		
				Lower Bound	Upper Bound	
Australia	Europe	0.015	0.006	0.186	-0.003	0.033
	India	-0.015	0.010	1.000	-0.045	0.014
	UK	0.009	0.006	1.000	-0.008	0.026
	USA	-0.012	0.005	0.115	-0.026	0.001
Europe	Australia	-0.015	0.006	0.186	-0.033	0.003
	India	-.030*	0.009	0.014	-0.057	-0.004
	UK	-0.006	0.007	1.000	-0.025	0.013
	USA	-.027*	0.006	0.000	-0.044	-0.010
India	Australia	0.015	0.010	1.000	-0.014	0.045
	Europe	.030*	0.009	0.014	0.004	0.057
	UK	0.024	0.011	0.240	-0.006	0.055
	USA	0.003	0.011	1.000	-0.027	0.033
UK	Australia	-0.009	0.006	1.000	-0.026	0.008
	Europe	0.006	0.007	1.000	-0.013	0.025
	India	-0.024	0.011	0.240	-0.055	0.006
	USA	-.021*	0.006	0.003	-0.038	-0.005
USA	Australia	0.012	0.005	0.115	-0.001	0.026
	Europe	.027*	0.006	0.000	0.010	0.044
	India	-0.003	0.011	1.000	-0.033	0.027
	UK	.021*	0.006	0.003	0.005	0.038

Based on estimated marginal means

*. The mean difference is significant at the .05 level.

b. Adjustment for multiple comparisons: Bonferroni.

Estimates

Dependent Variable: Lexical Density

industry	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Apparel	.614 ^a	0.008	0.599	0.629
Mining	.626 ^a	0.003	0.620	0.632
Oil	.630 ^a	0.003	0.624	0.637
Semiconductors	.634 ^a	0.005	0.623	0.644

a. Covariates appearing in the model are evaluated at the following values: env_perf = 54.80, soc_perf = 55.99, gov_perf = 67.94, ecn_perf = 53.27, total_assets = 19003738.45.

Pairwise Comparisons

Dependent Variable: Lexical Density

(I) industry		Mean Difference (I-J)	Std. Error	Sig. ^a	95% Confidence Interval for Difference ^a	
					Lower Bound	Upper Bound
Apparel	Mining	-0.012	0.008	0.768	-0.032	0.009
	Oil	-0.016	0.008	0.216	-0.037	0.004
	Semiconductors	-0.020	0.008	0.117	-0.042	0.003
Mining	Apparel	0.012	0.008	0.768	-0.009	0.032
	Oil	-0.005	0.004	1.000	-0.015	0.006
	Semiconductors	-0.008	0.006	1.000	-0.023	0.007
Oil	Apparel	0.016	0.008	0.216	-0.004	0.037
	Mining	0.005	0.004	1.000	-0.006	0.015
	Semiconductors	-0.003	0.006	1.000	-0.018	0.012
Semiconductors	Apparel	0.020	0.008	0.117	-0.003	0.042
	Mining	0.008	0.006	1.000	-0.007	0.023
	Oil	0.003	0.006	1.000	-0.012	0.018

Based on estimated marginal means

a. Adjustment for multiple comparisons: Bonferroni.

Estimates

Dependent Variable: Parse Tree Depth

region	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Australia	11.357 ^a	0.208	10.946	11.767
Europe	11.014 ^a	0.238	10.545	11.482
India	10.169 ^a	0.503	9.177	11.161
UK	11.144 ^a	0.264	10.625	11.664
USA	10.484 ^a	0.170	10.148	10.820

a. Covariates appearing in the model are evaluated at the following values: env_perf = 54.80, soc_perf = 55.99, gov_perf = 67.94, ecn_perf = 53.27, total_assets = 19003738.45.

Pairwise Comparisons

Dependent Variable: Parse Tree Depth

(i) region	Mean Difference (I-J)	Std. Error	Sig. ^b	95% Confidence Interval for Difference ^b		
				Lower Bound	Upper Bound	
Australia	Europe	0.343	0.319	1.000	-0.562	1.248
	India	1.188	0.533	0.270	-0.326	2.701
	UK	0.212	0.301	1.000	-0.642	1.067
	USA	.873 [*]	0.247	0.005	0.171	1.575
Europe	Australia	-0.343	0.319	1.000	-1.248	0.562
	India	0.845	0.474	0.764	-0.501	2.191
	UK	-0.131	0.341	1.000	-1.099	0.837
	USA	0.530	0.302	0.812	-0.328	1.388
India	Australia	-1.188	0.533	0.270	-2.701	0.326
	Europe	-0.845	0.474	0.764	-2.191	0.501
	UK	-0.975	0.546	0.758	-2.527	0.576
	USA	-0.315	0.536	1.000	-1.837	1.208
UK	Australia	-0.212	0.301	1.000	-1.067	0.642
	Europe	0.131	0.341	1.000	-0.837	1.099
	India	0.975	0.546	0.758	-0.576	2.527
	USA	0.660	0.297	0.275	-0.184	1.505
USA	Australia	-.873 [*]	0.247	0.005	-1.575	-0.171
	Europe	-0.530	0.302	0.812	-1.388	0.328
	India	0.315	0.536	1.000	-1.208	1.837
	UK	-0.660	0.297	0.275	-1.505	0.184

Based on estimated marginal means

*. The mean difference is significant at the .05 level.

b. Adjustment for multiple comparisons: Bonferroni.

Estimates

Dependent Variable: Parse Tree Depth

industry	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Apparel	10.864 ^a	0.386	10.103	11.625
Mining	10.821 ^a	0.152	10.521	11.122
Oil	10.873 ^a	0.164	10.551	11.196
Semiconductors	10.775 ^a	0.273	10.237	11.313

a. Covariates appearing in the model are evaluated at the following values: env_perf = 54.80, soc_perf = 55.99, gov_perf = 67.94, ecn_perf = 53.27, total_assets = 19003738.45.

Pairwise Comparisons

Dependent Variable: Parse Tree Depth

(I) industry		Mean Difference (I-J)	Std. Error	Sig. ^a	95% Confidence Interval for Difference ^a	
					Lower Bound	Upper Bound
Apparel	Mining	0.043	0.394	1.000	-1.007	1.093
	Oil	-0.009	0.397	1.000	-1.066	1.048
	Semiconductors	0.089	0.426	1.000	-1.046	1.224
Mining	Apparel	-0.043	0.394	1.000	-1.093	1.007
	Oil	-0.052	0.202	1.000	-0.589	0.485
	Semiconductors	0.046	0.290	1.000	-0.726	0.818
Oil	Apparel	0.009	0.397	1.000	-1.048	1.066
	Mining	0.052	0.202	1.000	-0.485	0.589
	Semiconductors	0.098	0.288	1.000	-0.669	0.865
Semiconductors	Apparel	-0.089	0.426	1.000	-1.224	1.046
	Mining	-0.046	0.290	1.000	-0.818	0.726
	Oil	-0.098	0.288	1.000	-0.865	0.669

Based on estimated marginal means

a. Adjustment for multiple comparisons: Bonferroni.

Estimates

Dependent Variable: Subordination

region	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Australia	.543 ^a	0.035	0.475	0.611
Europe	.583 ^a	0.039	0.505	0.660
India	.377 ^a	0.083	0.213	0.542
UK	.564 ^a	0.044	0.477	0.650
USA	.497 ^a	0.028	0.441	0.553

a. Covariates appearing in the model are evaluated at the following values: env_perf = 54.80, soc_perf = 55.99, gov_perf = 67.94, ecn_perf = 53.27, total_assets = 19003738.45.

Pairwise Comparisons

Dependent Variable: Subordination

(I) region		Mean Difference (I-J)	Std. Error	Sig. ^a	95% Confidence Interval for Difference ^a	
					Lower Bound	Upper Bound
Australia	Europe	-0.040	0.053	1.000	-0.190	0.110
	India	0.165	0.088	0.626	-0.085	0.416
	UK	-0.021	0.050	1.000	-0.162	0.121
	USA	0.046	0.041	1.000	-0.070	0.162
Europe	Australia	0.040	0.053	1.000	-0.110	0.190
	India	0.205	0.079	0.098	-0.018	0.428
	UK	0.019	0.057	1.000	-0.141	0.180
	USA	0.086	0.050	0.891	-0.057	0.228
India	Australia	-0.165	0.088	0.626	-0.416	0.085
	Europe	-0.205	0.079	0.098	-0.428	0.018
	UK	-0.186	0.091	0.412	-0.443	0.071
	USA	-0.120	0.089	1.000	-0.372	0.133
UK	Australia	0.021	0.050	1.000	-0.121	0.162
	Europe	-0.019	0.057	1.000	-0.180	0.141
	India	0.186	0.091	0.412	-0.071	0.443
	USA	0.067	0.049	1.000	-0.073	0.207
USA	Australia	-0.046	0.041	1.000	-0.162	0.070
	Europe	-0.086	0.050	0.891	-0.228	0.057
	India	0.120	0.089	1.000	-0.133	0.372
	UK	-0.067	0.049	1.000	-0.207	0.073

Based on estimated marginal means

a. Adjustment for multiple comparisons: Bonferroni.

Estimates

Dependent Variable: Subordination

industry	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Apparel	.569 ^a	0.064	0.443	0.695
Mining	.509 ^a	0.025	0.459	0.559
Oil	.462 ^a	0.027	0.408	0.515
Semiconductors	.511 ^a	0.045	0.422	0.600

a. Covariates appearing in the model are evaluated at the following values: env_perf = 54.80, soc_perf = 55.99, gov_perf = 67.94, ecn_perf = 53.27, total_assets = 19003738.45.

Pairwise Comparisons

Dependent Variable: Subordination

(I) industry		Mean Difference (I-J)	Std. Error	Sig. ^a	95% Confidence Interval for Difference ^a	
					Lower Bound	Upper Bound
Apparel	Mining	0.060	0.065	1.000	-0.114	0.234
	Oil	0.107	0.066	0.624	-0.068	0.283
	Semiconductors	0.058	0.071	1.000	-0.130	0.246
Mining	Apparel	-0.060	0.065	1.000	-0.234	0.114
	Oil	0.047	0.033	0.944	-0.042	0.136
	Semiconductors	-0.002	0.048	1.000	-0.130	0.126
Oil	Apparel	-0.107	0.066	0.624	-0.283	0.068
	Mining	-0.047	0.033	0.944	-0.136	0.042
	Semiconductors	-0.049	0.048	1.000	-0.176	0.078
Semiconductors	Apparel	-0.058	0.071	1.000	-0.246	0.130
	Mining	0.002	0.048	1.000	-0.126	0.130
	Oil	0.049	0.048	1.000	-0.078	0.176

Based on estimated marginal means

a. Adjustment for multiple comparisons: Bonferroni.

