

Simplifying, Reading, and Machine Translating Health Content:
An Empirical Investigation of Usability

Alessandra Rossetti

B.A., M.A.

Thesis submitted for the degree of
Doctor of Philosophy

School of Applied Language and Intercultural Studies
Dublin City University

April 2019

Supervisor:

Dr Sharon O'Brien

Declaration

I hereby certify that this material, which I now submit for assessment on the programme of study leading to the award of Doctor of Philosophy is entirely my own work, and that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge breach any law of copyright, and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

Signed: _____

Alessandra Rossetti

ID No.: 15211745

Date: _____

Acknowledgements

The first two years of this research have been funded by the Faculty of Humanities and Social Sciences at Dublin City University (DCU). The following two years have been funded through the Irish Research Council Government of Ireland Postgraduate Scholarship Programme. The secondments to Cochrane UK and Arizona State University (ASU) have taken place as part of the International Network on Crisis Translation (INTERACT), and have received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 734211.

My utmost gratitude goes, first and foremost, to my supervisor, Dr Sharon O'Brien, for her contagious enthusiasm about research, for her unfailing guidance and support, and for always being generous with her time and expertise. I cannot thank her enough for the endless opportunities that she has given me, including the opportunity to be part of INTERACT, which allowed me to grow as a person and as a researcher in ways that I could not have imagined. It has been a pleasure working with her.

I also wish to express my deep gratitude to all the groups of participants who kindly agreed to be involved in my research. This thesis would simply not have existed without your contributions.

I am grateful for the possibility to conduct my PhD research at the School of Applied Language and Intercultural Studies at DCU, where members of staff have always been very supportive and have provided me with helpful feedback along the way. In particular, I would like to thank my panel member, Dr Ryoko Sasamoto, and the Director of the Centre for Translation and Textual Studies, Dr Áine McGillicuddy.

I would also like to thank Dr Federico Gaspari for supporting me since I first decided to embark on this PhD adventure, and for always encouraging me throughout these years.

I am very grateful to Dr Maeve Olohan and to Dr Mary Phelan for agreeing to examine this thesis and for their constructive feedback.

At each stage of the PhD process, I could benefit from the invaluable assistance of a number of organisations. In particular, I would like to thank: Andrew Bredenkamp and Sabine Lehmann at Acrolinx for kindly allowing me to use their software for my authoring experiment; Juliane Reid, Hayley Hassan, and Andrea Cervera from the Cochrane translation community for helping me with the recruitment of participants for the machine translation evaluation study (also, a special word of thanks to Juliane Ried for her insightful comments at the design stage of the authoring and the machine translation experiments); Therese Docherty and staff at Cochrane UK (Oxford) for hosting me for two months during my secondment as part of INTERACT, and for helping me gain a better understanding of the authoring workflow at Cochrane; and Professor Danielle McNamara and her team at the Science of Learning and Educational Technology Lab at ASU (Phoenix) for hosting me for four months during my INTERACT secondment, for sharing their vast knowledge in areas of psychology,

education, and natural language processing that were new to me, and for making me feel part of the team.

On a more personal note, they say it takes a village, and I was extremely lucky to have a village of friends every step of the way. Eline, Pat, Carlos, Carla, Silvia, Hari, Mohsen, Moign, and Dani, thank you for the adventurous hikes in the remote areas of Ireland, the sharing of beautiful homemade food, the Sunday brunches, the table tennis matches, the movie marathons, the karaoke nights, and all the good laughs. More importantly, thank you for always being there for me. Your friendship and unconditional support made every weight lighter. A special word of thanks to Silvia for joining me in the secondment to Oxford, and for being both a mentor and a friend.

Finally, I would like to thank my family and friends in Italy for always encouraging me, and for being understanding with my radio silence at the busiest times of the PhD.

Grazie mille!

Publications and presentations from this research

Publications:

Rossetti, A. and O'Brien, S. (forthcoming). Helping the helpers: Evaluating the impact of a controlled language checker on the intralingual and interlingual translation tasks involving volunteer health professionals. *Translation Studies, Special Issue "Social Translation: New Roles, New Actors?"*.

Rossetti, A. and O'Brien, S. 2018. Seeking health content online: A survey of Internet users' habits and needs. *IN: Read, T., Montaner, S. and Sedano, B. (eds.) Proceedings of the Third Congress on Technological Innovation for Specialized Linguistic Domains (TISLID 18)*. Berlin: Éditions Universitaires Européennes, pp. 115-136.

Presentations:

Rossetti, A. 2018. Spanish translations of Cochrane plain language summaries: Assessing the impact of a controlled language checker on machine translation quality. Unpublished conference paper at: *25th Cochrane Colloquium*, 16-18 September, Edinburgh, United Kingdom.

Rossetti, A. 2018. Cochrane plain language summaries: A study of authors' satisfaction and users' comprehension. Unpublished conference paper at: *16th International Conference of the EARLI Special Interest Group on Writing (SIG Writing)*, 29-31 August, University of Antwerp, Belgium.

Rossetti, A., Rodríguez Vázquez, S., Ried, J. and O'Brien, S. 2017. A comparison of different approaches for editing health-related information: An author's satisfaction perspective. Unpublished poster session (presented by Juliane Ried) at: *Global Evidence Summit*, 13-16 September, Cape Town, South Africa.

Rossetti, A. 2016. Readability and machine translatability of health content. Unpublished conference paper at: *International Postgraduate Conference on Translation and Interpreting (IPCITI)*, 12-13 December, Dublin City University, Ireland.

Rossetti, A. 2016. Cultural appropriateness of healthcare information. Unpublished conference paper at: *International Association for Languages and Intercultural Communication (IALIC), Bridging Across Languages and Cultures in Everyday Lives: New Roles for Changing Scenarios*, 25-27 November, Autonomous University of Barcelona, Spain.

Table of contents

Declaration.....	ii
Acknowledgements.....	iii
Publications and presentations from this research.....	v
List of abbreviations.....	xi
List of tables.....	xiii
List of figures.....	xiv
Abstract.....	xv
1. INTRODUCTION.....	1
1.1 Background, Motivation, and Scope of the Thesis.....	1
1.2 Research Questions and Experimental Variables.....	6
1.3 Framework and Type of Research.....	10
1.4 Thesis Structure.....	11
2. REVIEW OF LITERATURE ON USABILITY.....	13
2.1 Content and Organisation of the Chapter.....	13
2.2 Usability and Related Concepts.....	13
2.3 Usability of Tools/Environments for Content Production and Editing.....	16
2.4 Conclusions Based on this Literature Review.....	20
3. REVIEW OF LITERATURE ON TEXT SIMPLIFICATION.....	21
3.1 Content and Organisation of the Chapter.....	21
3.2 Text Simplification and CL: Definitions, Goals, and Approaches.....	21
3.3 Simplification of Health-Related Texts.....	27
3.4 Conclusions Based on this Literature Review.....	30
4. ASSESSING THE SATISFACTION OF COCHRANE AUTHORS.....	31
4.1 Aims of the Study on Authors' Satisfaction and Overview of the Chapter.....	31
4.2 Related Work on Satisfaction of Volunteers.....	32

4.3 Motivation for Examining Cochrane Authors' Satisfaction, Research Question, Research Hypotheses, and Characteristics of the Acrolinx CL Checker.....	35
4.4 Recruitment of Cochrane Authors.....	41
4.5 Experimental Environment, Procedure and Tasks.....	42
4.6 Experimental Design and Selection of Experimental Materials.....	49
4.7 Methods Adopted for the Assessment of Authors' Satisfaction.....	51
4.7.1 Rationale behind the Selection of the SUS.....	52
4.7.2 Characteristics and Adoption of the SUS.....	56
4.8 Data Analysis and Results.....	58
4.8.1 Cochrane Authors' Background Characteristics.....	59
4.8.2 Authors' Interaction with Cochrane PLS Guidance.....	62
4.8.3 Authors' Future Simplification Preferences.....	69
4.8.4 SUS Scores.....	74
4.9 Discussion and Summary of the Study on Cochrane Authors' Satisfaction.....	76
5. ASSESSING THE READABILITY OF COCHRANE PLS AND ABSTRACTS....	81
5.1 Aims of the Study on Text Readability and Overview of the Chapter.....	81
5.2 Related Work on Readability.....	82
5.3 Motivation for Assessing the Readability of Cochrane PLS, Research Question, and Research Hypotheses.....	87
5.4 Experimental Materials.....	90
5.5 Method Adopted for the Measurement of Text Readability.....	93
5.5.1 Characteristics of Coh-Metrix and Rationale behind its Use.....	93
5.5.2 Coh-Metrix Measures.....	98
5.6 Data Analysis and Results.....	100
5.7 Discussion and Summary of the Study on the Readability of PLS and Abstracts.....	109
6. ASSESSING THE COMPREHENSIBILITY OF COCHRANE PLS AND ABSTRACTS.....	119

6.1 Aims of the Study on Comprehensibility and Overview of the Chapter.....	119
6.2 Related Work on Reading Comprehension of Health Content.....	120
6.3 Motivation for Assessing the Comprehensibility of Cochrane PLS, Research Question, and Research Hypotheses.....	124
6.4 Recruitment of Lay Readers.....	128
6.5 Experimental Environment, Procedure and Tasks.....	129
6.6 Experimental Materials and Experimental Design.....	134
6.7 Methods for Data Collection and Analysis on Text Comprehensibility.....	138
6.7.1 Characteristics of Text Retelling and Reasons for its Adoption.....	138
6.7.2 Procedure for the Analysis of Recall Protocols.....	142
6.8 Data Analysis and Results.....	149
6.8.1 Lay Readers' Screening and Background Characteristics.....	150
6.8.2 Free and Cued Recall.....	154
6.8.3 Ratings.....	165
6.9 Discussion and Summary of the Study on the Comprehensibility of PLS and Abstracts.....	168
7. ASSESSING THE MACHINE TRANSLATABILITY OF COCHRANE PLS.....	173
7.1 Aim of the Study on the Quality of Machine Translated Cochrane PLS and Overview of the Chapter.....	173
7.2 Related Work on MT in the Health Domain.....	174
7.3 Motivation for Assessing the Quality of Cochrane PLS Machine Translated into Spanish, Research Question, and Research Hypotheses.....	178
7.4 Recruitment of Cochrane MT Evaluators.....	182
7.5 Experimental Environment, Procedure and Tasks.....	183
7.6 Experimental Design and Experimental Materials.....	184
7.7 Method Adopted for Evaluating MT Quality.....	188
7.7.1 Overview of MT Quality Evaluation.....	188
7.7.2 Selection and Adoption of Adequacy and Fluency Measures.....	190

7.7.3 Rationale behind the Recruitment of Domain Experts and Implications	192
7.8 Data Analysis and Results	194
7.8.1 MT Evaluators' Background Characteristics	194
7.8.2 Quality of Cochrane PLS Machine Translated into Spanish	199
7.9 Discussion and Summary of the Study on the Machine Translatability of Cochrane PLS	211
8. CONCLUSIONS	215
8.1 Content and Organisation of the Chapter	215
8.2 An Overview of the Thesis: Goals, Findings, and Implications	215
8.2.1 Practical Implications	224
8.3 Contributions of the Thesis	225
8.3.1 Empirical Contribution	227
8.3.2 Methodological Contribution	230
8.3.3 Contribution to Knowledge of Practice	233
8.4 Limitations of the Thesis and Future Research	234
REFERENCES	241
Appendices	281
Appendix A: Research Ethics Committee Letter of Approval for Experiment on Cochrane Authors' Satisfaction	
Appendix B: Call for Participation Targeting Cochrane Authors of Plain Language Summaries	
Appendix C: Plain Language Statement, Informed Consent Form, and Pre-Task Questionnaire for Cochrane Authors	
Appendix D: Questionnaire for Cochrane Authors on Their Typical Interaction and Satisfaction with the Non-Automated Simplification Approach	
Appendix E: Instructions for Cochrane Authors on the Installation of TeamViewer	
Appendix F: Instructions for Cochrane Authors on Main Editing Task with Acrolinx	
Appendix G: Instructions for Cochrane Authors on Warm-Up Task with Acrolinx	

Appendix H: Post-Session Questionnaire for Cochrane Authors on Satisfaction Associated with Acrolinx and Future Editing Preferences.....

Appendix I: Ethical Approval for Reading Comprehension Study Received from Research Ethics Committee (Dublin City University) and Institutional Review Board (Arizona State University).....

Appendix J: Informed Consent Form and Background Questionnaire for Arizona State University Students Involved in Reading Comprehension Study.....

Appendix K: Instructions and Questions Submitted to Arizona State University Students Involved in Reading Comprehension Study.....

Appendix L: Prior Knowledge Questions Asked to Arizona State University Students after Reading Comprehension Study.....

Appendix M: Research Ethics Committee Letter of Approval for Experiment on Evaluation of Spanish Machine Translation Output.....

Appendix N: Call for Participation Targeting Cochrane Health Professionals (Native Speakers of Spanish).....

Appendix O: Plain Language Statement, Informed Consent Form, and Background Questionnaire for Cochrane Machine Translation Evaluators.....

Appendix P: Instructions and Questions for Machine Translation Evaluators.....

Appendix Q: Follow-Up Email Sent to Cochrane Machine Translation Evaluators.....

List of abbreviations

Ab	Abstract
AEMs	Automatic Evaluation Metrics
ANCOVA	Analysis of Covariance
ANOVA	Analysis of Variance
ASU	Arizona State University
AuPLS	Semi-Automated Plain Language Summary
BCG	Cochrane Breast Cancer Group
BLEU	Bi-Lingual Evaluation Understudy
CAT	Computer-Aided-Translation
CDC	Centers for Disease Control and Prevention
CEFR	Common European Framework of Reference for Languages
CFG	Cochrane Cystic Fibrosis and Genetic Disorders Group
CFP	Call for Participation
CHG	Cochrane Heart Group
CL	Controlled Language
CMS	Centers for Medicare and Medicaid Services
CWG	Cochrane Work Group
DCG	Cochrane Dementia and Cognitive Improvement Group
DCU	Dublin City University
DV	Dependent Variable
EVG	Cochrane Eyes and Vision Group
G-G	Greenhouse-Geisser
GMRT	Gates-MacGinitie Reading Test
GNG	Cochrane Gynaecological, Neuro-oncology and Orphan Cancer Group
GPHIN	Global Public Health Intelligence Network
GTM	General Text Matcher
H0	Null Hypothesis
H1	Alternative Hypothesis
HHS	Department of Health and Human Services
HimL	Health in My Language
HOCL	Human-Oriented Controlled Language
HSD	Honestly Significant Difference
HSE	Health Service Executive
HUME	Human Semantic Evaluation Measure
ICC	Intraclass Correlation Coefficient
ING	Cochrane Injuries Group
INTERACT	International Network on Crisis Translation

ISO	International Organization for Standardization
IV	Independent Variable
L1	First Language
L2	Second Language
LEP	Limited English Proficiency
MDG	Cochrane Common Mental Disorders Group
MOCL	Machine-Oriented Controlled Language
MT	Machine Translation
NHS	National Health Service
NIH	National Institutes of Health
NonAuPLS	Non-Automated Plain Language Summary
Npn	No page number
OVIX	Word Variation Index
PAHO	Pan American Health Organisation
PAHOMTS	Pan American Health Organisation Machine Translation System
PE	Post-Editing
PHAST	Public Health Automatic System for Translation
PL	Plain Language
PLEACS	Standards for the Reporting of Plain Language Summaries in New Cochrane Intervention Reviews
PLS	Plain Language Summary
PRISM	Program for Readability in Science and Medicine
QUARTET M	Qualité de l'Aide à la Rédaction et de la Traduction; Evaluation du Transfert d'information en Médecine
RevMan	Review Manager
RQ	Research Question
SD	Standard Deviation
STE	Simplified Technical English
STG	Cochrane Stroke Group
SUMI	Software Usability Measurement Inventory
SUS	System Usability Scale
TASA	Touchstone Applied Science Associates
TER	Translation Error Rate
T.E.R.A.	Text Ease and Readability Assessor
TQA	Translation Quality Assessment
UMUX	Usability Metric for User Experience
VAG	Cochrane Vascular Group

List of tables

Table 4.1: Descriptive statistics for Cochrane PLS guidance rankings.....	67
Table 4.2: Descriptive statistics for the SUS scores assigned by the total sample of participants.....	75
Table 4.3: Descriptive statistics for the SUS scores assigned by the 12 participants who conducted the Acrolinx editing task.....	75
Table 5.1: Descriptive statistics for text length (number of words), divided by corpus.....	92
Table 5.2: Descriptive statistics for narrativity scores divided by corpus.....	103
Table 5.3: Descriptive statistics for syntactic simplicity scores divided by corpus.....	104
Table 5.4: Descriptive statistics for word concreteness scores divided by corpus.....	105
Table 5.5: Descriptive statistics for referential cohesion scores divided by corpus.....	105
Table 5.6: Descriptive statistics for deep cohesion scores divided by corpus.....	106
Table 5.7: Descriptive statistics for L2 readability scores divided by corpus.....	108
Table 5.8: Descriptive statistics for Flesch-Kincaid Grade Level scores divided by corpus.....	109
Table 5.9: Descriptive and inferential statistics for measures analysed, per corpus.....	111
Table 6.1: Experimental design of reading comprehension experiment.....	137
Table 6.2: Descriptive statistics for lay readers' topic knowledge.....	153
Table 6.3: Descriptive statistics for participants' reading skills.....	154
Table 6.4: Descriptive statistics for lay readers' free recall scores.....	155
Table 6.5: Descriptive statistics for lay readers' cued recall scores.....	157
Table 6.6: Descriptive and inferential statistics for recall scores, per corpus and sample of participants.....	160
Table 7.1: Experimental design of MT evaluation experiment.....	187
Table 7.2: MT evaluators' (self-reported) level of English proficiency.....	198
Table 7.3: Descriptive and inferential statistics for adequacy scores.....	201
Table 7.4: Descriptive and inferential statistics for fluency scores.....	202
Table 7.5: Inter-rater agreement of MT evaluators (per group) on adequacy and fluency scores.....	210
Table 8.1: Contributions of this thesis.....	227

List of figures

Figure 1.1: Experimental variables in this thesis.....	9
Figure 4.1: Satisfaction as DV1.....	32
Figure 4.2: Acrolinx flagging readability/translatability issues in a sample text, and presenting suggestions and additional information in the sidebar in Microsoft Word...	38
Figure 4.3: Tasks assigned to Cochrane authors per session and collected evidence...	48
Figure 4.4: RevMan 5.3 interface.....	61
Figure 4.5: Cochrane PLS guidance provided to authors.....	62
Figure 4.6: Workflow of PLS production.....	63
Figure 4.7: Frequency of consultation of Cochrane PLS guidance.....	65
Figure 4.8: Rankings of completeness of Cochrane PLS guidance in terms of content.....	66
Figure 4.9: Rankings of completeness of Cochrane PLS guidance in terms of style....	67
Figure 4.10: Future use of authoring support for the production of PLS.....	70
Figure 5.1: Text readability as DV2.1.....	82
Figure 6.1: Comprehensibility as DV2.2.....	120
Figure 6.2: Tasks assigned to participants in reading comprehension study and collected evidence.....	133
Figure 6.3: Exemplification of links between ideas in incoherent (left) and coherent (right) mental representation.....	139
Figure 6.4: Regression slopes for free recall of native readers and reading skills.....	161
Figure 6.5: Regression slopes for free recall of non-native readers and reading skills.....	162
Figure 6.6: Regression slopes for cued recall of native readers and reading skills.....	163
Figure 6.7: Regression slopes for cued recall of non-native readers and reading skills.....	165
Figure 6.8: Ratings of text ease obtained from native readers.....	167
Figure 6.9: Ratings of text ease obtained from non-native readers.....	167
Figure 7.1: Machine translatability as DV2.3.....	174
Figure 7.2: Years MT evaluators had been reading English health-related texts.....	196
Figure 7.3: Hours (per month) spent by MT evaluators reading medical texts, on average.....	197
Figure 7.4: MT evaluators' frequency of use of MT systems.....	199

Abstract

Alessandra Rossetti

Simplifying, Reading, and Machine Translating Health Content: An Empirical Investigation of Usability

Text simplification, through plain language (PL) or controlled language (CL), is adopted to increase readability, comprehension and machine translatability of (health) content. Cochrane is a non-profit organisation where volunteer authors summarise and simplify health-related English texts on the impact of treatments and interventions into plain language summaries (PLS), which are then disseminated online to the lay audience and translated. Cochrane's simplification approach is non-automated, and involves the manual checking and implementation of different sets of PL guidelines, which can be an unsatisfactory, challenging and time-consuming task.

This thesis examined if using the Acrolinx CL checker to automatically and consistently check PLS for readability and translatability issues would increase the usability of Cochrane's simplification approach and, more precisely: (i) authors' satisfaction; and (ii) authors' effectiveness in terms of readability, comprehensibility, and machine translatability into Spanish.

Data on satisfaction were collected from twelve Cochrane authors by means of the System Usability Scale and follow-up preference questions. Readability was analysed through the computational tool Coh-Metrix. Evidence on comprehensibility was gathered through ratings and recall protocols produced by lay readers, both native and non-native speakers of English. Machine translatability was assessed in terms of adequacy and fluency with forty-one Cochrane contributors, all native speakers of Spanish.

Authors seemed to welcome the introduction of Acrolinx, and the adoption of this CL checker reduced word length, sentence length, and syntactic complexity. No significant impact on comprehensibility and machine translatability was identified. We observed that reading skills and characteristics other than simplified language (e.g. formatting) might influence comprehension. Machine translation quality was relatively high, with mainly style issues.

This thesis presented an environment that could boost volunteer authors' satisfaction and foster their adoption of simple language. We also discussed strategies to increase the accessibility of online health content among lay readers with different skills and language backgrounds.

*'Never use a foreign phrase, a scientific word or a jargon word
if you can think of an everyday English equivalent. [...]
These rules sound elementary, and so they are, but they demand a deep change
of attitude in anyone who has grown used to writing in the style now fashionable.'*

George Orwell (1903-1950)

CHAPTER 1

INTRODUCTION

1.1 Background, Motivation, and Scope of the Thesis

The Internet has become a widely consulted source of health information, particularly among lay¹ people (often patients) interested in answering their own health-related questions, making informed healthcare decisions, or improving the self-management of an illness (Hall, Stollefson and Bernhardt 2012; Basch et al. 2018; Stollefson et al. 2018). This widespread reliance on online resources for health-related purposes has been observed worldwide, and across different languages (Renahy, Parizot and Chauvin 2008; Novillo-Ortiz, Hernández-Pérez and Saigí-Rubió 2017). However, research has shed light on two issues related to online health information seeking. The first issue is represented by the difficulties that lay people often encounter when trying to read and comprehend medical² texts, characterised by specialised vocabulary, complex syntactic structures, and cohesion gaps (Lachance et al. 2010; Mičić 2013). The second issue affects non-native speakers of English, and particularly those with no or limited knowledge of English, as it has been observed that the majority of online health information is available in English only (Adams and Fleck 2015; Heilman and West 2015). To address these issues — thus in turn fostering the accessibility³ of online health content and reducing the vulnerability of lay readers (with different language backgrounds) — numerous organisations aim to simplify the language of medical texts and to translate them.