Estimates

Dependent Variable: Passivisation

region	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Australia	.236 ^a	0.014	0.207	0.264
Europe	.228 ^a	0.016	0.195	0.260
India	.234 ^a	0.035	0.166	0.303
UK	.231 ^a	0.018	0.196	0.267
USA	.170 ^a	0.012	0.147	0.193

a. Covariates appearing in the model are evaluated at the following values: env_perf = 54.80, soc_perf = 55.99, gov_perf = 67.94, ecn_perf = 53.27, total_assets = 19003738.45.

Pairwise Comparisons

Dependent Variable: Passivisation

(I) region		Mean Difference (I-J)	Std. Error	Sig. ^b	95% Confidence Interval for Difference ^b	
					Lower Bound	Upper Bound
Australia	Europe	0.008	0.022	1.000	-0.054	0.070
	India	0.001	0.037	1.000	-0.103	0.105
	UK	0.004	0.021	1.000	-0.055	0.063
	USA	.066*	0.017	0.001	0.018	0.114
Europe	Australia	-0.008	0.022	1.000	-0.070	0.054
	India	-0.007	0.033	1.000	-0.099	0.086
	UK	-0.004	0.023	1.000	-0.070	0.063
	USA	0.058	0.021	0.058	-0.001	0.117
India	Australia	-0.001	0.037	1.000	-0.105	0.103
	Europe	0.007	0.033	1.000	-0.086	0.099
	UK	0.003	0.038	1.000	-0.104	0.110
	USA	0.065	0.037	0.808	-0.040	0.170
UK	Australia	-0.004	0.021	1.000	-0.063	0.055
	Europe	0.004	0.023	1.000	-0.063	0.070
	India	-0.003	0.038	1.000	-0.110	0.104
	USA	.062*	0.020	0.029	0.004	0.120
USA	Australia	-.066*	0.017	0.001	-0.114	-0.018
	Europe	-0.058	0.021	0.058	-0.117	0.001
	India	-0.065	0.037	0.808	-0.170	0.040
	UK	-.062*	0.020	0.029	-0.120	-0.004

Based on estimated marginal means

*. The mean difference is significant at the .05 level.

b. Adjustment for multiple comparisons: Bonferroni.

Estimates

Dependent Variable: Passivisation

industry	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Apparel	.237 ^a	0.027	0.184	0.289
Mining	.213 ^a	0.010	0.192	0.233
Oil	.215 ^a	0.011	0.192	0.237
Semiconductors	.215 ^a	0.019	0.178	0.252

a. Covariates appearing in the model are evaluated at the following values: env_perf = 54.80, soc_perf = 55.99, gov_perf = 67.94, ecn_perf = 53.27, total_assets = 19003738.45.

Pairwise Comparisons

Dependent Variable: Passivisation

(I) industry		Mean Difference (I-J)	Std. Error	Sig. ^a	95% Confidence Interval for Difference ^a	
					Lower Bound	Upper Bound
Apparel	Mining	0.024	0.027	1.000	-0.048	0.096
	Oil	0.022	0.027	1.000	-0.051	0.095
	Semiconductors	0.022	0.029	1.000	-0.056	0.100
Mining	Apparel	-0.024	0.027	1.000	-0.096	0.048
	Oil	-0.002	0.014	1.000	-0.039	0.035
	Semiconductors	-0.002	0.020	1.000	-0.056	0.051
Oil	Apparel	-0.022	0.027	1.000	-0.095	0.051
	Mining	0.002	0.014	1.000	-0.035	0.039
	Semiconductors	0.000	0.020	1.000	-0.053	0.052
Semiconductors	Apparel	-0.022	0.029	1.000	-0.100	0.056
	Mining	0.002	0.020	1.000	-0.051	0.056
	Oil	0.000	0.020	1.000	-0.052	0.053

Based on estimated marginal means

a. Adjustment for multiple comparisons: Bonferroni.

Sustainability Report Letters to Stakeholders - Readability Formulae

Estimates

Dependent Variable: Flesch Reading Ease Index

region	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Australia	32.121 ^a	2.999	26.143	38.099
Europe	34.004 ^a	2.760	28.504	39.504
India	37.560 ^a	6.037	25.529	49.592
UK	34.458 ^a	2.531	29.413	39.502
USA	23.813 ^a	1.948	19.931	27.695

a. Covariates appearing in the model are evaluated at the following values: env_perf = 73.05, soc_perf = 79.84, gov_perf = 75.23, ecn_perf = 68.31, total_assets = 34853831.81.

Pairwise Comparisons

Dependent Variable: Flesch Reading Ease Index

(I) region		Mean Difference (I-J)	Std. Error	Sig. ^b	95% Confidence Interval for Difference ^b	
					Lower Bound	Upper Bound
Australia	Europe	-1.883	3.910	1.000	-13.203	9.437
	India	-5.439	6.845	1.000	-25.255	14.377
	UK	-2.336	3.067	1.000	-11.214	6.542
	USA	8.308	3.177	0.108	-0.888	17.504
Europe	Australia	1.883	3.910	1.000	-9.437	13.203
	India	-3.556	5.348	1.000	-19.039	11.926
	UK	-0.454	3.537	1.000	-10.692	9.785
	USA	10.191 [*]	3.193	0.021	0.949	19.433
India	Australia	5.439	6.845	1.000	-14.377	25.255
	Europe	3.556	5.348	1.000	-11.926	19.039
	UK	3.103	6.589	1.000	-15.973	22.178
	USA	13.747	6.447	0.363	-4.916	32.410
UK	Australia	2.336	3.067	1.000	-6.542	11.214
	Europe	0.454	3.537	1.000	-9.785	10.692
	India	-3.103	6.589	1.000	-22.178	15.973
	USA	10.644 [*]	2.784	0.003	2.584	18.705
USA	Australia	-8.308	3.177	0.108	-17.504	0.888
	Europe	-10.191 [*]	3.193	0.021	-19.433	-0.949
	India	-13.747	6.447	0.363	-32.410	4.916
	UK	-10.644 [*]	2.784	0.003	-18.705	-2.584

Based on estimated marginal means

*. The mean difference is significant at the .05 level.

b. Adjustment for multiple comparisons: Bonferroni.

Estimates

Dependent Variable: Flesch Reading Ease Index

industry	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Apparel	39.907 ^a	5.298	29.348	50.465
Mining	28.064 ^a	1.652	24.771	31.356
Oil	29.924 ^a	1.892	26.155	33.694
Semiconductors	31.670 ^a	3.064	25.563	37.778

a. Covariates appearing in the model are evaluated at the following values: env_perf = 73.05, soc_perf = 79.84, gov_perf = 75.23, ecn_perf = 68.31, total_assets = 34853831.81.

Pairwise Comparisons

Dependent Variable: Flesch Reading Ease Index

(I) industry		Mean Difference (I-J)	Std. Error	Sig. ^a	95% Confidence Interval for Difference ^a	
					Lower Bound	Upper Bound
Apparel	Mining	11.843	5.431	0.195	-2.887	26.573
	Oil	9.982	5.315	0.386	-4.431	24.396
	Semiconductors	8.237	5.576	0.864	-6.886	23.359
Mining	Apparel	-11.843	5.431	0.195	-26.573	2.887
	Oil	-1.861	2.243	1.000	-7.943	4.221
	Semiconductors	-3.606	3.420	1.000	-12.883	5.670
Oil	Apparel	-9.982	5.315	0.386	-24.396	4.431
	Mining	1.861	2.243	1.000	-4.221	7.943
	Semiconductors	-1.746	3.393	1.000	-10.948	7.457
Semiconductors	Apparel	-8.237	5.576	0.864	-23.359	6.886
	Mining	3.606	3.420	1.000	-5.670	12.883
	Oil	1.746	3.393	1.000	-7.457	10.948

Based on estimated marginal means

a. Adjustment for multiple comparisons: Bonferroni.

Estimates

Dependent Variable: Flesch-Kincaid Grade Level

region	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Australia	14.854 ^a	0.738	13.383	16.325
Europe	14.269 ^a	0.679	12.915	15.622
India	12.396 ^a	1.485	9.436	15.356
UK	14.205 ^a	0.623	12.964	15.446
USA	15.964 ^a	0.479	15.009	16.920

a. Covariates appearing in the model are evaluated at the following values: env_perf = 73.05, soc_perf = 79.84, gov_perf = 75.23, ecn_perf = 68.31, total_assets = 34853831.81.

Pairwise Comparisons

Dependent Variable: Flesch-Kincaid Grade Level

(I) region		Mean Difference (I-J)	Std. Error	Sig. ^a	95% Confidence Interval for Difference ^a	
					Lower Bound	Upper Bound
Australia	Europe	0.586	0.962	1.000	-2.200	3.371
	India	2.458	1.684	1.000	-2.417	7.334
	UK	0.650	0.755	1.000	-1.535	2.834
	USA	-1.110	0.782	1.000	-3.373	1.152
Europe	Australia	-0.586	0.962	1.000	-3.371	2.200
	India	1.873	1.316	1.000	-1.937	5.682
	UK	0.064	0.870	1.000	-2.455	2.583
	USA	-1.696	0.786	0.342	-3.970	0.578
India	Australia	-2.458	1.684	1.000	-7.334	2.417
	Europe	-1.873	1.316	1.000	-5.682	1.937
	UK	-1.809	1.621	1.000	-6.502	2.885
	USA	-3.568	1.586	0.275	-8.160	1.023
UK	Australia	-0.650	0.755	1.000	-2.834	1.535
	Europe	-0.064	0.870	1.000	-2.583	2.455
	India	1.809	1.621	1.000	-2.885	6.502
	USA	-1.760	0.685	0.123	-3.743	0.224
USA	Australia	1.110	0.782	1.000	-1.152	3.373
	Europe	1.696	0.786	0.342	-0.578	3.970
	India	3.568	1.586	0.275	-1.023	8.160
	UK	1.760	0.685	0.123	-0.224	3.743

Based on estimated marginal means

a. Adjustment for multiple comparisons: Bonferroni.

Estimates

Dependent Variable: Flesch-Kincaid Grade Level

industry	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Apparel	13.194 ^a	1.304	10.597	15.792
Mining	15.164 ^a	0.406	14.354	15.974
Oil	14.585 ^a	0.465	13.658	15.513
Semiconductors	14.407 ^a	0.754	12.904	15.909

a. Covariates appearing in the model are evaluated at the following values: env_perf = 73.05, soc_perf = 79.84, gov_perf = 75.23, ecn_perf = 68.31, total_assets = 34853831.81.