An example is represented by the Cochrane Collaboration (henceforth *Cochrane*), a non-profit organisation that relies on an international network of volunteer contributors for the preparation, maintenance, and dissemination of online systematic

¹ In line with Patel and Kaufman (2006, p. 152), we define a lay person as a person having “only common sense or everyday knowledge of a domain”.

² Drawing upon previous research showing the similarities between the terms *Health 2.0* and *Medicine 2.0* (Hughes, Joshi and Wareham 2008), in this thesis we use the terms *health/healthcare* and *medicine/medical* interchangeably.

³ Here we define *accessibility* as comprehensibility or understandability of content, in line with one of the four principles of the Web Accessibility Initiative (Rodríguez Vázquez 2016). The Web Accessibility Initiative website is available at: <https://bit.ly/2BtvMAY> [Accessed 12 December 2018].

reviews of studies on the effects of health interventions and treatments (Smith 2013). Each systematic review aims to address a specific research question (Green et al. 2011) — examples of systematic reviews are *Vaccines for Preventing Influenza in People with Asthma* (Cates and Rowe 2013), or *Pharmacological Interventions for Alcoholic Liver Disease* (Buzzetti et al. 2017). To produce systematic reviews, volunteer authors at Cochrane conduct the following main tasks: (i) systematic search of databases for medical studies dealing with the impact/effectiveness of a specific treatment or intervention; (ii) selection of eligible studies; (iii) assessment of the risk of bias in the included studies; (iv) combination of the statistical results from the included studies (known as meta-analysis); (v) discussion of potential bias; and (vi) presentation and interpretation of the results (Higgins and Green 2011). Cochrane Systematic Reviews are often produced by a team of contributors — this teamwork ensures that potential errors at the different stages of the authoring process are more easily identified (ibid.). Cochrane contributors have a health background, and belong to different review groups, each dealing with a specific area (*Cochrane Review Groups* 2018). Examples of Cochrane Review Groups are: Drugs and Alcohol Group; Infectious Diseases Group; and Skin Group.

By producing systematic reviews which are then published online on the Cochrane Library website⁴, Cochrane aims to promote informed decisions and evidence-based practice in healthcare (Smith 2013). Green et al. (2011, no page number [nnp]) explain the rationale behind systematic reviews:

Healthcare providers, consumers, researchers, and policy makers are inundated with unmanageable amounts of information, including evidence from healthcare research. It is unlikely that all will have the time, skills and resources to find, appraise and interpret this evidence and to incorporate it into healthcare decisions. Cochrane reviews respond to this challenge by identifying, appraising and synthesizing research-based evidence and presenting it in an *accessible* format. (Emphasis added)

As emerges from this quote, accessibility is part of Cochrane's mission, and is linked with the characteristics of Cochrane's target audience, which does not only include

⁴ The Cochrane Library website is available at: <https://bit.ly/2CrMUX3> [Accessed 12 December 2018].

health professionals, but also patients/lay people with no health (research) background. The organisation seems to adopt a view of accessibility that mainly coincides with the provision of comprehensible content for lay users, achieved either by simplification or translation. This view of accessibility is reflected in the document outlining Cochrane's strategy to 2020:

We will *simplify* and standardise the language used across our content to improve readability and reduce ambiguity. [...] We will *translate* key content into at least the five other official languages of the World Health Organisation (Spanish, French, Russian, Chinese, and Arabic); and make it *accessible* in the same way as English-language content. (*Strategy to 2020* 2013, pp. 14-15, emphasis added)

Plain language summaries (PLS) play a key role in Cochrane's accessibility strategy. Each systematic review is preceded by a PLS which both summarises and simplifies its content by using a plain language (PL). Summarisation and simplification are needed because systematic reviews are lengthy and characterised by specialised medical language (Harvey 2018). Moreover, PLS are translated from English (the language in which they are written) into multiple languages, such as Spanish, Croatian, or Japanese (*Knowledge Translation in Multi-Languages* 2018).

Chisholm and Henry (2005) point out that achieving web accessibility requires the cooperation and integration of technical components (e.g. authoring/evaluation tools) and human components (e.g. content producers). However, at Cochrane, the approach adopted for the production of accessible information in the form of PLS is non-automated⁵, and involves the manual checking and implementation of different sets of guidelines dealing with both content (summarisation) and language/style (simplification). These guidelines are sometimes characterised by contradictions and vagueness, their implementation is likely to depend on the authors' memory, and authors do not receive feedback that might improve their PL writing skills (Section 4.3).

⁵ As the data reported in Section 4.8.1 will show, Cochrane authors write PLS in either Microsoft Word or Review Manager, where they can avail of functionalities such as automatic spell checking or warning if the PLS exceeds 400 words. However, we treat their approach as non-automated/manual because none of the two software tools allows for the automatic checking of texts against the wide range of Cochrane guidelines, e.g. on the use of acronyms (Section 4.3). For a description of Review Manager, see Section 4.8.1.

Moreover, Cochrane PLS guidelines are spread across different documents, which is likely to make their checking difficult and time-consuming for authors (Temnikova 2012), particularly for volunteers with no linguistics background. Accordingly, the PLS resulting from this non-automated summarisation/simplification approach have shown inconsistencies, low readability (Kadic et al. 2016; Karačić et al. 2017), and reduced comprehensibility (Maguire and Clarke 2014; Santesso et al. 2015). Furthermore, there is a lack of empirical evidence on the extent to which the PL used in the PLS makes them (machine) translatable (Section 7.3).

Against this background, this thesis examines the *usability* of the *simplification* approach currently adopted at Cochrane for the production of PLS, and the impact of introducing the Acrolinx controlled language (CL)⁶ checker⁷ on the usability of the aforementioned approach. We adopt the definition of usability provided by the International Organization for Standardization (ISO 9241-11:2018, 3.1.1)⁸, namely the “extent to which a system, product or service can be used by specified users to achieve specified goals with *effectiveness*, efficiency and *satisfaction* in a specified context of use” (emphasis added). Specifically, we tested if and to what extent providing Cochrane authors with Acrolinx to check and edit the language in the PLS would increase: (i) their satisfaction; and (ii) their effectiveness in terms of readability, comprehensibility and machine translatability achieved in the PLS⁹ (Section 1.2). As the literature review in Chapter 2 will show, there is a dearth of studies on the usability of text simplification approaches for health content.

A detailed description of the Acrolinx CL checker used in this study is available in Section 4.3. Here we specify that this software: (i) checks texts against a predefined set of readability and translatability rules whose goal is to simplify texts and make them

⁶ In Section 3.2, we will delve into the notion of PL and its relationship with the broader notion of CL.

⁷ The name assigned by the Acrolinx company to their software is *content optimisation software/platform*. However, in line with scholarly tradition (e.g. Roturier 2009; Rodríguez Vázquez 2016), we call the software *CL checker*. The Acrolinx website is available at: <https://bit.ly/2AnXn4A> [Accessed 12 December 2018].

⁸ In Section 2.2, we will explain why the ISO 9241-11:2018 definition was deemed suitable for the purposes of our investigation.

⁹ The reason why the efficiency component of usability was not included in this investigation will be explained in Section 4.6.

more processable by both humans and machine translation (MT) systems; (ii) automatically and consistently flags issues in texts when these rules are contravened; and (iii) provides examples and suggestions on how to solve readability and translatability issues by editing the texts (Reuther and Schmidt-Wigger 2000). To signal the difference with the non-automated simplification approach currently used at Cochrane, we define text simplification conducted with Acrolinx as *semi-automated* — even though readability and translatability issues are automatically and consistently flagged, the author needs to manually apply or select the edits.

Regarding the scope of this thesis, our focus is on the simplification, rather than the summarisation, conducted at Cochrane as part of the workflow of PLS production (Section 4.3). This decision was taken because the sets of guidelines on summarisation (i.e. dealing with the content of systematic reviews that should be included in PLS) could not be formalised into a rule, and therefore could not be integrated into Acrolinx's error-type oriented approach to text checking (Bredenkamp, Crysmann and Petrea 2000).

The scope of this thesis is also limited to the health content produced at Cochrane. This non-profit organisation was chosen firstly because, as this section has shown, it represents a good example of an organisation relying on volunteers who aim for (multilingual) accessibility of online health content. Furthermore, as the initiatives reviewed in Section 3.3 will show, numerous organisations currently simplify health-related texts using a manual/non-automated approach similar to the one adopted by Cochrane. Therefore, despite the limited scope of this thesis, the findings could have implications for other environments (Section 8.3). It is also worth noting that Cochrane has developed a partnership with Wikipedia, currently one of the most widely consulted online sources of health information (Heilman and West 2015). The Cochrane-Wikipedia initiative involves using content from Cochrane Systematic Reviews (usually paraphrasing or summarising it) to improve the evidence base of Wikipedia's medical articles (Shafee et al. 2017). Similar to Cochrane, Wikipedia relies on volunteers for the production, maintenance, editing and translation of content (Heilman and West 2015). Moreover, like Cochrane, Wikipedia often needs to address the issue of readability of its

articles (Shafee et al. 2017), which led to the development of Simple English Wikipedia (Section 8.3). Therefore, despite limiting its scope to the usability of the simplification approach at Cochrane, this thesis might uncover issues and propose solutions that could also benefit Wikipedia’s volunteer contributors and lay users.

A final reason for selecting Cochrane is the fact that this organisation is one of the partners involved in the EU-funded project *International Network on Crisis Translation* (INTERACT)¹⁰, in which the author of this thesis and her supervisor are also involved (Section 1.3). As will be explained in Chapters 4 and 7, this collaboration facilitated the recruitment of participants (authors and evaluators), and provided us with training and valuable insights into the workflow of content production and evaluation at Cochrane, thus enhancing the ecological validity of our experiments.

1.2 Research Questions and Experimental Variables

As reported in Section 1.1, this thesis examines the *usability* of the *text simplification* approach currently adopted at Cochrane for the production of PLS, and the impact of introducing the Acrolinx CL checker on the usability of the aforementioned approach. Concretely, this thesis seeks to answer the following overarching research question (RQ), where usability represents the dependent variable (DV) and the text simplification approach adopted represents the independent variable (IV):

RQ: Does semi-automating a non-automated simplification approach by introducing a CL checker increase usability?

With regard to our IV, namely the “circumstance or characteristic that is manipulated or systematically controlled” (MacKenzie 2013, p. 161), this had two possible values or levels, respectively characterised by the absence or presence of the CL checker. In other words, the simplification approach could be either non-automated (prior to the introduction of the CL checker) or semi-automated (after the introduction of the CL checker). In the field of human-computer interaction (of which usability testing is a

¹⁰ The INTERACT website is available at: <https://goo.gl/NkLhxZ> [Accessed 12 December 2018].

component), the IV is usually linked with technology and users (Lazar, Feng and Hochheiser 2010).

As far as our DV is concerned, the ISO definition reported in Section 1.1 (ISO 9241-11:2018, 3.1.1) shows that usability is a broad concept in which different components (namely, satisfaction, effectiveness, and efficiency) can be identified (Hornbæk 2006). In line with previous works on usability in the domain of translation (Doherty and O'Brien 2012), this study divided the ISO 9241-11:2018 definition into its components, and analysed each of them separately. Therefore, our overarching RQ was segmented into the following RQs:

*RQ1: Does semi-automating a non-automated simplification approach by introducing a CL checker increase authors' **satisfaction**?*

*RQ2: Does semi-automating a non-automated simplification approach by introducing a CL checker increase authors' **effectiveness**?*

We define *satisfaction* as “extent to which the user’s physical, cognitive and emotional responses that result from the use of a system, product or service meet the user’s needs and expectations” (ISO 9241-11:2018, 3.1.14). *Effectiveness* is defined as “the accuracy and completeness with which users achieve certain goals” (ISO 9241-11:2018, 3.1.12).

As this latter definition shows, effectiveness coincides with goal completion. In order to achieve the goal of *accessibility* for lay users with different language backgrounds, Cochrane PLS need to be easy to read, comprehend, and (machine) translate (Section 1.1). Therefore, we treated readability, comprehensibility and machine translatability as goals of the simplification approach (for an explanation of the difference between readability and comprehensibility, see Section 5.2). Bevan, Carter and Harker (2015, p. 144) remark that “it is necessary to identify the goals and to decompose effectiveness, efficiency and satisfaction and the components of the context of use into sub-components with measurable and verifiable attributes”. In line with this observation, our RQ2 was further segmented into the following questions, each corresponding to a component of effectiveness/goal completion:

*RQ2.1: Does semi-automating a non-automated simplification approach by introducing a CL checker increase **readability**?*

*RQ2.2: Does semi-automating a non-automated simplification approach by introducing a CL checker increase **comprehensibility**?*

*RQ2.3: Does semi-automating a non-automated simplification approach by introducing a CL checker increase **machine translatability**?*

With a view to answering the aforementioned RQs, we conducted four different experimental studies. Experimental research involves the creation, control or manipulation of conditions or variables to determine if a causal relationship between two factors (i.e. an IV and a DV) exists (Lazar, Feng and Hochheiser 2010; Mellinger and Hanson 2017). The DVs, or observed components of usability, in our four experiments were:

DV1: Satisfaction

DV2.1: Readability

DV2.2: Comprehensibility

DV2.3: Machine translatability

By measuring the impact of the Acrolinx CL checker on DV1, DV2.1, DV2.2, and DV2.3, we aimed to answer the broader RQ on the impact of the Acrolinx CL checker on usability. In the interest of clarity, Figure 1.1 visually summarises the two levels of the IV and the usability components that represent the DVs of this thesis.

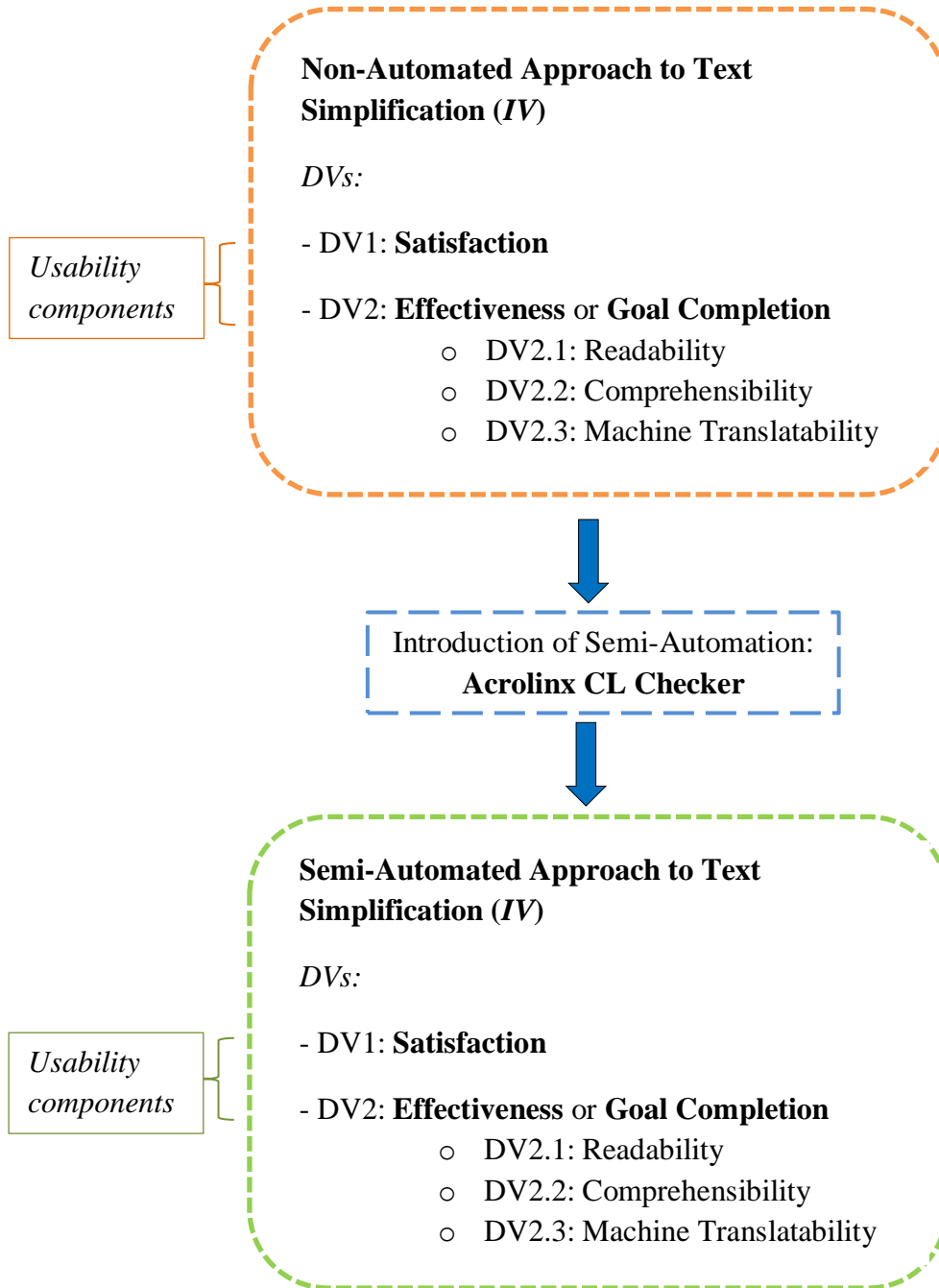


Figure 1.1: Experimental variables in this thesis

As a final remark, it is worth mentioning here that, when conducting the experimental studies on the impact of Acrolinx on readability (Chapter 5) and comprehensibility (Chapter 6), we also expanded our analysis to include Cochrane abstracts, namely non-simplified summaries of systematic reviews (Section 5.4). Abstracts were included in

the analysis to determine whether and to what extent text simplification (regardless of being non-automated or semi-automated) is beneficial in terms of readability and comprehensibility.

1.3 Framework and Type of Research

This research was conducted at the intersection of disciplines as diverse as accessibility, usability/human-computer interaction, health communication, computational linguistics, psycholinguistics, and translation quality assessment. Usability — i.e. our main DV (Section 1.2) — brings together the different areas, and represents the framework within which this investigation was conducted. Drawing upon Matthews and Ross (2010, p. 34), we describe a framework as a set of theoretical ideas and approaches that are adopted to view and collect knowledge. Specifically, usability represented the lens through which our findings were interpreted. As the discussions at the end of each experimental chapter (Chapters 4-7) and at the end of the entire thesis (Chapter 8) will show, findings from the four experiments were interpreted in terms of authors' satisfaction or effectiveness/goal completion, i.e. the usability components under investigation.

In Section 1.2 we described the research in this thesis as experimental. Here we specify that our research can also be treated as empirical since it “seeks new data, new information derived from the observation of data and from *experimental work*; it seeks evidence which supports or disconfirms hypotheses, or generates new ones” (Williams and Chesterman 2002, p. 58, emphasis added). According to MacKenzie (2013), the experimental methodology, along with the observational methodology and the correlational methodology, can be assigned to the broader category of empirical research. The author (*ibid.*, p. 129) goes on to specify that “empirical means *capable of being verified or disproved by observation or experiment*” (emphasis in original).

The research described in this thesis can also be classified as evaluative (Saldanha and O'Brien 2013), since it sought to evaluate the impact of introducing a CL checker into a text simplification approach. More precisely, our investigation involved a *summative* (rather than a formative) usability *evaluation*. Summative usability

evaluation is conducted against a set of criteria — in our study, satisfaction and effectiveness — after a product has already been released (Tullis and Albert 2013). Interestingly, satisfaction and effectiveness are also listed by Brajnik (2008) as the measures of summative accessibility evaluation (Section 1.1).

Hochheiser and Lazar (2007) argue that the field of usability/human-computer interaction has traditionally been concerned with practical results that could improve the quality of life. With its focus on practical results (described in Section 1.1), our investigation represents no exception and can therefore be assigned to the category of applied research (Saldanha and O’Brien 2013). In Section 1.1, we also mentioned that the research described in this thesis has been conducted as part of INTERACT, a EU-funded project focusing on communication in crisis and disaster scenarios. Specifically, our investigation was part of the work package on simplification of health content, which was included because crises and disaster often have a health component (Shiu-Thornton et al. 2007). In addition to Cochrane, one of the partners involved in the same work package was Arizona State University (ASU), where we conducted the study on comprehensibility (Chapter 6).

1.4 Thesis Structure

This introductory chapter has provided a high-level overview of this thesis by describing its background, motivation, and scope. The RQs and the experimental variables have also been presented. Moreover, this chapter has provided a classification of this research and an introduction to its usability framework.

Chapter 2 and Chapter 3 constitute the literature review analysing works that deal, respectively, with our main DV (i.e. usability) and our IV (i.e. text simplification) (Figure 1.1). In particular, Chapter 2 offers a literature review on usability and related concepts, followed by a special focus on the usability of authoring/editing environments and tools. Chapter 3 presents publications and initiatives dealing with text simplification, its different approaches, and its practical applications to health content. A review of the literature on the specific components of usability under investigation (i.e.

satisfaction, readability, comprehensibility, and machine translatability) is presented in Chapters 4-7, respectively.

Chapter 4 describes the first of our four experimental studies, which revolves around the impact of introducing the Acrolinx CL checker on the satisfaction of Cochrane authors of PLS. Chapter 5 deals with the readability of Cochrane PLS and abstracts, and presents the changes in text characteristics resulting from the use of Acrolinx. Chapter 6 describes the experiment aimed at collecting data on the comprehensibility of Cochrane PLS (before and after the introduction of Acrolinx) and abstracts. Chapter 7 outlines the experiment on the impact of Acrolinx on the machine translatability of Cochrane PLS.

Each of these four experimental chapters contains a short literature review section, followed by a section explaining the rationale behind the experiment. In each experimental chapter, we also describe: the experimental materials (i.e. texts) used; the method adopted; and the analysis of the data gathered. Each of these experimental chapters has a final section where the main findings are summarised and discussed. For the experiments involving human participants (i.e. the experiments reported in Chapters 4, 6, and 7), we also outline the recruitment of participants, along with the experimental design and the experimental environment, procedure, and tasks.

Finally, Chapter 8 provides an overview of the entire thesis by discussing and summarising its goals, findings, and implications. We also outline the main contributions of this work. Before concluding, we discuss the limitations and provide ideas for future research directions.

CHAPTER 2

REVIEW OF LITERATURE ON USABILITY

2.1 Content and Organisation of the Chapter

This chapter presents and discusses publications dealing with our main DV, i.e. *usability* (Section 1.2). Specifically, we will start by delving into the notion of usability, the different definitions that scholars have assigned to it, and its differences and similarities with related concepts. Subsequently, and in line with the goals of this thesis (Section 1.2), we will discuss works on the usability of tools/environments for content production (i.e. authoring and translation) and editing (e.g. by means of CL checkers). Finally, we will summarise the main implications emerging from this literature review.