Pairwise Comparisons

Dependent Variable: Flesch-Kincaid Grade Level

(I) industry		Mean Difference (I-J)	Std. Error	Sig. ^a	95% Confidence Interval for Difference ^a	
					Lower Bound	Upper Bound
Apparel	Mining	-1.970	1.336	0.869	-5.594	1.654
	Oil	-1.391	1.308	1.000	-4.937	2.156
	Semiconductors	-1.212	1.372	1.000	-4.933	2.509
Mining	Apparel	1.970	1.336	0.869	-1.654	5.594
	Oil	0.579	0.552	1.000	-0.917	2.076
	Semiconductors	0.758	0.842	1.000	-1.525	3.040
Oil	Apparel	1.391	1.308	1.000	-2.156	4.937
	Mining	-0.579	0.552	1.000	-2.076	0.917
	Semiconductors	0.179	0.835	1.000	-2.086	2.443
Semiconductors	Apparel	1.212	1.372	1.000	-2.509	4.933
	Mining	-0.758	0.842	1.000	-3.040	1.525
	Oil	-0.179	0.835	1.000	-2.443	2.086

Based on estimated marginal means

a. Adjustment for multiple comparisons: Bonferroni.

Estimates

Dependent Variable: Gunning Fog Index

region	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Australia	18.142 ^a	0.794	16.558	19.725
Europe	17.630 ^a	0.731	16.173	19.087
India	15.398 ^a	1.599	12.211	18.584
UK	17.533 ^a	0.670	16.197	18.869
USA	19.475 ^a	0.516	18.447	20.503

a. Covariates appearing in the model are evaluated at the following values: env_perf = 73.05, soc_perf = 79.84, gov_perf = 75.23, ecn_perf = 68.31, total_assets = 34853831.81.

Pairwise Comparisons

Dependent Variable: Gunning Fog Index

(I) region		Mean Difference (I-J)	Std. Error	Sig. ^a	95% Confidence Interval for Difference ^a	
					Lower Bound	Upper Bound
Australia	Europe	0.511	1.036	1.000	-2.487	3.510
	India	2.744	1.813	1.000	-2.505	7.993
	UK	0.608	0.812	1.000	-1.743	2.960
	USA	-1.333	0.841	1.000	-3.769	1.103
Europe	Australia	-0.511	1.036	1.000	-3.510	2.487
	India	2.233	1.417	1.000	-1.868	6.334
	UK	0.097	0.937	1.000	-2.615	2.809
	USA	-1.845	0.846	0.324	-4.293	0.603
India	Australia	-2.744	1.813	1.000	-7.993	2.505
	Europe	-2.233	1.417	1.000	-6.334	1.868
	UK	-2.136	1.745	1.000	-7.188	2.917
	USA	-4.077	1.708	0.195	-9.021	0.866
UK	Australia	-0.608	0.812	1.000	-2.960	1.743
	Europe	-0.097	0.937	1.000	-2.809	2.615
	India	2.136	1.745	1.000	-2.917	7.188
	USA	-1.942	0.738	0.103	-4.077	0.194
USA	Australia	1.333	0.841	1.000	-1.103	3.769
	Europe	1.845	0.846	0.324	-0.603	4.293
	India	4.077	1.708	0.195	-0.866	9.021
	UK	1.942	0.738	0.103	-0.194	4.077

Based on estimated marginal means

a. Adjustment for multiple comparisons: Bonferroni.

Estimates

Dependent Variable: Gunning Fog Index

industry	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Apparel	16.116 ^a	1.403	13.320	18.913
Mining	18.571 ^a	0.438	17.699	19.443
Oil	18.153 ^a	0.501	17.154	19.151
Semiconductors	17.702 ^a	0.812	16.084	19.320

a. Covariates appearing in the model are evaluated at the following values: env_perf = 73.05, soc_perf = 79.84, gov_perf = 75.23, ecn_perf = 68.31, total_assets = 34853831.81.

Pairwise Comparisons

Dependent Variable: Gunning Fog Index

(I) industry		Mean Difference (I-J)	Std. Error	Sig. ^a	95% Confidence Interval for Difference ^a	
					Lower Bound	Upper Bound
Apparel	Mining	-2.454	1.439	0.554	-6.356	1.448
	Oil	-2.036	1.408	0.914	-5.854	1.781
	Semiconductors	-1.586	1.477	1.000	-5.591	2.420
Mining	Apparel	2.454	1.439	0.554	-1.448	6.356
	Oil	0.418	0.594	1.000	-1.193	2.029
	Semiconductors	0.869	0.906	1.000	-1.588	3.326
Oil	Apparel	2.036	1.408	0.914	-1.781	5.854
	Mining	-0.418	0.594	1.000	-2.029	1.193
	Semiconductors	0.451	0.899	1.000	-1.987	2.888
Semiconductors	Apparel	1.586	1.477	1.000	-2.420	5.591
	Mining	-0.869	0.906	1.000	-3.326	1.588
	Oil	-0.451	0.899	1.000	-2.888	1.987

Based on estimated marginal means

a. Adjustment for multiple comparisons: Bonferroni.

Sustainability Report Letters to Stakeholders - Lexicosyntactic Features

Estimates

Dependent Variable: Lexical Density

region	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Australia	.613 ^a	0.009	0.594	0.632
Europe	.594 ^a	0.009	0.577	0.612
India	.609 ^a	0.019	0.571	0.647
UK	.587 ^a	0.008	0.571	0.603
USA	.610 ^a	0.006	0.597	0.622

a. Covariates appearing in the model are evaluated at the following values: env_perf = 73.05, soc_perf = 79.84, gov_perf = 75.23, ecn_perf = 68.31, total_assets = 34853831.81.

Pairwise Comparisons

Dependent Variable: Lexical Density

(I) region		Mean Difference (I-J)	Std. Error	Sig. ^a	95% Confidence Interval for Difference ^a	
					Lower Bound	Upper Bound
Australia	Europe	0.019	0.012	1.000	-0.017	0.054
	India	0.003	0.022	1.000	-0.059	0.066
	UK	0.025	0.010	0.103	-0.003	0.053
	USA	0.003	0.010	1.000	-0.026	0.032
Europe	Australia	-0.019	0.012	1.000	-0.054	0.017
	India	-0.015	0.017	1.000	-0.064	0.034
	UK	0.007	0.011	1.000	-0.025	0.039
	USA	-0.016	0.010	1.000	-0.045	0.014
India	Australia	-0.003	0.022	1.000	-0.066	0.059
	Europe	0.015	0.017	1.000	-0.034	0.064
	UK	0.022	0.021	1.000	-0.038	0.082
	USA	0.000	0.020	1.000	-0.059	0.059
UK	Australia	-0.025	0.010	0.103	-0.053	0.003
	Europe	-0.007	0.011	1.000	-0.039	0.025
	India	-0.022	0.021	1.000	-0.082	0.038
	USA	-0.022	0.009	0.126	-0.048	0.003
USA	Australia	-0.003	0.010	1.000	-0.032	0.026
	Europe	0.016	0.010	1.000	-0.014	0.045
	India	0.000	0.020	1.000	-0.059	0.059
	UK	0.022	0.009	0.126	-0.003	0.048

Based on estimated marginal means

a. Adjustment for multiple comparisons: Bonferroni.

Estimates

Dependent Variable:

Lexical Density

industry	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Apparel	.574 ^a	0.017	0.541	0.607
Mining	.607 ^a	0.005	0.596	0.617
Oil	.612 ^a	0.006	0.600	0.624
Semiconductors	.618 ^a	0.010	0.599	0.637

a. Covariates appearing in the model are evaluated at the following values: env_perf = 73.05, soc_perf = 79.84, gov_perf = 75.23, ecn_perf = 68.31, total_assets = 34853831.81.

Pairwise Comparisons

Dependent Variable:

Lexical Density

(I) industry		Mean Difference (I-J)	Std. Error	Sig. ^a	95% Confidence Interval for Difference ^a	
					Lower Bound	Upper Bound
Apparel	Mining	-0.032	0.017	0.375	-0.079	0.014
	Oil	-0.038	0.017	0.160	-0.083	0.008
	Semiconductors	-0.044	0.018	0.090	-0.092	0.004
Mining	Apparel	0.032	0.017	0.375	-0.014	0.079
	Oil	-0.006	0.007	1.000	-0.025	0.014
	Semiconductors	-0.011	0.011	1.000	-0.041	0.018
Oil	Apparel	0.038	0.017	0.160	-0.008	0.083
	Mining	0.006	0.007	1.000	-0.014	0.025
	Semiconductors	-0.006	0.011	1.000	-0.035	0.023
Semiconductors	Apparel	0.044	0.018	0.090	-0.004	0.092
	Mining	0.011	0.011	1.000	-0.018	0.041
	Oil	0.006	0.011	1.000	-0.023	0.035

Based on estimated marginal means

a. Adjustment for multiple comparisons: Bonferroni.

Estimates

Dependent Variable: Parse Tree Depth

region	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Australia	11.535 ^a	0.492	10.555	12.516
Europe	10.713 ^a	0.453	9.811	11.615
India	10.067 ^a	0.990	8.094	12.040
UK	11.313 ^a	0.415	10.486	12.140
USA	11.136 ^a	0.319	10.499	11.772

a. Covariates appearing in the model are evaluated at the following values: env_perf = 73.05, soc_perf = 79.84, gov_perf = 75.23, ecn_perf = 68.31, total_assets = 34853831.81.

Pairwise Comparisons

Dependent Variable: Parse Tree Depth

(I) region		Mean Difference (I-J)	Std. Error	Sig. ^a	95% Confidence Interval for Difference ^a	
					Lower Bound	Upper Bound
Australia	Europe	0.823	0.641	1.000	-1.034	2.679
	India	1.468	1.123	1.000	-1.781	4.718
	UK	0.223	0.503	1.000	-1.233	1.678
	USA	0.400	0.521	1.000	-1.108	1.908
Europe	Australia	-0.823	0.641	1.000	-2.679	1.034
	India	0.645	0.877	1.000	-1.894	3.184
	UK	-0.600	0.580	1.000	-2.279	1.079
	USA	-0.423	0.524	1.000	-1.939	1.093
India	Australia	-1.468	1.123	1.000	-4.718	1.781
	Europe	-0.645	0.877	1.000	-3.184	1.894
	UK	-1.246	1.081	1.000	-4.374	1.883
	USA	-1.068	1.057	1.000	-4.129	1.992
UK	Australia	-0.223	0.503	1.000	-1.678	1.233
	Europe	0.600	0.580	1.000	-1.079	2.279
	India	1.246	1.081	1.000	-1.883	4.374
	USA	0.177	0.457	1.000	-1.145	1.499
USA	Australia	-0.400	0.521	1.000	-1.908	1.108
	Europe	0.423	0.524	1.000	-1.093	1.939
	India	1.068	1.057	1.000	-1.992	4.129
	UK	-0.177	0.457	1.000	-1.499	1.145

Based on estimated marginal means

a. Adjustment for multiple comparisons: Bonferroni.

Estimates

Dependent Variable: Parse Tree Depth

industry	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Apparel	11.388 ^a	0.869	9.657	13.119
Mining	11.128 ^a	0.271	10.588	11.668
Oil	10.548 ^a	0.310	9.930	11.167
Semiconductors	10.746 ^a	0.503	9.745	11.748

a. Covariates appearing in the model are evaluated at the following values: env_perf = 73.05, soc_perf = 79.84, gov_perf = 75.23, ecn_perf = 68.31, total_assets = 34853831.81.