Our goal with this chapter is to define the scope of the usability framework within which this research is being conducted (Section 1.3), and to identify the key concepts that underpin it. More specifically, we aim to identify a perspective on usability that is broad enough to be applicable to plain language (PL) guidelines and a controlled language (CL) checker, and that allows us to investigate multiple aspects of the interaction with a product.

2.2 Usability and Related Concepts

The term *usability* was originally coined in the 1980s to substitute the term *user friendly* (Bevan, Kirakowski and Maissel 1991). Generally speaking, *usability* can be defined as the “capability of being used” (Bevan, Carter and Harker 2015, p. 143). Usability is an important component in the broader area of human-computer interaction (Lazar, Feng and Hochheiser 2010). Carroll (2002, p. xxvii) argues that human-computer interaction

is about understanding and creating software and other technology that people will want to use, will be able to use, and will find effective when used. And the *usability* concept and the methods and tools to encourage it, achieve it, and measure it are now touchstones in the culture of computing. (Emphasis added)

Since the concept of usability can be approached from different angles, different definitions have been proposed. For instance, Bevan, Kirakowski and Maissel (1991)

report that that some definitions of usability derive from a product- and user-oriented view — i.e. they focus on the product’s characteristics and the user’s mental effort/attitude — while other definitions are more contextually oriented, as they derive from the idea that the usability level of a product can vary, depending on the users, the task, and the environment. In line with this view, Booth (1989) remarks that usability is determined by the complex interaction of a number of factors, such as task, user, and characteristics of a system. For instance, regarding the task, Issa and Isaias (2015) stress the importance of ease of learning and task match. Ease of learning/learnability is determined by the effort required by the user to familiarise themselves with and operate a system (we will discuss learnability of Acrolinx in Section 4.7.1). Task match refers to the “extent to which the information and functions that a system provides matches the needs of the user” (Booth 1989, p. 107).

In addition to changing depending on the view adopted, the concept of usability varies in terms of the components or principles that have been assigned to it. Krug (2014) points out that the umbrella adjective *usable* often includes traits as different as usefulness, learnability, effectiveness, efficiency or desirability. Rubin and Chisnell (2008, p. 4) argue that “[t]o be usable, a product or service should be useful, efficient, effective, satisfying, learnable, and accessible”. Other scholars identify an overlap between usability and ease of use (Nielsen 2012), or between usability and the ability to conduct a task naturally (Dix et al. 2004). Dix et al. (ibid.) also discuss flexibility, robustness, and learnability as the three main principles of usability, and present the sub-principles underlying each one of them. For example, robustness includes: observability; recoverability; responsiveness; and task conformance.

Despite the different components that have been assigned to the notion of usability, and the different perspectives that have been adopted to define this notion, there seems to be some agreement that the user’s ability to conduct a task effectively and efficiently when using a product or a system is as important as their willingness to use them (Dix et al. 2004, p. 156). Krug (2014, p. 9) tries to condense and simplify the different views on usability by stating that something is usable when “[a] person of

average (or even below average) ability and experience can figure out how to use the thing to accomplish something without it being more trouble than it's worth”.

While Krug's (ibid.) definition has the advantage of being short and simple, it lacks specificity as for the criteria against which usability should be evaluated. In contrast, the ISO 9241-11:2018 definition of usability has the advantage of summarising the different views and components of usability, while also specifying its evaluation criteria. In particular, this ISO definition includes what Issa and Isaias (2015) describe as *performance measures* (related e.g. to the success and effort involved in a task) and *preference measures* (e.g. users' opinions). We repeat the ISO definition here: “extent to which a system, product or service can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use” (ISO 9241-11:2018, 3.1.1).

In addition to being comprehensive and defining measurable outcomes (Bevan, Carter and Harker 2015), the ISO 9241-11:2018 expands its focus to include products, systems and services, with *system* defined as “combination of interacting elements organized to achieve one or more stated purposes” and “all of the associated equipment, facilities, material, computer programs, firmware, technical documentation, services and personnel required for operations and support to the degree necessary for self-sufficient use in its intended environment” (ISO 9241-11:2018, 3.1.4). With its broad definition of what could represent the object of a usability study, this ISO definition was deemed particularly suitable for the purposes of our investigation, which focused on products/systems as diverse as documentation containing PL guidelines and a CL checker (Section 1.1).

As a final remark on *usability*, this term is often used interchangeably with *user experience* and *accessibility*¹¹. Here we highlight the differences and similarities between these terms with a view to clarifying the framework within which our investigation was conducted (Section 1.3). Regarding user experience, Tullis and Albert (2013) and Bevan, Carter and Harker (2015) point out that its focus is on the user's

¹¹ Here we refer to the accessibility of the product, service, or system (e.g. a CL checker) used to achieve a goal, rather than the accessibility/comprehensibility of texts resulting from that usage, as discussed in Chapter 1.

emotional experience (e.g. their thoughts and feelings) and motivations, rather than their effectiveness and efficiency. There are however some commonalities between user experience and the satisfaction component of usability, since they both deal with personal factors in the interaction with a system. Unsurprisingly then, the definition of satisfaction in ISO 9241-11:2018 has been expanded to also account for user experience (Bevan, Carter and Harker 2015). With regard to the relationship between accessibility and usability, Petrie and Kheir (2007, p. 397) argue that accessibility could be regarded as “usability for people with disabilities”. Similarly, Thatcher et al. (2003) state that usability issues seem to affect all users (both disabled and non-disabled), while accessibility problems are particularly detrimental to the interests of disabled people. In other words, accessibility issues seem to be regarded as a subset of usability issues.

2.3 Usability of Tools/Environments for Content Production and Editing

Our searches of publication databases have shown that there are no previous studies on the usability of PL guidelines (both in general and for health content). This result, which provides further motivation for our investigation, is surprising when considering that numerous organisations rely on written sets of PL guidelines for the simplification of health-related texts (Section 3.3).

There is also a dearth of empirical evidence on the usability of CL checkers applied to health content. However, a study conducted by Miyata et al. (2017) investigates the usability of a CL authoring assistant employed for the production of machine translatable Japanese municipal texts. Similarly to our investigation, Miyata et al. (ibid.) adopted the ISO 9241-11 definition of usability (Section 2.2), measured authors’ satisfaction using the System Usability Scale (SUS) (Section 4.7), treated machine translatability as a component of effectiveness (Section 7.1), and used human evaluation of the MT output (Section 7.7.2). However, differently from our investigation, Miyata et al. (ibid.) did not evaluate the readability and comprehensibility of the source texts, and conducted their experiment with statistical, rather than neural, MT systems (Section 7.3). Finally, the language and the domain of their source texts were also different (Section 4.6).

A study conducted by Thomas et al. (2015) is also relevant to our investigation since it addresses: (i) the acceptance of a CL specifically developed for health professionals; and (ii) the improvement of a writing assistant software prototype (Prolipsia CL Authoring Software). The authors (ibid., p. 96) explain the rationale behind their study as follows:

[h]ealth professionals were influenced by the style of the medical writing in scientific publications they were used to reading. They were rarely able to organize the pieces of information in a complete, *simple*, hierarchical and unambiguous way. (Emphasis added)

Thomas et al. (ibid.) also specify that their authoring software might reduce the workload of writers because they will not have to rely on their memory of simplification rules, which might also be an issue with Cochrane sets of PLS guidelines (Section 4.3). Participants in Thomas et al. (ibid.) found the Prolipsia software to be good, particularly because it reduced the time needed to select a recommended term/sentence/phrase while at the same time allowing authors to have control over the writing task — the author was alerted when a CL rule was contravened, and then guided through different steps to decide if and how the non-compliance to the CL should have been fixed. Despite the similarities with our investigation, Thomas et al. (ibid.) conducted their study with French (rather than English) texts (Section 4.6). Moreover, their software was used to write texts from scratch, rather than checking their compliance with CL rules and editing them accordingly (as the authors in our study did with the Acrolinx CL checker) (Section 4.5). Finally, the authors (ibid.) did not evaluate the impact of using their software on text characteristics such as readability (Section 5.7), comprehensibility (Section 6.9), or machine translatability (Section 7.9).

Further discussion on the relationship between usability and CL can be found in Mitamura and Nyberg (2001, p. 3), who argue that

when deploying a controlled language, author usability and productivity are very important factors for CL acceptance. If CL is too strict and/or time-consuming, then authors may have difficulty using it *effectively*. (Emphasis added)

This remark from Mitamura and Nyberg (*ibid.*) seems to suggest that authors' willingness to use a CL might be correlated with their effectiveness.

Khodambashi and Nytrø (2017) conducted a systematic review of literature on the evaluation of software tools for the authoring of clinical guidelines, namely statements that guide health professionals and patients in decisions regarding appropriate care (Shiffman et al. 2012). The software tools reviewed have a variety of functionalities — from the development of clinical guidelines to their dissemination (Khodambashi and Nytrø 2017) — and have been evaluated along a variety of dimensions, including usability (Shiffman et al. 2012) and tool performance (Rathbone, Hoffmann and Glasziou 2015). In particular, Shiffman et al. (2012) describe the usability evaluation of BRIDGE-Wiz, a wizard that guides authors in the development of clinical guidelines by providing them with a template relying on a CL approach that prompts authors to use, among others, transitive verbs and active voice.

We have also identified several publications which are more loosely related to the objects of study in our investigation (Section 1.2). In particular, these works have looked, among others, at the usability of authoring tools/environments for: the writing of policies and instructional/educational content (Reeder et al. 2007; Dağ, Durdu and Gerdan 2014; Gordillo, Barra and Quemada 2017); the development of serious games that support learning (Slootmaker, Hummel and Koper 2017); the fostering of game design skills in children (Yatim 2008); the creation of a limited domain communication scenario between deaf people and hearing people, particularly in medical contexts (Duma et al. 2015); or the writing of narratives to promote preparation for seismic events (Gaeta et al. 2014). Interestingly, Murray (2016) points out that, when designing authoring tools, there is a trade-off between, on the one hand, usability (defined as efficiency and ease of use) and, on the other hand, power, defined as the flexibility, breadth and depth of the tool — e.g. the extent to which a tool can support different domains and types of content.

Several of these usability studies have adopted the same ISO definition that was selected for our investigation (Section 1.1), thus focusing on the three components of effectiveness, efficiency, and satisfaction. However, these researchers often expanded

their analysis to include other measures of usability. For example, Yatim (2008) also focused on the fun experienced by children when using the game authoring tool; and Slootmaker, Hummel and Koper (2017) analysed, among other aspects, learnability, operability, user error protection, and user interface aesthetics. In terms of methodology, it is worth mentioning that several studies (e.g. Gaeta et al. 2014; Gordillo, Barra and Quemada 2017) collected data on users' satisfaction by means of the SUS, as we did in our investigation (Section 4.7).

Although not explicitly addressing usability or evaluation, Lee et al. (2005) describe the design and implementation of an authoring system, based on Synchronized Multimedia Integration Language, for multimedia health content. Similarly, Di Marco et al. (2006) describe the development of a tool for the automatic authoring and tailoring of health education materials based on a Natural Language Generation system. However, their tool was intended for the programmer that would develop the system, rather than the authors/end users.

Finally, it should be mentioned that the usability of Computer-Aided-Translation (CAT) tools has also been investigated and discussed in recent publications. These works are reviewed here because, similar to the simplification/editing/authoring process of specialised medical texts, translation can be assigned to the broader category of the "process chain of specialised communication" (Schubert 2007, quoted in Krüger 2016, p. 120). Krüger (2016) developed a usability model based, again, on the ISO 9241-11 definition of usability and the additional component of learnability. Although the main focus of this model was on translation memory systems, the author (*ibid.*) argues that it could be used as a starting point for empirical research on the usability of CAT tools in general. O'Brien et al. (2017) conducted a survey among professional translators with a view to collecting evidence on CAT tool features regarded as irritating. Their results showed that professional translators still find CAT tool irritating because of factors such as the complexity of the user interface and segmentation (e.g. segmented view of a text).