Pairwise Comparisons

Dependent Variable: Parse Tree Depth

(I) industry		Mean Difference (I-J)	Std. Error	Sig. ^a	95% Confidence Interval for Difference ^a	
					Lower Bound	Upper Bound
Apparel	Mining	0.260	0.891	1.000	-2.156	2.675
	Oil	0.840	0.871	1.000	-1.524	3.203
	Semiconductors	0.642	0.914	1.000	-1.838	3.122
Mining	Apparel	-0.260	0.891	1.000	-2.675	2.156
	Oil	0.580	0.368	0.715	-0.418	1.577
	Semiconductors	0.382	0.561	1.000	-1.139	1.903
Oil	Apparel	-0.840	0.871	1.000	-3.203	1.524
	Mining	-0.580	0.368	0.715	-1.577	0.418
	Semiconductors	-0.198	0.556	1.000	-1.707	1.311
Semiconductors	Apparel	-0.642	0.914	1.000	-3.122	1.838
	Mining	-0.382	0.561	1.000	-1.903	1.139
	Oil	0.198	0.556	1.000	-1.311	1.707

Based on estimated marginal means

a. Adjustment for multiple comparisons: Bonferroni.

Estimates

Dependent Variable: Subordination

region	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Australia	.682 ^a	0.097	0.489	0.876
Europe	.658 ^a	0.089	0.480	0.836
India	.395 ^a	0.195	0.006	0.784
UK	.761 ^a	0.082	0.598	0.924
USA	.690 ^a	0.063	0.564	0.815

a. Covariates appearing in the model are evaluated at the following values: env_perf = 73.05, soc_perf = 79.84, gov_perf = 75.23, ecn_perf = 68.31, total_assets = 34853831.81.

Pairwise Comparisons

Dependent Variable: Subordination

(I) region		Mean Difference (I-J)	Std. Error	Sig. ^a	95% Confidence Interval for Difference ^a	
					Lower Bound	Upper Bound
Australia	Europe	0.024	0.126	1.000	-0.342	0.390
	India	0.287	0.221	1.000	-0.354	0.928
	UK	-0.078	0.099	1.000	-0.365	0.209
	USA	-0.007	0.103	1.000	-0.305	0.290
Europe	Australia	-0.024	0.126	1.000	-0.390	0.342
	India	0.263	0.173	1.000	-0.238	0.763
	UK	-0.103	0.114	1.000	-0.434	0.228
	USA	-0.032	0.103	1.000	-0.330	0.267
India	Australia	-0.287	0.221	1.000	-0.928	0.354
	Europe	-0.263	0.173	1.000	-0.763	0.238
	UK	-0.365	0.213	0.905	-0.982	0.251
	USA	-0.294	0.208	1.000	-0.898	0.309
UK	Australia	0.078	0.099	1.000	-0.209	0.365
	Europe	0.103	0.114	1.000	-0.228	0.434
	India	0.365	0.213	0.905	-0.251	0.982
	USA	0.071	0.090	1.000	-0.190	0.332
USA	Australia	0.007	0.103	1.000	-0.290	0.305
	Europe	0.032	0.103	1.000	-0.267	0.330
	India	0.294	0.208	1.000	-0.309	0.898
	UK	-0.071	0.090	1.000	-0.332	0.190

Based on estimated marginal means

a. Adjustment for multiple comparisons: Bonferroni.

Estimates

Dependent Variable:

Subordination

industry	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Apparel	.889 ^a	0.171	0.548	1.231
Mining	.631 ^a	0.053	0.524	0.737
Oil	.500 ^a	0.061	0.379	0.622
Semiconductors	.529 ^a	0.099	0.331	0.726

a. Covariates appearing in the model are evaluated at the following values: env_perf = 73.05, soc_perf = 79.84, gov_perf = 75.23, ecn_perf = 68.31, total_assets = 34853831.81.

Pairwise Comparisons

Dependent Variable:

Subordination

(I) industry		Mean Difference (I-J)	Std. Error	Sig. ^a	95% Confidence Interval for Difference ^a	
					Lower Bound	Upper Bound
Apparel	Mining	0.258	0.176	0.875	-0.218	0.734
	Oil	0.389	0.172	0.160	-0.077	0.855
	Semiconductors	0.361	0.180	0.295	-0.128	0.850
Mining	Apparel	-0.258	0.176	0.875	-0.734	0.218
	Oil	0.130	0.073	0.456	-0.066	0.327
	Semiconductors	0.102	0.111	1.000	-0.198	0.402
Oil	Apparel	-0.389	0.172	0.160	-0.855	0.077
	Mining	-0.130	0.073	0.456	-0.327	0.066
	Semiconductors	-0.028	0.110	1.000	-0.326	0.269
Semiconductors	Apparel	-0.361	0.180	0.295	-0.850	0.128
	Mining	-0.102	0.111	1.000	-0.402	0.198
	Oil	0.028	0.110	1.000	-0.269	0.326

Based on estimated marginal means

a. Adjustment for multiple comparisons: Bonferroni.

Estimates

Dependent Variable: Passivisation

region	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Australia	.211 ^a	0.032	0.148	0.274
Europe	.216 ^a	0.029	0.158	0.274
India	.106 ^a	0.064	-0.021	0.233
UK	.213 ^a	0.027	0.160	0.266
USA	.150 ^a	0.021	0.109	0.191

a. Covariates appearing in the model are evaluated at the following values: env_perf = 73.05, soc_perf = 79.84, gov_perf = 75.23, ecn_perf = 68.31, total_assets = 34853831.81.

Pairwise Comparisons

Dependent Variable: Passivisation

(I) region		Mean Difference (I-J)	Std. Error	Sig. ^a	95% Confidence Interval for Difference ^a	
					Lower Bound	Upper Bound
Australia	Europe	-0.005	0.041	1.000	-0.125	0.114
	India	0.105	0.072	1.000	-0.104	0.314
	UK	-0.002	0.032	1.000	-0.096	0.091
	USA	0.060	0.034	0.765	-0.037	0.157
Europe	Australia	0.005	0.041	1.000	-0.114	0.125
	India	0.110	0.056	0.545	-0.053	0.274
	UK	0.003	0.037	1.000	-0.105	0.111
	USA	0.066	0.034	0.552	-0.032	0.163
India	Australia	-0.105	0.072	1.000	-0.314	0.104
	Europe	-0.110	0.056	0.545	-0.274	0.053
	UK	-0.107	0.070	1.000	-0.309	0.094
	USA	-0.045	0.068	1.000	-0.242	0.152
UK	Australia	0.002	0.032	1.000	-0.091	0.096
	Europe	-0.003	0.037	1.000	-0.111	0.105
	India	0.107	0.070	1.000	-0.094	0.309
	USA	0.063	0.029	0.363	-0.022	0.148
USA	Australia	-0.060	0.034	0.765	-0.157	0.037
	Europe	-0.066	0.034	0.552	-0.163	0.032
	India	0.045	0.068	1.000	-0.152	0.242
	UK	-0.063	0.029	0.363	-0.148	0.022

Based on estimated marginal means

a. Adjustment for multiple comparisons: Bonferroni.

Estimates

Dependent Variable: Passivisation

industry	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Apparel	.228 ^a	0.056	0.117	0.339
Mining	.184 ^a	0.017	0.149	0.219
Oil	.165 ^a	0.020	0.126	0.205
Semiconductors	.139 ^a	0.032	0.074	0.203

a. Covariates appearing in the model are evaluated at the following values: env_perf = 73.05, soc_perf = 79.84, gov_perf = 75.23, ecn_perf = 68.31, total_assets = 34853831.81.

Pairwise Comparisons

Dependent Variable: Passivisation

(I) industry		Mean Difference (I-J)	Std. Error	Sig. ^a	95% Confidence Interval for Difference ^a	
					Lower Bound	Upper Bound
Apparel	Mining	0.044	0.057	1.000	-0.112	0.199
	Oil	0.063	0.056	1.000	-0.090	0.215
	Semiconductors	0.089	0.059	0.804	-0.070	0.249
Mining	Apparel	-0.044	0.057	1.000	-0.199	0.112
	Oil	0.019	0.024	1.000	-0.045	0.083
	Semiconductors	0.045	0.036	1.000	-0.053	0.143
Oil	Apparel	-0.063	0.056	1.000	-0.215	0.090
	Mining	-0.019	0.024	1.000	-0.083	0.045
	Semiconductors	0.027	0.036	1.000	-0.071	0.124
Semiconductors	Apparel	-0.089	0.059	0.804	-0.249	0.070
	Mining	-0.045	0.036	1.000	-0.143	0.053
	Oil	-0.027	0.036	1.000	-0.124	0.071

Based on estimated marginal means

a. Adjustment for multiple comparisons: Bonferroni.

Appendix 2: Manipulation¹

¹ Participants read these texts as presented below, i.e. with some extra markup to improve legibility and better approximate the layout of a Letter to Stakeholders as it might appear in a report.

Readability Survey (adapted for offline display)

Archive Corporate Reporting Questionnaire

This questionnaire will present you with a chairman's letter that introduces a company's annual report. We would like to ask you some questions about that letter.

We will ask you to read such a letter. After you read the letter, we'll ask you a few questions about it and, finally, about yourself.

The entire questionnaire should take about 15 minutes.

Thank you very much for your time! We greatly appreciate your assistance.

1. Informed Consent

By taking this survey, you are agreeing with the following:

I give permission to the researcher and any possible future researchers to use the recorded materials and written surveys for scientific research. I agree that my personal information will be processed and used, and I know that I have the right to access and correct this information. The data will be processed anonymously and my privacy will be respected at any time.