2.4 Conclusions Based on this Literature Review

This review of the literature on usability and on tools/environments for content production and editing has two main implications for this thesis. First of all, it confirmed a research gap to be filled by shedding light on the lack of empirical evidence on the usability of PL guidelines and CL checkers when applied to health-related texts with a view to simplifying them. Secondly, this review informed the framework of this thesis by helping us identify the ISO 9241-11:2018 definition of usability as particularly suitable for the purposes of our investigation due to its specificity and breadth. In other words, the ISO definition explicitly mentions and defines both the subjective and objective components of usability (Section 1.2), while also accounting for a broad range of products, systems, and services that might represent the object of a usability investigation. The next chapter will review the literature and practical applications related to our IV, i.e. text simplification.

CHAPTER 3

REVIEW OF LITERATURE ON TEXT SIMPLIFICATION

3.1 Content and Organisation of the Chapter

This chapter presents and discusses publications and initiatives dealing with our IV, i.e. *text simplification* (Section 1.3). Concretely, we will start by defining text simplification, outlining its goals, and describing the different approaches that have been adopted to simplify texts. Then, in line with the goals of this thesis (Section 1.2), we will discuss how simplification is usually applied to health-related texts produced by different organisations. Finally, we will summarise the main implications emerging from this literature review.

Our main goal with this chapter is to clarify the type of text modifications that we will treat as *simplification*, by distinguishing them from other forms of text alteration (e.g. summarisation or elaboration). Defining text simplification in the health domain will also allow us to further define our objects of study and the scope of our investigation.

3.2 Text Simplification and CL: Definitions, Goals, and Approaches

As argued in Mitamura and Nyberg (2001), texts are often characterised by complex and ambiguous language that might hinder their processing by both humans and computer applications (e.g. MT systems). Text simplification addresses this issue as it involves the modification of natural language aimed at increasing its readability, comprehensibility¹², and machine translatability (Shardlow 2014; Štajner and Popović 2016). Siddharthan (2014) also specifies that, in the process of text simplification, the meaning and information of the original text should be retained. The author (ibid.) points out that a broad view of simplification should also include other types of text modifications, such as summarisation (i.e. the deletion of irrelevant or peripheral content) and elaboration/explicitation. However, for the purposes of this thesis, we adopted a narrower view of simplification, which allowed us to distinguish between, on the one hand, the edits made by Cochrane authors on language/style (simplification) and, on the

¹² For an explanation of the difference between readability and comprehensibility, see Section 5.2.

other hand, the edits they made on content (summarisation) when producing PLS (Section 4.6).

Text simplification addresses a wide audience, which includes first language (L1) readers with reduced literacy (e.g. caused by dyslexia), non-native speakers/second language (L2) learners, or children (Siddharthan 2014). As explained in Bingel (2018, p. 8),

text simplification serves the central purpose of promoting accessibility of written language for people who would otherwise not be able to understand it fully or be able to do so only to some degree, or who would have to invest excessive amounts of energy to do so.

Simplifying texts has also been shown to increase the quality of MT output (Aikawa et al. 2007; Štajner and Popović 2016) and to reduce post-editing¹³ (PE) effort (O’Brien and Roturier 2007), particularly for technical documents.

Simplifying already existing complex texts or writing simple texts from scratch has traditionally involved the adoption of sets of guidelines/rules resulting in a controlled language (CL) (Temnikova 2012). Kuhn (2014, p. 123) defines a CL as “a constructed language that is based on a certain natural language, being more restrictive concerning lexicon, syntax, and/or semantics, while preserving most of its natural properties”¹⁴. The author (ibid.) specifies that CLs have also been described as *simplified* or *basic* languages, among others. In line with the aforementioned goals of text simplification (i.e. facilitating processing of information for humans and computers), CLs can be classified as Human-Oriented Controlled Languages (HOCLs), aiming at improving comprehension by humans, or Machine-Oriented Controlled Languages (MOCLs), whose goal is to facilitate text processing by computers (Huijsen 1998). Temnikova (2012) also reviews mixed-purpose CLs, developed with the twofold objective of being both human-oriented and machine-oriented. An example of a mixed-

¹³ In line with Allen (2003), we define PE as the editing and/or correction of MT output.

¹⁴ A clarification is needed regarding the difference between CL and sublanguage. While the former is artificially restricted through sets of rules, the latter emerges spontaneously from highly specialised communication, e.g. in the medical domain (Kittredge 2003, quoted in Kuhn 2014; Doing-Harris et al. 2013).

purpose CL is ASD Simplified Technical English (ASD STE), which is applied to aircraft documents and is based on around 60 rules (ibid.; Kuhn 2014).

The guidelines/rules that lead to a CL can be varied. As reported in Nyberg, Mitamura and Huijsen (2003, p. 245), “there is no single CL, say for English, which is approved by some global authority”. In her analysis of eight sets of rules for controlled English, O’Brien (2003) found that only one rule (dealing with sentence shortening) was shared by all sets, and that only seven rules were shared by the majority of rule sets. The author (ibid.) also observed that CL rules addressed various text characteristics: from lexicon (e.g. avoidance of polysemy), to syntax (e.g. restriction of the size of noun cluster), textual structure (e.g. reduction of sentence length), and pragmatics (e.g. avoidance of slang). Other examples of rules are reported in Nyberg, Mitamura and Huijsen (2003), and they address, for example, the maximum number of words allowed in a sentence, the avoidance of passive voice, or the use of bulleted lists.

Kuhn (2014) provides a comprehensive review and classification of 100 existing English-based CLs, from Ogden’s (1930) Basic English to Attempto Controlled English (Fuchs, Kaljurand and Kuhn 2008). Such comprehensive review and classification is outside the scope of this thesis. However, for the purposes of our investigation, it is worth pointing out that Kuhn (2014) and Cardey, Greenfield and Wu (2004) list plain language (PL) among the existing CLs if PL guidelines aim at defining a constructed/restricted natural language. This is the case for Cochrane PL guidelines, which will be further described in Section 4.3. For instance, the *Standards for the Reporting of Plain Language Summaries in New Cochrane Intervention Reviews (PLEACS)* contain a list of dos and don’ts dealing with, among others, the type of vocabulary (e.g. medical jargon) and syntactic structures (e.g. passive voice) to be avoided (The Cochrane Collaboration 2013). Therefore, in line with other scholars, we treated Cochrane PL guidelines as a form of CL¹⁵.

As a further remark on the relationship between CL and PL, Kuhn (2014) specifies that the main goal of PL guidelines is to facilitate comprehension, rather than

¹⁵ The PL guidelines developed for health-related texts and described in Section 3.3 can also be assigned to the broader category of CLs.

machine translatability (Section 3.3). In other words, bearing in mind that CLs can be classified as HOCLs or MOCLs, a PL is usually a HOCL. Stableford and Mettger (2007, p. 75) define the adoption of PL as “using evidence-based standards in structuring, writing, and designing to create *reading ease*” (emphasis added). Interestingly, we observed a lack of focus on (machine) translatability also in Cochrane PL guidelines (Section 4.3). This only partial overlap in goals between CL and PL is likely to be due to historical reasons. More precisely, while CLs were often developed by industries which had to translate their technical documents in order to reach out to an international audience (Rychtycky 2002), the PL movement began in the United States in the 1940s with a view to reducing the use of unclear and pretentious language in communications from the government to industries and the public, including limited English proficiency (LEP) readers (Schriver 2017).

There are different approaches to text simplification, from manual/non-automated, to semi-automated, to fully automatic. Manual simplification involves the editing of texts without any technological support, and is common in educational environments, where teachers often adapt texts to match the skills of their students (Petersen and Ostendorf 2007; Candido et al. 2009). Manual simplification is also adopted by numerous organisations which produce health-related texts in PL, including Cochrane (Sections 3.3 and 4.3). As a result of the time commitment and effort that manual simplification involves, more automated approaches have been developed.

CL checkers like the Acrolinx software in our authoring study can be described as a semi-automated approach to simplification because, despite automatically flagging readability and translatability issues, “these checkers [...] cannot do the hard work and transform a non-STE compliant text automatically into a compliant one” (Schwitter 2015, p. 453). Nyberg, Mitamura and Huijsen (2003, pp. 252-253) provide a detailed description of what CL checkers are and how they work:

CL checkers are programs which assist authors in determining whether their text complies with the specification of a CL. This assistance is generally given as a series of **critiques** or issues that are raised with respect to the text, communicated to the user as text messages by

the software. [...] In addition to pointing out violations of the CL, a checker may also offer help in the form of proposed **corrections**. (Emphasis in original)

In addition to Acrolinx (which will be described in Section 4.3), examples of other CL checkers are: MAXit Checker¹⁶, developed by Smart Communications Inc. (Kuhn 2014); the Boeing Simplified English Checker¹⁷ (Nyberg, Mitamura and Huijsen 2003); or the Controlled Automotive Service Language Checker¹⁸, developed at General Motors (Godden 2000). Most CL checkers are applied to technical texts. However, Acrolinx has already been applied to health content in (Simple English) Wikipedia articles (Ojala 2013; Azzam et al. 2017), which makes it particularly relevant for our investigation. Scarton et al. (2010) describe SIMPLIFICA, a web-based authoring tool developed for the lexical and syntactic simplification of Portuguese texts. Interestingly, this tool also allows authors to run a readability assessment of the texts based on Coh-Metrix measures (Section 5.5.1). Yimam and Biemann (2018) describe Par4Sim, a simplification tool that resembles a text editor and shows authors suggestions on how to simplify difficult words and phrases (that are automatically flagged). The authors (ibid.) specify that the tool, available on the Amazon Mechanical Turk crowdsourcing platform, can learn from the author's edits.

With regard to automatic text simplification systems, a comprehensive review is provided in Siddharthan (2014). The author (ibid.) explains that systems for automatic text simplification can be either rule-based, data-driven, or hybrid. In particular, data-driven approaches use large parallel corpora of complex texts and their simplified counterparts in order to develop monolingual MT systems. One of the most widely used parallel corpora is Wikipedia and its simplified version, Simple English Wikipedia, whose editors use a modified version of Ogden's Basic English (Coster and Kauchak 2011; Schwitter 2015). Despite the advantage of automatically delivering a simplified text, automatic text simplification systems entail several limitations. In particular, similar to interlingual MT outputs, intralingual MT outputs might be flawed, e.g. as a

¹⁶ Information on MAXit Checker is available at: <https://bit.ly/2U8JS0o> [Accessed 12 December 2018].

¹⁷ Information on the Boeing Simplified English Checker is available at: <https://bit.ly/2zHoEhu> [Accessed 12 December 2018].

¹⁸ The Controlled Automotive Service Language Checker seems to be no longer in use.

result of long sentences (Aluísio et al. 2008). Moreover, automatic text simplification mainly focus on vocabulary, syntax, and explanation generation, while not addressing cohesion (Siddharthan 2014; Shardlow 2014), which is one of the text characteristics affecting comprehension (Section 6.2). In addition, as far as we are aware, the automatic simplification systems described in academic publications are often not freely available to the public.