(Participant reads one of three texts. They may refer back to it while answering questions)

2. Text

- What is the name of the company you just read about? **(Text box)**
- Which industry does it operate in?
 - Diamond industry
 - Chemical industry
 - Retail
 - Oil industry

3. Familiarity

	Not at all familiar	Somewhat familiar	Familiar	Very familiar
Corporate Reporting				
CEO letters				
Corporate Sustainability				

- If you indicate anything other than 'not at all familiar', please indicate the source of this familiarity (for example studies, work, etc.). Please be as specific as possible **(Text box)**

4. Adjectives

- To which extent did you find the company portrayed in the text... **(Random order for every respondent)**

	Strongly disagree	Disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Agree	Strongly agree
Open							
Honest							
Sincere							
Responsible							
Trustworthy							
Complex							
Competent							
Sustainable							
Professional							

5. Statements

- Based on the text you have just read, please indicate to which extent you agree with the following:
(Random order for every respondent)

	Strongly disagree	Disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Agree	Strongly agree
I have a good feeling about this company							
I respect this company							
I trust this company							
This company has a clear vision for its future							
This company is well managed							
This looks like a company that would have good employees							
This is an environmentally responsible company							
This company maintains a high standard in the way it treats people							
This looks like a company with strong prospects for future growth							

6. Sentiment

- How positive or negative did you find the text overall? (Weighed -3 through +3)
 - Very negative
 - Negative
 - Somewhat negative
 - Neither positive nor negative
 - Somewhat positive
 - Positive
 - Very positive

7. Readability

- Please indicate to which extent you thought the text was.... (Random order for every respondent)

	Strongly disagree	Disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Agree	Strongly agree
Clear							
Readable							
Written by an expert							
Complicated							
Well-written							
Easy to understand							
Persuasive							
Accessible							
Pleasant to read							

8. Difficulty

- How difficult did you find the text? (**Weighed -3 through +3**)
 - Very difficult
 - Difficult
 - Somewhat difficult
 - Neither easy nor difficult
 - Somewhat easy
 - Easy
 - Very easy

9. About you

- What is your gender?
 - Male
 - Female
 - Neither of the above
 - I'd rather not say
- What is your age? (**Text box**)
- Are you a native speaker of English?
 - Yes
 - No (specify native language)
- If you are a native speaker of English, what is your native variety of English?
 - American
 - British
 - Australian
 - Indian
 - Other (please specify)
- How would you rate your English?
 - Very weak
 - Weak
 - Quite weak
 - Average
 - Quite good
 - Good
 - Very good
 - Excellent
- What is the highest degree you have obtained so far?
 - Doctoral (PhD, MD,...)
 - Master (MA, MSc,...)
 - Bachelor (BA, BSc,...)
 - Secondary education (or equivalent)
 - Primary education
 - Other (please specify)
- Was/is English a part of your curriculum after secondary school?
 - Yes
 - No
 - Other (please specify)

LtS - Original Version: FRE 13.8

Chief Executive Officer's review

Ensuring a safe working environment for all our employees is of primary importance to us at Lustre Minerals and it is with great sadness that we need to report that three fatal incidents occurred during 2012 at our operations. Our heartfelt condolences go out to the families, friends and loved ones of our colleagues and team members, Opelo Mmolai, Mogakolodi Monthe and Mateboho S'kosana. We will continue to work hard to continuously improve our systems and eliminate risk in as far as is practicable in our workplace, thereby driving to achieve our target of zero harm.

Three business drivers of sustainable development

Our sustainable development framework is our response to three key business drivers.

- **Retaining our social licence to operate:** conducting our business in an ethical, transparent and responsible manner, will help us retain our social licence to operate. This requires a particular focus on managing and controlling risk and consequential impacts through understanding risk drivers and how these relate to our business processes.
- **Continuing to attract high quality customers:** our customers' expectations rise inexorably. To meet them, we continue to improve our sustainability performance as an integral characteristic of quality management. We are committed to a programme of continuous improvement to increase business value.
- **Continuously improve our reputation:** public perception of our Company and brand remains a key component of our business success.

Our strategy for sustainability ensures that business is conducted ethically and responsibly. We are committed to taking progressive steps towards aligning our vision and principles with sustainability best practice.

Sustainable development materiality

We believe that sustainability demonstrates our adaptability to a changing socioeconomic and bio-physical environment. We develop relationships with stakeholders based on trust, enabling us to continue to conduct our business in a responsible manner. During 2012, a collective effort across all business units resulted in the conceptualisation of 'The Lustre Way', that clearly communicates our philosophy of zero tolerance and our commitment to responsible care. Through 'The Lustre Way', we will secure the health and safety of our workforce, the responsible use of natural resources and the optimisation of benefits for those who may be affected by our operations.

We continue to closely assess our operations' impact, both positive and potentially negative, on our workforce, host communities, economies and the receiving environment. This risk-based approach informs our business strategy enabling it to continue to deliver sustainable value to our stakeholders.

Our sustainable development and corporate social responsibility strategy is focused on six core aspects:

- Creating a safe and healthy work environment for the workforce, including own and contractor employees.
- Ensuring an operationally intelligent and productive workforce by implementing appropriate strategies to develop and retain our employees.
- Reducing resource consumption in order to increase operational efficiency and profitability, whilst reducing dependency on natural resources that are increasingly constrained.
- Identifying, mitigating and managing our impacts on the natural environment.
- Leaving a positive legacy for our Company's Project Affected Communities, contributing to sustainable economic growth of the countries where we conduct our business and delivering sustainable value to shareholders.
- Maintaining the highest levels of product integrity and ensuring that all diamonds recovered are certified under the most stringent ethical standards.

By actively managing these material aspects in an integrated manner, we aim to minimise harm and optimise benefit.

What does the future hold?

At Lustre Minerals, we are committed to, and remain focused on continually improving performance and eliminating unacceptable risk to the business and all our stakeholders. Across all operations, we will expand initiatives to minimise resource consumption and optimise opportunities to create a lasting positive legacy in our Project Affected Communities. As a Company, we encourage two-way dialogue with all stakeholders to ensure that we continue to meet their expectations and truly uphold our commitment to responsible care.

Clifford Elphick

Chief Executive Officer

LtS - 'More Readable' Version: FRE 36.6

Chief Executive Officer's review

Lustre Minerals prioritises a safe working environment for all our employees, and it is with great sadness that we need to report that three fatal incidents occurred during 2012 at our operations. Our heartfelt condolences go out to the families, friends and loved ones of our colleagues and team members, Opelo Mmolai, Mogakolodi Monthe and Mateboho S'kosana. We will continue to work hard to improve our systems and eliminate risk. As far as is practicable in our workplace, we strive for our target of zero harm.

Three business drivers of sustainable development

Our sustainable development framework responds to three key business drivers. We must...

- **Keep our social licence to operate:** doing our business in an ethical, transparent and responsible manner will help us keep our social licence to operate. We must focus on managing and controlling risk and its impacts. If we understand risk drivers and how these relate to our business processes, we will be better able to control them.
- **Keep attracting high quality customers:** our customers' expectations keep rising. To meet them, we continue to improve our sustainability performance. This is a core part of quality management. We commit to keep improving, to increase business value.
- **Keep improving our reputation:** how the public sees our Company and brand is always a key part of our business success.

Our sustainability strategy makes sure that we conduct business ethically and responsibly. We are committed to moving towards aligning our vision and principles with sustainability best practice.

Sustainable development materiality

We believe that sustainability shows we can adapt to a changing social, economic, biological and physical environment. We develop relationships with stakeholders based on trust, which helps us conduct our business in a responsible manner. During 2012, an effort across all business units together outlined 'The Lustre Way', that explains our philosophy of zero tolerance and our commitment to responsible care. Through 'The Lustre Way', we will secure the health and safety of our workforce, the responsible use of natural resources and the optimal use of benefits for those whom our operations may affect.

We keep a close eye on our operations' positive and potential negative impact. Lustre's operations can affect our workforce, host communities, economies and the environment we exist in. Keeping these risks in mind helps us plan so we can keep delivering sustainable value to our stakeholders.

Our sustainable development and corporate social responsibility strategy focuses on six core aspects:

- Creating a safe and healthy work environment for the workforce. This includes own and contractor employees.
- Ensuring a productive workforce that knows how we work. We must implement the right strategies to develop and retain our employees.
- Using fewer resources to work more efficiently and profitably. We want to rely less on more constrained natural resources.
- Identifying, reducing and managing our impacts on the natural environment.
- Delivering sustainable value to shareholders. We want to leave a positive legacy for our Company's Project Affected Communities, and contribute to sustainable economic growth of the countries where we conduct our business.
- Maintaining the highest levels of product integrity and ensuring that those certifying diamonds we recover obey the highest ethical standards.

If all parts of our company bear these key issues in mind, we can minimise harm and optimise benefit.

What does the future hold?

Lustre Minerals continues to focus on improving its performance and stopping unacceptable risk to the business and all stakeholders. Across all operations, we will do more to minimise resource consumption and find more ways to create a lasting positive legacy in our Project Affected Communities. As a Company, we encourage two-way dialogue with all stakeholders. We want to keep meeting their expectations and upholding our commitment to responsible care.

Clifford Elphick

Chief Executive Officer

LtS - 'Most Readable' Version: FRE 47.1

Chief Executive Officer's review

Employee work safety is crucial to us and we are deeply sad having to report three fatal incidents during 2012. We offer our heartfelt condolences to the families, friends and loved ones of our colleagues and team members, Opelo Mmolai, Mogakolodi Monthe and Mateboho S'kosana. We will keep improving our systems to stop risk as much as our workplace allows. Our target remains zero harm.

Three business drivers of sustainable development

Three aspects drive our sustainable development framework. We want to:

- **Keep our social licence to operate.** Our business must be ethical, transparent and responsible. Only then can we keep our social licence to operate. We must manage and control risks and impacts. Understanding risks helps us control them.
- **Keep attracting high quality customers.** Our customers keep expecting more, so we keep performing more sustainably as part of quality management. We keep improving to increase business value.
- **Keep improving our reputation.** How the public sees our Company and brand is always key to our success.

Our sustainability strategy makes sure we do ethical and responsible business. We keep matching our vision and principles to sustainability best practice.

Sustainable development materiality

Sustainability means we adapt to different changes:

- Social;
- Economic;
- Biological; and
- Physical

We strive to build trust with stakeholders. This helps us do business responsibly. During 2012, all business units outlined 'The Lustre Way' together. It collects our thoughts on zero tolerance, and commits us to responsible care. 'The Lustre Way' helps us:

- Keep our workforce safe and healthy;
- Use natural resources responsibly; and
- Do the best we can for those our operations affect.

We keep close watch on how our operations do good, and potentially do harm. That can be to our workforce, host communities, economies and environment. If we know the risks, we can better plan sustainable value growth for our stakeholders.

We develop sustainably and socially responsibly in six ways:

- We create a safe and healthy work environment for own and contractor employees.
- We ensure a productive workforce that understands how we work. We must carry out the right plans to train and keep our employees.
- We use fewer resources to work more efficiently and profitably. We want to rely less on limited natural resources.
- We identify, reduce and manage our impacts on the natural environment.
- We deliver sustainable value to shareholders. We try to do lasting good for our Company's Project Affected Communities. We help the countries where we do business grow their economy sustainably.
- We maintain the highest levels of product integrity. Those certifying diamonds we recover must obey the highest ethical standards.

These points help all parts of our company do the least harm and benefit the most.

What does the future hold?

Lustre Minerals keeps focusing on performing better. We seek to end unacceptable risk to the business and all our stakeholders. All our operations will do more to use fewer resources and find more ways to do lasting good in our Project Affected Communities. As a Company, we promote two-way dialogue with all stakeholders. We want to keep meeting their expectations and committing to responsible care.

Clifford Elphick

Chief Executive Officer

Appendix 3: Sentiment Annotation

Sentiment Annotation Guidelines

The aim of these guidelines is to help you determine and annotate use of sentiment in CEO letters that are part of sustainability reports. They will first describe the annotation interface, and then explain how to annotate the texts contained therein.