While manual simplification can take place at either the authoring stage (i.e. when producing a text from scratch) or at the editing stage (i.e. when simplifying an already existing text), semi-automated and automatic approaches to text simplification tend to be applied to already existing complex texts. For example, in relation to CL checkers, Nyberg, Mitamura and Huijsen (2003, p. 246) state that “CL can be used with software which performs a complete check of each new text to verify conformance”. Regarding our authoring study, the Acrolinx CL checker was used on already existing texts that had been simplified with a manual approach (Section 4.5). However, a few tools for CL authoring (rather than checking) have also been developed — in Section 2.3, we reported on authoring software that helps authors write CL-conformant texts from scratch (Thomas et al. 2015). Similarly, Max (2006) describes an interactive text simplification system whereby authors are shown simplified rephrasing that they can accept or modify when producing texts for readers with language impairments (e.g. aphasia). Another tool, called ProphetMT, is described in Wu et al. (2016) — this system guides authors in the application of a CL by showing auto-suggestions during the composition of in-domain sentences that are then translated by a statistical MT system.

Finally, it should be mentioned that the majority of research and practical applications of CLs have focused on English. As reported in Spaggiari, Beaujard and Cannesson (2003, p. 152), “English is a very productive natural language for CLs’ creation as it is the current international language used for trade and science”. Similarly, Renahy et al. (2009) explain that CLs are traditionally developed for intercultural communication, which usually takes place with English as a lingua franca, hence the prominence of CL for English. However, it is worth mentioning that CLs have also been developed for other languages, such as German, Chinese, Swedish, and French

(Spaggiari, Beaujard and Cannesson 2003). Additionally, research has been done on automatic text simplification systems for a variety of languages, e.g. Bulgarian, Portuguese, Korean, and Italian (Siddharthan 2014). However, as Siddharthan (ibid.) points out, languages other than English cannot avail of corpora of simplified texts as large as Simple English Wikipedia.

3.3 Simplification of Health-Related Texts

As discussed in Section 3.2, the main rationale behind the adoption of PL (which is a form of CL) is to facilitate comprehension, while the broader category of CL often has the additional goal of increasing machine translatability. Unsurprisingly then, when it comes to the simplification of health-related texts, the term *CL* is traditionally adopted when the simplification of health content aims at facilitating both its comprehensibility and machine translatability (Cardey, Greenfield and Wu 2004; Renahy et al. 2015), while the term *PL* seems to be used when the goal of simplification is to increase comprehensibility of health content, particularly among individuals with low health literacy¹⁹ (Rudd et al. 2004; Grene, Cleary and Marcus-Quinn 2017). The body of literature on PL for health information is vast, and numerous initiatives have been undertaken to simplify health content through PL (Schriver 2017). This prominence attributed to the comprehension (rather than the machine translatability) of health-related texts explains the lack of empirical evidence on the effects of CL on the quality of MT output of health content, as will be discussed in Section 7.3. Below we report on some of the initiatives²⁰ aimed at applying PL to health-related texts, their guidelines/rules, and the simplification approach adopted. Our main focus will be on health content in English.

¹⁹ 250 different definitions of health literacy have been identified (Malloy-Weir et al. 2016). In this thesis, we adopt the definition in Berkman, Davis and McCormack (2010, p. 16), namely “[t]he degree to which individuals can obtain, process, understand, and communicate about health-related information needed to make informed health decisions”. We selected this definition for its focus on understandability (or comprehensibility), and for its link between *understandability* and *informed* health decisions. Other definitions mention *appropriate* health decisions instead — as Malloy-Weir et al. (2016) point out, the degree to which a health decision is appropriate might depend on the point of view.

²⁰ Cochrane’s approach to text simplification will be discussed in detail in Section 4.3. Here we deal with the simplification initiatives undertaken by other organisations that share the same mission as Cochrane, i.e. providing lay readers with accessible health content (Section 1.1).

Schraver (2017) describes how health literacy and PL gained prominence in the United States starting from the early 2000s, with the US Department of Health and Human Services (HHS) applying PL on one of their websites²¹. The PL efforts of the HHS are still ongoing across its various Operating and Staff Divisions, including the Centers for Disease Control and Prevention (CDC), the Centers for Medicare and Medicaid Services (CMS), and the National Institutes of Health (NIH) — in the HHS's (2018) *Plain Writing Act Compliance Report*, it is stated that training on PL writing is provided to staff members, and that numerous agencies expect their employees to attend additional agency-specific training. Moreover, training materials are provided. For example, the CDC (2016) published the *Everyday Words for Public Health Communication*, in which federal employees and contractors are shown examples of words and sentences rewritten in PL to be used as examples. Similarly, the CMS elaborated style rules and checklists to guide their staff in their PL writing activities, such as the *Toolkit for Making Written Material Clear and Effective* (CMS 2012). It is worth mentioning that a section of this toolkit deals with translation, and it briefly mentions that, by following the guidelines for PL writing, translation quality might improve. However, their focus is exclusively on human translation.

Other initiatives and resources for PL writing are discussed in Gilliver (2015). These include: *How to Write Medical Information in Plain English*, developed as a result of the Plain English Campaign (2011) for workers and companies in the health sector; and the Program for Readability in Science and Medicine (PRISM) Toolkit of the Group Health Research Institute (2009). The PRISM Toolkit was developed with a view to making information on clinical trials comprehensible for participants (ibid.). In Ireland, the Health Service Executive (HSE) published a style guide and checklist whose goal is to guide health professionals in the production and checking of their written and spoken communications in PL (HSE 2017). In the United Kingdom, a recent initiative of the Academy of Medical Royal Colleges is encouraging doctors to write emails and letters directly to their patients (rather than to their GPs), and to simplify their language (Campbell 2018).

²¹ The HHS website in PL is available at: <https://bit.ly/2s4B8we> [Accessed 12 December 2018].

Guidelines/rules often mentioned in these training documents deal with vocabulary (e.g. avoidance of medical jargon); syntax (e.g. use of active voice and of short sentences); organised content structure; and avoidance of unnecessary information. The importance of cohesion is also mentioned occasionally (CMS 2012). Warde et al. (2018, p. e54) provide a comprehensive summary of the main characteristics that a text written in PL should have:

Plain language is defined as communication that can be understood the first time it is seen or heard, that uses succinct active-voiced grammatically correct complete sentences to better enable patients and caregivers to engage with information, using a more informal tone and common terms whenever possible.

For none of these initiatives aimed at simplifying health-related texts there is a mention of technological assistance or support for authors — similarly to the simplification approach currently adopted at Cochrane (Section 4.3), the implementation of PL on health-related texts at these organisations is also non-automated. An exception seems to be the adoption of the Acrolinx CL checker to help Wikipedia volunteers comply with Simple English Wikipedia guidelines when simplifying health content (Acrolinx 2012). Interestingly, reporting the results of a study conducted among health communication professionals from the US government, Harper and Zimmerman (2009, npn) wrote that:

(1) Many guidelines are open to interpretation by the writer/editor (some of the NIH guidelines are vague; such as “write simply and clearly” — which can result in different interpretations as proven in this exploratory study). (2) Even within an organization, communicators can have very different ideas on what it means to write and edit in plain language. Therefore, the variance could be even more extreme when compared across organizations.

This remark is particularly relevant to the purposes of our investigation since it sheds light on the inconsistencies that can emerge from a non-automated approach to text simplification, where implementation and interpretation of guidelines/rules is deeply influenced by the medical authors’ subjectivity (Section 4.3). It also shows that, while

beneficial (Warde et al. 2018), training authors with a health background on PL writing might not be enough.

3.4 Conclusions Based on this Literature Review

This review of the literature and initiatives dealing with simplification (of health content) has several implications for this thesis. First of all, reviewing the literature helped us clarify the relationship between CL and PL — the identification of their differences and similarities was particularly relevant for the purposes of this thesis since our authors were asked to semi-automatically apply a CL to texts produced following PL guidelines (Section 4.5). Secondly, this review also clarified the differences between text simplification and text summarisation, thus allowing us to make a distinction between the edits made at the level of content (as part of a summarisation task) and the edits made at the level of language/style (as part of a simplification task) by Cochrane authors of PLS (Section 4.6). Thirdly, the need to provide authors and editors of health content with technological assistance emerged. This need, which is also discussed in Sections 4.2 and 4.3, is part of the rationale behind this investigation on the impact of introducing a CL checker into a manual simplification approach. Finally, this literature review shed light on the practical applications of text simplification, thus helping us categorise this research as applied (Section 1.3).

To conclude, this chapter and Chapter 2 respectively reviewed works and initiatives linked with the two main experimental variables of this thesis, i.e. text simplification (the IV) and usability (the main DV). These two chapters also allowed us to define the scope of our investigation in relation to the concepts of usability and text simplification. The following four experimental chapters (Chapters 4-7) will also present a review of related work on the specific components of usability under investigation (Section 1.2). The next chapter contains the first of our four experimental studies, dealing with the satisfaction of Cochrane authors.

CHAPTER 4

ASSESSING THE SATISFACTION OF COCHRANE AUTHORS

4.1 Aims of the Study on Authors' Satisfaction and Overview of the Chapter

This chapter describes an experimental study which was conducted with two main aims. The first aim was to gain a deeper understanding of Cochrane authors' typical interaction and satisfaction with the non-automated simplification approach currently adopted for the production of PLS. The second aim was to determine if Cochrane authors' satisfaction could be boosted by the introduction of the Acrolinx CL checker into the current simplification approach. As reported in Section 1.2, satisfaction represents the DV1 of the empirical investigation described in this thesis. To show how this experimental study relates to our broader investigation, in Figure 4.1 we highlighted satisfaction as the DV.

This chapter will begin with a summary of related research. Subsequently, we will present the rationale for assessing Cochrane authors' satisfaction, the RQ, the research hypotheses of this experiment, and the characteristics of the Acrolinx software used in this study. We will then describe the recruitment of participants and the experimental environment, design, and materials. Subsequently, we will delve into the methods adopted to measure authors' satisfaction. Finally, the analysis of the data will be presented, and the results will be discussed.