All of these excerpts are from **2012**, so please think from that perspective. You will be asked to evaluate a number of questions which will be situated on sentence level and/or text level. Please try to read these texts **from the perspective of a shareholder or stakeholder**.

Please proceed as follows:

1. **Read the entire text** and answer the text level questions.
2. **Read the sentences and answer the sentence level questions.**

If anything in the text is unclear to you, feel free to **look up any information you need** (for instance on the internet).

Deciding between categories

With the exception of assigning scores and ranking elements, you can make your choices as follows:

1. Start at the first/upper category in the list of multiple choices
2. Consider whether this category applies. If yes, choose this category and move on to the next question.
3. Move one step down and repeat the previous step.

In other words, stop at the soonest applicable answer.

Using WebAnno

Go to anno.lt3.ugent.be/webanno and enter your username and password (if you do not have a username or password, please request one). From there, proceed to 'Annotation' and choose a text.

In 'Settings', enable both the 'Sentence' and 'Text' annotation layers. Set 'Number of Sentences' to whatever you prefer (you can change this later if it would be more convenient). You can also tweak the sidebar size and font to whatever you find most convenient. Note that if you display many sentences per page, the last line is sometimes hidden by the interface. You can fix this by going to the previous or next document, and back, or selecting another annotation.

In this interface, you annotate sentences and other spans of text by selecting (double clicking on) the first word of the sentence. On the right side of the screen, you can choose between the 'text' and 'sentence' layer, and answer the questions for the various sentences as soon as you have selected a layer.

The first span of text you annotate will be the document's title. Apply the 'text' layer to it and answer the text-level questions. After that, answer the sentence-level questions for every sentence.

An important note on how the texts were generated, and how to deal with errors: this corpus is the result of Optical Character Recognition applied to PDFs published by the reporting companies. While we found this to yield the best possible results, it is still not error-free. In some cases, characters will be misread. If this occurs, try to infer what the correct sequence of characters would be and treat the text as if that sequence is present. If you cannot infer the correct form and would be unable to analyse the sentence due to the error, simply skip the sentence and notify us.

In other cases, these errors may make a sentence go on too long (for example if there are commas instead of full stops). If this occurs, annotate the first word of every actual sentence, as if the errors were not there. In these cases, you will annotate multiple words per line.

Finally, it is also possible for sentences to terminate earlier than they should. When this occurs, simply annotate all of the fragments separately but identically.

The 'Done' button: please refrain from pressing the 'done' button. It locks the document for further editing, and requires a curator to undo. Use 'open', 'next' and 'previous' to navigate between files.

Deciding between categories

With the exception of assigning scores and ranking elements, you can make your choices as follows:

1. Start at the first/upper category in the list of multiple choices
2. Consider whether this category applies. If yes, choose this category and move on to the next question.
3. Move one step down and repeat the previous step.

In other words, stop at the soonest applicable answer.

Text level-questions

When answering text-level questions, **annotate the title** at the start of the document, using the 'Title' layer. This title will be formed as 'COMPANY NAME – DOCUMENT TYPE', e.g. 'PANORAMIC RESOURCES – CEO LETTER FROM SUSTAINABILITY REPORT'.

1. Which performance perspectives/aspects receive most attention in the text?

We focus on the four primary performance perspectives/aspects of sustainability reporting, and how much of the attention is devoted to these aspects. Those four performance perspectives are:

1. Financial
2. Environmental
3. Social
4. Governance

The addendum 'Sentiment Performance Aspects.docx' describes these perspectives in more detail. Note especially that 'Governance' will be the rarest performance aspect, and you should not expect every document to address it.

For each text, sort the aspects by how much of the attention is devoted to (1=most attention compared to the other aspects, 2 = less than 1 but less than the others, etc. - = no attention). Note that when all aspects are mentioned, you will have a range between 1 and 4.

2. How positive or negative is the text with respect to these aspects of sustainability?

2a. Does the text cast a positive or negative light on these aspects?

Does the text mainly express positive or negative sentiment regarding these aspects?

2b. Degree to which the sentiment is positive or negative

If the sentiment is generally positive, use a number from 1 (somewhat positive) through 3 (very positive). If the sentiment is generally negative, use a number from -1 (somewhat negative) through -3 (very negative).

Use 0 in cases where there is sentiment, but it is neither positive nor negative, or unclear based on the context. Such cases will be fairly rare on a sentence level, and extremely rare on a text level.

If a text contains no sentiment for a given perspective, use the value ‘-’, which means ‘no sentiment’. This will be your default choice in many cases.

There may be various reasons why a particular positive or negative sentiment can be considered to be stronger or weaker. What follows are a few (non-exhaustive) reasons to consider sentiment very strong (3 or -3) or very weak (1 or -1). Their presence or absence in the discussion of the different aspects of sustainability may be of guidance to your general impression regarding the degree to which a positive or negative sentiment is presented. However, just because one or more of these are present, this does not mean you *must* consider the sentiment to be stronger or weaker; they can simply help you decide.

Stronger:

- The presence of intensifiers in combination with positive/negative adjectives, e.g. ‘**very** good/bad performance’ instead of ‘good performance’.
- A markedly stronger choice of words, such as ‘staggering’ instead of ‘impressive’ or ‘deplorable’ instead of ‘poor’.
- Unusual (factual) content, such as ‘we tripled our profits/debts’ instead of simply ‘increased’.
- Etc.

Weaker:

- Downtoners; e.g. Results were somewhat disappointing (weaker than: results were disappointing);
- Modality markers, such as ‘might cause damage’ rather than ‘causes damage’; ‘possibly affects results’ rather than ‘affects results’
- The sentiment only being implicitly present in the sentence

Examples (based on sentences):

“The global oil and gas industry is facing a major challenge: satisfying growing energy demand in a strained labour market.” – *negative (-2) social sentiment (the energy market is growing, and energy production is typically environmentally destructive, and employment conditions are strained); possibly also slightly negative (-1) financial and environmental sentiment by implication.*

“Thus, Maurel & Prom has the duty to enhance, as much as possible, the beneficial aspects of its activities for the economic development of the regions that host them and to maintain the highest degree of protective vigilance in managing the potential impacts of its activities on the environment.” – *very strong positive (+3) environmental sentiment due to intensity of language.*

“In West Texas and Southeast New Mexico, we are drilling new wells in the historic Permian Basin, using technology to create opportunities for significant economic growth.” – *positive (+2) financial sentiment, but negative (-2) environmental sentiment (drilling wells in historic areas tends to be destructive).*

“The SB&I team felt, after focusing our reporting on key impacts and business targets, that NIKE had significant work to do internally on assurance.” – *very negative (-3) for governance; the company itself indicates that it had a reporting issue, and intensifies it with ‘significant’.*

3. Which type of linguistic act is the company performing in this text?

Choose which of the following best describes what type of message the company is basically delivering in the text. If nothing fits, choose the 'other' category. Restrict yourself to a maximum of three options (if any) (1 being the most prominent, 3 being the least prominent). When in doubt, prefer one of the more specific options, which are higher up in the table.

3. The company is mainly...
1. ... apologising for something
2. ... making a request
3. ... raising questions
4. ... expressing gratitude
5. ... presenting a declaration of intent
6. ... expressing a desire
7. ... making predictions about the future
8. ... expressing (a) belief(s)/opinion(s)
9. ... describing a state of affairs
... doing none of the above. (Other)

You can consider commitments or other forms of engagement to fall within the purview of 'intentions'.

In deciding whether the company is presenting a **belief/opinion** as opposed to a state of affairs, consider the primary purpose of the sentence: does the company mostly try to communicate an evaluation, impression or subjective position in general? If yes, we say that the company is expressing a belief or opinion. If the emphasis is more on the company describing a situation or events, those fall within a state of affairs.

EXAMPLES (based on sentences)

"In 2012, the Group continued to implement its sustainable development strategy formalised since 2006 in its "Safety, Environment and Quality" Charter." – **state of affairs**

"The global oil and gas industry is facing a major challenge: satisfying growing energy demand in a strained labour market." – **State of affairs**

"The sole aim is to eliminate fatalities and serious injuries from AWAC's operations." – **declaration of intent; the company makes an (implied) commitment to eliminate fatalities and serious injuries.**

"We expect low commodity prices to negatively impact our industry throughout the following years." – **Prediction**

"As ever, we are grateful to our shareholders for their trust in the company." – **Gratitude**

"We hope to start work on this development in the next year, although this will depend on the results of the survey." – **Desire**

"What is 'hydraulic fracturing' or 'fracking'?" – **Question**

4. How forward-looking is the text?

As you are evaluating at a text level, choose which timeframe the bulk of the text applies to.

Start at the top of the table and work your way down, choosing the first appropriate option. In other words, prefer the higher (more specific) option.

4. The text is situated mostly in...
1. ... the future, and expresses an intention on the company's part.
2. ... the future.
3. ... the past.
4. ... the present.
5. ... none of the above time frames. (Atemporal)

If the text mainly expresses a clear **intent** on the company's part to do something in the future, e.g. 'we will double profits by the next decade' or 'we expect to implement this zero-emissions policy by 2018', we consider this to be more forward-looking.

Intentions expressed as **habits, rhetorical moves or general truths**, such as 'we commit to being a good corporate citizen') do not fall within the most forward-looking category. We consider them situated in the **present** or **atemporal** categories, depending on the context.

EXAMPLES

"We maintained a well qualified emergency response and rescue team able to be immediately and efficiently mobilized." – *Past*

"Approximately 50% of our employees are from Upper Egypt, the area where Sukari is situated, which typically has less economic activity than the richer areas around the Nile delta." – *Present; we can assume this percentage to be variable and not a general truth*

"Training up a skilled local workforce, and promoting access for women to positions at every level of the organisation, are some of the issues to be addressed by the oil and gas industry." – *Future*

"From a 2005 baseline, a 25% reduction in average freshwater-use intensity by 2020 and 30% by 2030." – *Future, with intent (only a sentence fragment, however)*

Sentence level questions

Answer the following questions on sentence level. Do not be surprised to find a fair share of overlap with text-level questions. The overlap is needed for coarse-grained and fine-grained analysis. For examples in case of overlap, see above.

1. Which performance perspectives/aspects receive attention in the sentence?

The addendum 'Sentiment Performance Aspects.docx' describes these perspectives in more detail.

Proportion of attention devoted to each aspect NEED NOT be measured per sentence. However, if, for instance, two aspects are addressed, values for sentiment can be attributed to each aspect.

1b. Does the sentence cast a positive or negative light on these aspects?

i.e. Does the sentence express positive or negative sentiment?

1c. How positive or negative is the sentence with respect to these aspects of sustainability?

If the sentiment is generally positive, use a number from 1 (somewhat positive) through 3 (very positive). If the sentiment is generally negative, use a number from -1 (somewhat negative) through -3 (very negative).

Use 0 in cases where there is sentiment, but it is neither positive nor negative, or unclear based on the context. Such cases will be fairly rare.

If a sentence contains no sentiment for a given perspective, keep the default value ‘-’, which means ‘no sentiment’.

There may be various reasons why a particular positive or negative sentiment can be considered to be stronger or weaker. What follows are a few (non-exhaustive) reasons to consider sentiment very strong (3 or -3) or very weak (1 or -1). Their presence or absence in the discussion of the different aspects of sustainability may be of guidance to your general impression regarding the degree to which a positive or negative sentiment is presented. However, just because one or more of these are present, this does not mean you *must* consider the sentiment to be stronger or weaker; they can simply help you decide.

Stronger:

- The presence of intensifiers in combination with positive/negative adjectives, e.g. ‘**very** good/bad performance’ instead of ‘good performance’;
- A markedly stronger choice of words, such as ‘staggering’ instead of ‘impressive’ or ‘deplorable’ instead of ‘poor’;
- Unusual (factual) content, such as ‘we tripled our profits/debts’ instead of simply ‘increased’;
- Etc.

Weaker:

- Downtoners; e.g. Results were somewhat disappointing (weaker than: results were disappointing);
- Modality markers, such as ‘might cause damage’ rather than ‘causes damage’; ‘possibly affects results’ rather than ‘affects results’.
- Information being only implicitly present.
- Etc.

2. Which type of linguistic act is the company performing in this sentence?

Choose which of the following best describes what type of message the company delivers in this sentence. If nothing fits, choose the ‘other’ category. You do not need to rank the various options; simply choose the option that fits best.

When in doubt, prefer the more specific option, i.e. the one higher up in the table.

2. The company is mainly...
1. ... apologising for something
2. ... making a request
3. ... posing (a) question(s)
4. ... expressing gratitude
5. ... declaring an intention
6. ... expressing a desire
7. ... making (a) prediction(s)
8. ... expressing belief(s)/opinion(s)
9. ... describing a state of affairs
... doing none of the above. (Other)

3. To what extent is the company explicitly presented as the agent?

Next, determine to which extent the company is presented as the agent in the sentence. There are different ways in which agency may be presented, which are discussed below. Make a general assessment with regard to explicit presence based on the different manifestations you attest in your text.

Start at the top of the table and work your way down, choosing the first appropriate option. In other words, prefer the higher (more specific) option.

3. The agent is the sentence is...
1. ... (part of) the company as agent through a first-person pronoun (incl. possessive).
2. ... the company itself.
3. ... the company, through metonymy.
4. ... the company, but hidden or non-explicit (e.g. through passivisation).
5. ... a non-company agent.
6. ... a non-company agent, but hidden or non-explicit.

The company being the agent through a **first-person pronoun** can manifest through a personal pronoun ('we implemented new policies across...') or possessive pronoun ('our human resources division ensures...' or 'our profits enable us to...')

If the **company itself** is the explicit agent, they will generally refer to themselves by name or by 'the company', for instance in 'The company has made substantial improvements...' or 'Nike takes great care to...'

By '**metonymy**,' we mean that the **company is represented explicitly** by something that it is a part of (e.g. 'the oil industry') or something that is a part of it (e.g. 'the Human Resources division claims that...') Note that '*our* human resources division' falls within the 'first-person pronoun' category. Examples include 'the global energy industry is facing a major crisis' or 'AK Steel's grain-oriented steels conduct electricity in hydroelectric projects' (as a company's products are a part of it).

A sentence might contain the company (through metonymy or otherwise) as a **hidden agent** if it uses, for example, the passive voice. If the company is implied but not explicit as doing the action (i.e. not represented in the sentence), choose this option. For instance: 'A substantial setback was incurred in this development.' (rather than 'The company incurred...' or 'this development suffer a substantial setback...')

A **non-company agent** generally indicates that something other than the company is doing the action central to the sentence. For instance: "local action groups voted to go on strike after the incident".

A **hidden non-company agent** means that the agent is not present in the sentence, but is not implied to be the company. For example: "The company was fined €5m in December for discharging into rivers." (as opposed to 'Courts fined the company...')

EXAMPLES

"NIKE continues to seek quality and transparency in our performance management and reporting." – *Company itself, by name*

"As we do this, we have explored additional ways to provide confidence in our processes and our reported data." – *company through first-person pronoun*

"Following our FY05/06 and FY07-09 reports, NIKE's internal audit team was asked to review our sustainability reporting processes." – *hidden company agent (presumably, the company is asking)*

“Training up a skilled local workforce, and promoting access for women to positions at every level of the organisation, are some of the issues to be addressed by the oil and gas industry.” – *company through metonymy, as it belongs to this industry.*

“The Group's attractiveness was confirmed in 2012 by the expansion of its teams, Gabon leading the recruitment drive with 93 out of a total 110 new hires.” – *company through metonymy, as both the teams and Gabon (a subsidiary) are a part of the company.*

“This presentation is made in accordance with the terms of the Decree of 24 April 2012 relating to the obligation of corporate transparency in social and environmental matters.” – *hidden company agent – the company made the presentation, but is not explicitly present in the text.*

“[Hydrokinetic] technology uses river and ocean currents to generate renewable energy, which can be transferred to the power grid using equipment made with AK Steel's highly efficient electrical steels.” – *non-company agent; the technology in question does not belong to the company.*

“AK Steel's products are even being used to support the latest hydrokinetic power transmission and distribution systems.” – *hidden non-company agent; it is not entirely clear who uses the products, but it is unlikely to be the company itself.*

4. How forward-looking is the sentence?

Start at the top of the table and work your way down, choosing the first appropriate option. In other words, prefer the higher (more specific) option.

4. The sentence is situated mostly in...
1. ... the future, and expresses an intention on the company's part.
2. ... the future.
3. ... the past.
4. ... the present.
5. ... none of the above time frames. (Atemporal)

To reiterate: if the sentence mainly expresses a clear **intent** on the company's part to do something in the future, e.g. ‘we will double profits by the next decade’ or ‘we expect to implement this zero-emissions policy by 2018’, we consider this to be more forward-looking.

Intentions situated in the present (e.g. expressed as habits or rhetorical moves or general truths, such as ‘we commit to being a good corporate citizen’) do not fall within the most forward-looking category.

5. To what extent is the sentence (presented as) an opinion?

Start at the top of the table and work your way down, choosing the first appropriate option. In other words, prefer the higher (more specific) option.

5. The sentence is ...
1. ... presented an opinion.
2. ... presented as a fact with positive or negative colouring
3. ... presented as factual, but contains subjectivity.
4. ... (almost) entirely factual
5. ... not assertive.

We consider a sentence to **explicitly present itself as an opinion** when the speaker highlights their

own position in the discursive frame. This can occur (non-exhaustively) through ‘parenthetical verbs’ or ‘comment clauses’ (e.g. ‘we believe...’, ‘I think’, ‘I suppose’), which can also take an adverbial form (‘to be honest’, ‘as we mentioned’), or other means that modify modality or stance towards the idea expressed or claim made, such as adverbial constructions (‘supposedly’, ‘allegedly’).

In the preceding cases, indicate this as an **explicit opinion**. Generally, **if one or more linguistic elements’ primary purpose is to increase distance between the speaker and the idea expressed or claim made**, we consider the sentence an opinion in the wide sense of the word.

A sentence might **not explicitly (i.e. indirectly)** signal itself as an opinion, but still contain **words that provide extra colouring, e.g. through rich vocabulary, boosters or downtoners** (e.g. ‘we achieved stellar performance’ or ‘profits rose by 7%’ vs. we made a profit of \$3.5 million this year’.). These are presented as facts with positive or negative colouring. Even if a sentence is otherwise factual, if it contains positively or negatively coloured words (within this context), such as ‘profits rose’ or ‘safety improved’, you may place the sentences in this category.

Some sentences will have the appearance of a fact, but contain **subjective elements that are not necessarily positively or negatively coloured**. These will be elements that would (either or both) be difficult to achieve consensus on, or to measure or prove; they are derived from someone judging or assessing them to be so. These are, for instance, the word ‘natural’ in ‘GHG emissions are the natural corollary to our operations’ or ‘substantial’ in ‘These assets form a substantial part of our business’ (depending on the context, this may also be positively or negatively coloured). We consider these sentences to be **presented as factual without being entirely so**.

If none of the above applies, consider the sentence to be (almost) entirely factual, except if it makes no assertion at all. We consider a sentence to be **factual** when all the information contained within is **verifiable by measuring it and/or universally accepted**; in other words, it is objective, in a fairly loose sense of the word. We add ‘(almost) entirely’ to leave some margin for error, for example for information that is somewhat relative or subjective, but would still be universally accepted. For instance, in the sentence ‘Metal processing is energy-intensive’, ‘energy-intensive’ is still *technically* subjective in that it relies on human judgment, but not subjective in any practical sense of the word as virtually no-one would argue the opposite. We count these cases as facts. (Depending on the context, ‘energy-intensive’ could also be negatively coloured).

Finally, some sentences are **not assertions** (e.g. a simple question such as ‘How did we achieve this?’). In other cases, they will make an implicit assertion despite not looking like one (e.g. ‘How did we achieve such amazing results?’ implies that results are amazing.). Such cases would be labelled as a fact with positive or negative colouring.

EXAMPLES

“Following our FY05/06 and FY07-09 reports, NIKE's internal audit team was asked to review our sustainability reporting processes.” – *Fact*

As the operator, Alcoa has invested substantial intellectual, financial and system resources over several decades to understand the key drivers behind safety behaviour- Presented as a fact with positive or negative colouring

The Group's attractiveness was confirmed in 2012 by the expansion of its teams, Gabon leading the recruitment drive with 93 out of a total 110 new hires.” Presented as a fact with positive or negative colouring

“The Group's recruitment policy is aimed at providing it with the best skills to support its development.” Presented as a fact with positive or negative colouring

We understand there are opportunities to improve our data collection processes, especially where information comes from third parties such as contract factories or material vendors that supply to such factories.” – *Explicitly an opinion due to ‘understand’*.

Sentiment Annotation Guide – How to Identify Different Aspects of Performance

This document gives a brief overview of the four performance perspectives relevant to this annotation task.

Financial performance encompasses any aspect to do with the company's value and profitability, both in the short and the long term. For example, if a company's share price increases or decreases, this is relevant to financial performance. If a company posts a profit or loss, or forecasts one for the future, this is relevant to financial performance. Virtually any news that would raise or lower the company's share price without fitting in any of the other categories would be positive or negative financial news, respectively, as financial performance remains the primary benchmark by which companies are measured.

Environmental performance refers to the impact the company has on the environment. Positive news regarding environmental performance would be any change that lowers the company's (negative) impact on the environment, while negative news would be any change where it increases its (negative) impact on the environment. For instance, lowered carbon emissions would be positive news, while a company starting a new mining site or causing an oil spill would be negative news.

Social performance refers to the impact the company has on communities, employees and consumers. Positive news regarding social performance would be any change that lowers the company's negative impact on these stakeholders or increases the positive impact on them, while negative news would be the opposite. For instance, a decrease in worker injuries or increase in community program funding would be good news, while fatalities or protest would be negative news.

Governance performance refers to how the company is directed and organised; specifically, it is concerned with to which extent company organisation and leadership ensure that the company works towards representing shareholders' and stakeholders' interest fairly and ethically (as opposed to, primarily, placing management and other leadership's interests above them). Unsurprisingly, governance performance is most difficult to pin down amongst these four performance measures, especially because it requires knowledge of the company's organisation. In broad terms, any organisational change or action that ensures more ethical, equitable or transparent behaviour on the part of those leading the company will be favourable news from a governance perspective; the opposite is also true. For instance, the company performing an audit (which improves transparency) or implementing a whistleblower policy (which improves ethical behaviour) would be positive news, while the company's involvement in a corruption scandal would be (very) negative news. Simply because a text reports management's actions does not make it positive news from a governance perspective; a phrase such as 'Management is implementing new strategy to increase profitability' is positive news from a financial perspective, but not relevant to the governance perspective.

Sentiment Annotation Guide – Reference Flowchart

A. For every text, determine the following and annotate for the title:

1. Which performance aspects receive the most attention?				
1. <i>Financial</i>				
2. <i>Environmental</i>				
3. <i>Social</i>				
4. <i>Governance</i>				
1	2	3	4	0
Most	Less than 1, more than 3	Less than 2, more than 4	Less than 3, but some	None



2. What is the general evaluation expressed or implied about each of these performance aspects?							
1. <i>Financial</i>							
2. <i>Environmental</i>							
3. <i>Social</i>							
4. <i>Governance</i>							
-3	-2	-1	-	0	1	2	3
Very negative	Negative	Slightly negative	No sentiment	Sentiment, but neither positive nor negative	Slightly positive	Positive	Very positive



3. The company is primarily...									
1: <i>Most</i>									
2: <i>Second most</i>									
3: <i>Third most</i>									
1. Apologising	2. Making a request	3. Posing (a) question(s)	4. Expressing gratitude	5. Declaring an intention	6. Expressing a desire	7. Making (a) prediction(s)	8. Expressing belief(s) opinion(s)	9. Describing a state of affairs	(Other)



4. The text is primarily situated...				
1. Mostly in future; signals intent	2. Mostly in future	3. Mostly in past	4. Mostly in present	5. Atemporal

B. For every sentence, determine the following:

1. What is the evaluation expressed or implied about each of these performance aspects?							
1. <i>Financial</i>							
2. <i>Environmental</i>							
3. <i>Social</i>							
4. <i>Governance</i>							
-3	-2	-1	-	0	1	2	3

Very negative	Negative	Slightly negative	No sentiment	Sentiment, but neither positive nor negative	Slightly positive	Positive	Very positive
---------------	----------	-------------------	--------------	--	-------------------	----------	---------------



2. The company is... (work left to right)									
1. Apologising	2. Making a request	3. Posing (a) question(s)	4. Expressing gratitude	5. Declaring an intention	6. Expressing a desire	7. Making (a) prediction(s)	8. Expressing belief(s) / opinion(s)	9. Describing a state of affairs	(Other)



3. To which extent is the agent the company itself? (work left to right)					
1. (Part of) company through first person pronoun	2. Company itself is agent	3. Company through metonymy (i.e. part of company, or company as part)	4. Company, but hidden or non-explicit (e.g. passivisation)	5. Non-company agent	6. Hidden or non-explicit non-company agent



4. How forward-looking is the sentence? (work left to right)				
1. Mostly in future; signals intent	2. Mostly in future	3. Mostly in past	4. Mostly in present	5. Atemporal



5. To which extent is the claim presented as an opinion? (work left to right)				
1. Presented as opinion	2. Positive or negative colouring	3. Contains subjectivity	4. Factual	5. Not assertive

Appendix 4: Human Assessment & Machine Learning¹

¹ Assessors received these reports in plaintext format; we present them with some markup here for legibility purposes.

PVH – Excerpt from Sustainability Report

2013 Plans and Challenges

As we continue to grow, so too will our carbon footprint. Our main challenge will be to reduce the company's carbon footprint year-over-year. As we review and renew our CSR strategic objectives, we strive to reduce our resource use across the enterprise.

In 2013, we will expand our environmental baselines to include all North American operations, and we will seek to conduct a carbon emissions inventory for all employee air travel.

Other specific energy efficiency projects will include:

- Capturing energy consumption of building mechanical systems. Currently, we do not have a mechanism to capture this data. In 2013, we will begin to inventory mechanical systems in our owned facilities to determine opportunities for energy savings.
- Continuing to replace plumbing fixtures with more efficient ones across all U.S. offices. We will also conduct a waterless urinal analysis for our New York offices and a water fixture inventory for our warehouses. Additionally, we will explore the feasibility of rainwater harvesting at our Bridgewater office.
- Creating an inventory of cleaning chemicals at all offices, integrating pest management documentation, and adopting a green cleaning policy at the corporate level.
- Expanding conservation and energy efficiency improvement efforts at U.S. retail locations. Conservation guidelines that outline best practices for temperature set points, recycling policies and front door policies that conserve heating and cooling were given to retail managers. In addition, we are installing energy-efficient lighting as renovations at our retail stores occur.

In 2013, through PVH Europe's Associate Ambassador program several teams investigated the potential of reducing the amount and nature of packaging on our products. The outcome of these projects are currently under review. The CSR Steering Committee will determine how best to move forward with these great ideas in the near future.

Additionally in 2013, we will participate in the Environmental Defense Fund's Climate Corps program, a summer fellowship that places specially-trained graduate students in companies, cities and universities to build the business case for energy efficiency. We will be dedicating a resource to work with an EDF Climate Corps Fellow to help us improve energy efficiency.

Freeport-McMoRan – Excerpt from Sustainability Report

FREEPORT-Mcmoran COPPER & GOLD INC.

Freeport-Mcmoran Copper & Gold Inc. (Freeport-Mcmoran or the Company) is a leading international mining company with headquarters in Phoenix, Arizona. We operate large, long-lived, geographically diverse assets with significant proven and probable reserves of copper, gold and molybdenum. The Company has a dynamic portfolio of operating, expansion and growth projects in the copper industry and is the world's largest producer of molybdenum.

Freeport-Mcmoran's portfolio of assets includes the Grasberg minerals district in Indonesia, one of the world's largest copper and gold deposits; significant mining operations in the Americas, including the large-scale Morenci minerals district in North America and the Cerro Verde and El Abra operations in South America; and the Tenke Fungurume minerals district in the Democratic Republic of Congo (DRC).

In the second-quarter of 2013 the Company completed its three-way combination with Plains Exploration & Production Company (NYSE: PXP) and Mcmoran Exploration Co. (NYSE: MMR). The transactions add a high quality portfolio of oil and gas assets to Freeport-Mcmoran's global mining business to create a premier U.S.-based natural resource company. Freeport-Mcmoran's portfolio of oil and gas assets include strong oil production facilities in California, a growing production profile in the onshore Eagle Ford trend in Texas, significant production facilities and growth potential in the Deepwater Gulf of Mexico and large onshore resources in the Haynesville natural gas trend in Louisiana. In addition, Freeport-Mcmoran is an industry leader in the emerging ultra-deep gas trend with sizeable potential, located offshore in the shallow waters of the Gulf of Mexico and onshore in South Louisiana.

ABOUT THIS REPORT

This 2012 Working Toward Sustainable Development (WTSD) report provides our stakeholders with summary information on our sustainability programs, including policies, systems and performance data. Additional information is located on our website at www.fcx.com including specific topical reports, performance data and fact sheets.

Data presented in the report includes the primary operations of Freeport-Mcmoran's principal subsidiaries: PT Freeport Indonesia (PTFI) and Freeport-Mcmoran Corporation for the period January 1, 2012 to December 31, 2012 (oil and gas assets acquired in 2013 are not included in report boundary).

As a result of methodology changes or corrections, prior year data may be updated. Data presentation and comparisons may not meet the direct needs of all stakeholders, and we encourage users of this information to contact our Sustainable Development Department at sustainability@fmi.com with inquiries about our report. We appreciate

receiving feedback that will help us identify the topics that are of most interest to you and thus improve the quality of future reporting.

Cautionary Statement

This report contains forward-looking statements in which we discuss factors we believe may affect our performance in the future. Forward-looking statements are all statements other than statements of historical facts, such as statements regarding projected production and sales volumes. We caution readers that our actual results may differ materially from those anticipated or projected in the forward-looking statements. Important factors that can cause our actual results to differ are described in Freeport-Mcmoran's Annual Report on Form 10-K for the year ended December 31, 2012, filed with the Securities and Exchange Commission and available on our website at www.fcx.com.

Chevron – Excerpt from Sustainability Report

Executing With Excellence

Chevron is one of the largest producers of crude oil and natural gas in the U.S. Gulf of Mexico. A continuing commitment to safety, leading Operational Excellence (OE) programs and new technology allow us to tap into needed energy supplies. At the same time, our work in the Gulf creates jobs and grows businesses.

Above: Bob Miller is a field Health, Environment and Safety specialist in the Gulf of Mexico aboard the Pacific Santa Ana, a deepwater drillship built to Chevron's specifications and the first drillship with dual-gradient drilling capabilities, which can enhance the safety of deepwater drilling.

Chevron's Billy Varnado knows that the time for easily finding oil is gone. Twelve years ago, he worked on our first discovery in the deepwater region of the Gulf of Mexico, the Genesis Field, located approximately 150 miles (241 km) south of New Orleans, Louisiana, where we tapped resources 12,000 feet (3,658 m) below sea level. Today, his work takes him another 130 miles (209 km) south of Genesis to the Chevron-operated discoveries Jack and St. Malo, where we will seek energy at depths of 27,000 feet (8,230 m) below the water's surface.

"Jack and St. Malo are being developed at extreme depths and amid challenging temperatures, currents, pressures and drilling complexity," said Varnado, the Jack/St. Malo project director. "We implement processes to help ensure the health and safety of our people and the environment, from design to production and through the life of the field's operation, which can last for decades."

Jack and St. Malo highlight the complexity of finding new energy sources. The project involves two fields 25 miles (40 km) apart. Each field will have separate clusters of wellheads on the seafloor that will be connected to a single floating production unit located between the two facilities. When the \$7.5 billion Jack/St. Malo project comes on line in 2014, it is expected to supply energy resources for 30 to 40 years.

"Our growth depends on our ability to maintain the region's confidence in our deepwater drilling projects and practices. People expect that the energy the world needs will be produced safely and reliably," said Warner Williams, vice president of the Gulf of Mexico business unit. "There is no room for complacency in our operations."