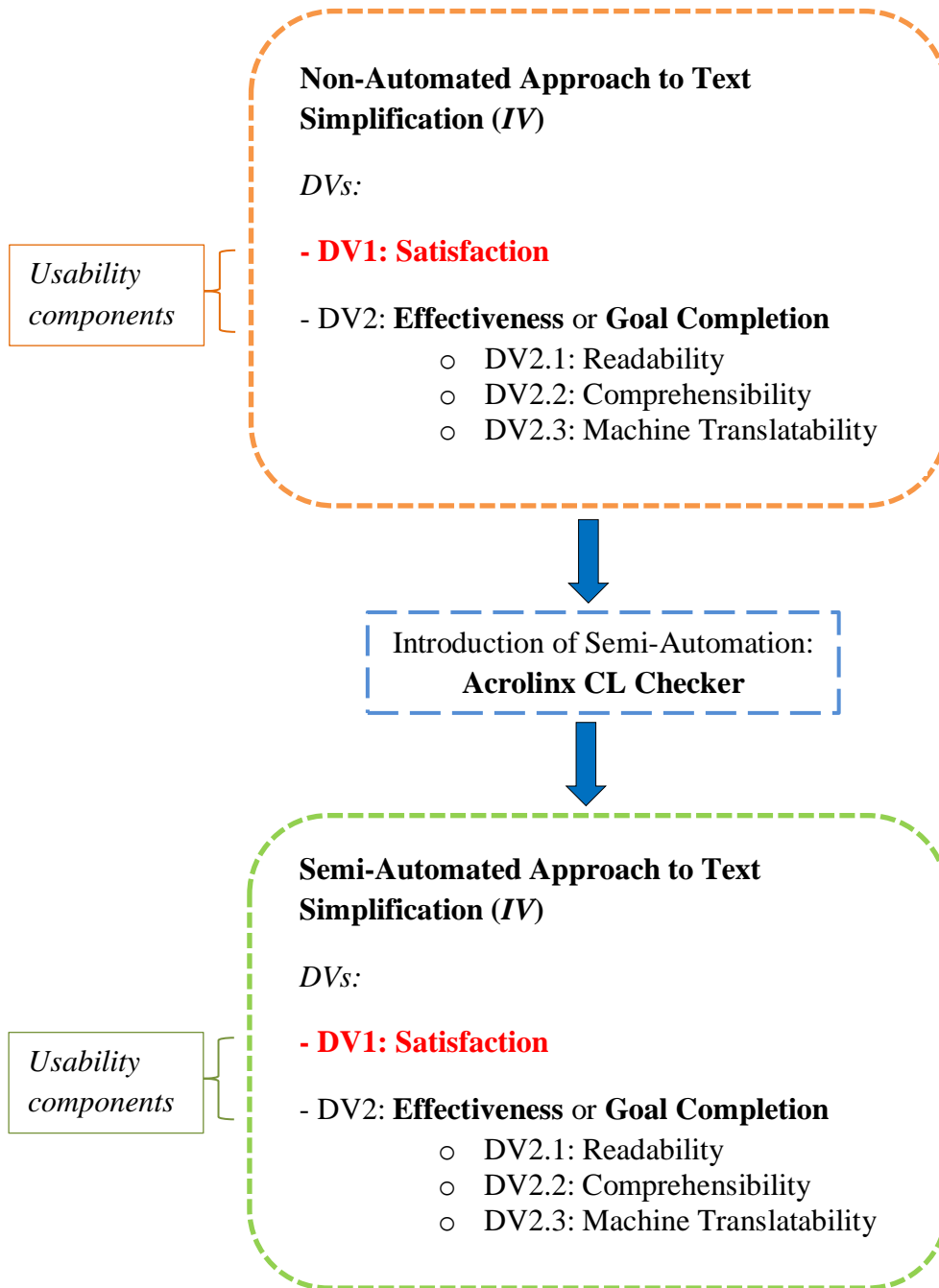


Figure 4.1: Satisfaction as DV1

4.2 Related Work on Satisfaction of Volunteers

Text simplification potentially represents a difficult and time-consuming task, especially when authors/content writers are expected to remember or check long lists of authoring guidelines (Aikawa et al. 2007), as in the case of Cochrane guidelines on PLS (Section

4.3). Schwitter (2015) discusses the difficulty of rewriting a text in Basic English while maintaining its meaning (Section 3.2). In her analysis of the time and difficulties which characterise manual text simplification based on a CL in the crisis management domain, Temnikova (2012) found that her participants simplified, on average, between 3.42 and 65.61 words per minute. The author's (ibid.) conclusions were that manual simplification represents a time-consuming task. In line with this observation, Nyberg, Mitamura and Huijsen (2003, p. 248) argue that "CLs which are not supported by automatic checking require self-vigilance on the part of the author, which can also be time-consuming". Unsurprisingly, Temnikova (2012) points out that a semi-automatic simplification assistive tool might prove beneficial to reduce authors' time commitment and effort. The introduction of a tool which provides assistance during the simplification task might be especially necessary when authors are volunteers with medical (rather than linguistics) background. In the case of Cochrane, not all review groups have an editorial team with PL writing skills (Higgins and Green 2011). Therefore, the onus is usually on the authors to ensure that texts are written in PL.

Since volunteers working for non-profit organisations (like Cochrane) do not receive monetary reward for their work (Millette and Gagné 2008), the motivations behind their commitment have been widely investigated. Clary et al. (1998, p. 1518) present a model according to which volunteering can serve six functions, which are: expressing altruistic values; engaging in favourably viewed activities; advancing one's career; reducing one's sense of guilt; enhancing positive mood; and understanding, which is defined as the possibility for volunteers to be involved in new learning experiences, and to put their skills and abilities into practice. Below we mainly present studies that deal with motivations for volunteering content production (either authoring or translation).

The function of understanding (Clary et al. 1998) emerged as a motivating factor in studies conducted with volunteer translators. O'Brien and Schäler (2010) found that the possibility to receive feedback from the non-profit organisation The Rosetta Foundation and from qualified translators on translated texts was the leading motivating factor for continued contributions among the volunteer translators that they surveyed, as

it allowed them to improve their translation skills. Similarly, in her study on the motivations of volunteer translators of TEDTalks, Olohan (2014) observed that learning from the talks was one of the motivating factors for contributing to their translations. In the context of Wikipedia, Baytiyeh and Pfaffman (2010) argue that volunteer authors might be attracted by the possibility of familiarising themselves with new software, and learning new features might increase their satisfaction.

When volunteering involves content production (as in the case of Cochrane or Wikipedia), Nov and Rao (2008, p. 85) identify asymmetry — defined as “a lack of contributed resources for maintaining and improving the common pool of resources” — as a threat. In other words, when content production relies on volunteers, there is some risk that volunteers’ contributions will be scarce and not sufficient to expand and update content. Ensuring the commitment of volunteers to the authoring task (e.g. by boosting their satisfaction) might therefore reduce the threat of asymmetry, and result in an increase in the amount of content produced, i.e. in “volunteering rates” (Olohan 2014, p. 19).

In summary, previous works seem to indicate that: (i) the requirement to remember, check and manually apply a list of simplification guidelines to a text might prove daunting and time-consuming for authors; (ii) the introduction of semi-automation might facilitate the simplification process; (iii) the introduction of semi-automation in the form of an assistive tool might represent a satisfactory learning or training opportunity that allows volunteer authors to develop their PL writing skills and, in turn, encourages them to continue simplifying texts. In other words, receiving semi-automated assistance and feedback (for instance, from a CL checker) during the simplification task might increase authors’ satisfaction (and, in turn, motivation) by reducing the time and effort required to check/remember simplification guidelines, and by turning the simplification task into a learning experience.

4.3 Motivation for Examining Cochrane Authors' Satisfaction, Research Question, Research Hypotheses, and Characteristics of the Acrolinx CL Checker

The non-automated simplification approach currently adopted at Cochrane involves the manual checking and implementation of different sets of guidelines. Our analysis of this approach shed light on the need to introduce a form of technological assistance for authors. First of all, we observed that Cochrane guidelines on PLS can be found in a variety of manuals online, which is likely to increase the time and effort required to check and remember them. More precisely, guidelines can be found in the *Cochrane Handbook for Systematic Reviews of Interventions* (Higgins and Green 2011), in the *PLEACS* (The Cochrane Collaboration 2013), in the *Cochrane Style Manual* (The Cochrane Collaboration 2016), and in the checklist *How to Write a Plain Language Summary of a Cochrane Intervention Review* (Cochrane Norway 2017).

Furthermore, we observed several contradictions both between and within manuals. Regarding contradictions within the same manual, in the checklist *How to Write a Plain Language Summary of a Cochrane Intervention Review* (Cochrane Norway 2017, p. 2), authors of PLS are encouraged to consider the characteristics of their target audience (e.g. whether they are parents, health workers, or policy makers) when choosing their writing style. However, in the same page, authors are also instructed to always assume that their readers are non-native speakers of English, and are not familiar with the topic and the methods. Regarding contradictions between manuals, in Higgins and Green (2011), it is reported that PLS should contain up to 400 words, while in the other manuals, it is written that PLS can contain between 400 and 700 words. Moreover, while in the checklist authors are instructed to either avoid acronyms or explain them (Cochrane Norway 2017), in the *PLEACS*, authors are encouraged to use acronyms for repeated use (The Cochrane Collaboration 2013).

Sometimes Cochrane PL guidelines also lack specificity or examples and, as a consequence, it might be difficult for authors to determine if their edits accomplish the goal of text simplification. For instance, authors are instructed to use short paragraphs and to address one key point per sentence (The Cochrane Collaboration 2013). However, no indications on the number of words recommended in paragraphs or

sentences are provided. Similarly, authors are encouraged to replace technical terms that would be difficult to understand for a lay audience with their PL counterparts (Cochrane Norway 2017). However, no lists or examples of hard/technical words that should be replaced are provided. Glenton (2017, p. 8) reports the following from her pilot study with Cochrane authors and review groups:

[s]everal people pointed to a need for more guidance on how to write in plain language. For instance, they called for guidance regarding the use of active versus passive voice with *examples*, a *reminder* to use short rather than long sentences, and more *suggestions* about how common terms could be expressed in plain language. (Emphasis added)

Finally, since authors do not get feedback on the impact (or lack thereof) of their edits, applying Cochrane PL guidelines is unlikely to be regarded by volunteer authors as a training opportunity aimed to develop their PL writing skills.

This analysis of the limitations that characterise the non-automated simplification approach²² currently adopted at Cochrane led us to hypothesise that the introduction of semi-automation in the form of the Acrolinx CL checker might turn the simplification task into a more satisfactory experience for authors (thus, in turn, increasing their commitment to simplify a higher amount of Cochrane health content). Developed at the German Research Center for Artificial Intelligence, the Acrolinx CL checker is a commercial tool which claims to ensure the readability and translatability of content by checking texts against a set of CL rules on style, spelling, grammar, tone of voice, and terminology (Rodríguez Vázquez 2016). In addition to automatically and consistently identifying and flagging issues in texts when CL rules are contravened, the Acrolinx CL checker provides suggestions on how to solve them, as well as an overall text score based on style, spelling, grammar, tone of voice, and terminology. We selected this CL checker for our study because it has already been applied to health content in (Simple English) Wikipedia articles (Ojala 2013; Azzam et al. 2017).

By automatically flagging readability and translatability issues in a text, a CL checker is likely to reduce authors' effort since they are not required to remember or

²² In line with the goals of this thesis (Section 1.2), this analysis only focused on Cochrane guidelines dealing with language/style (i.e. simplification) rather than content (i.e. summarisation).

check the guidelines. Moreover, by scanning documents against only one set of CL rules (rather than sets of guidelines scattered across different manuals), the CL checker avoids contradictions and inconsistencies that might irritate the authors. In addition, the use of a CL checker might represent a learning opportunity for authors, since they are presented with a real-time score based on text characteristics and suggestions on how to solve readability and translatability issues. In summary, we tested the hypothesis that these aspects of the use of a CL checker (i.e. reduced effort on the part of authors for checking/remembering guidelines, increased consistency of rules and learning opportunity) would boost authors' satisfaction. *Satisfaction* is defined here as "extent to which the user's physical, cognitive and emotional responses that result from the use of a system, product or service meet the user's needs and expectations" (ISO 9241-11:2018, 3.1.14).

As discussed in Section 1.2, the RQ on satisfaction that this investigation sought to answer is:

RQ1: Does semi-automating a non-automated simplification approach by introducing a CL checker increase authors' satisfaction?

The research hypotheses associated with RQ1 are the following:

H0 (null hypothesis): Semi-automating a non-automated simplification approach by introducing a CL checker does not increase authors' satisfaction.

H1 (alternative hypothesis): Semi-automating a non-automated simplification approach by introducing a CL checker increases authors' satisfaction.

For our study, we adopted the Acrolinx plugin for Microsoft Word (Sidebar Edition, Version 1.5 SR2), which shows readability/translatability issues and provides suggestions on how to solve them in the sidebar. When selecting a specific issue in the sidebar, the issue is also flagged in the text. In addition, for some of the issues, a *MORE INFORMATION* option is available. By clicking on it, a help file containing examples and more comprehensive information on the readability/translatability issue will open.

Help files can be visualised in the web browser (Rodríguez Vázquez 2016). Therefore, two different types of content are made available by the Acrolinx software: linguistic content (i.e. terminology, tone of voice, style, spelling, and grammar rules that, when contravened, result in issues/errors being flagged to users); and didactic content (namely, the display and explanation of issues to users) (Reuther and Schmidt-Wigger 2000). Figure 4.2 shows the Acrolinx sidebar in Microsoft Word and how issues were flagged in a sample PLS.

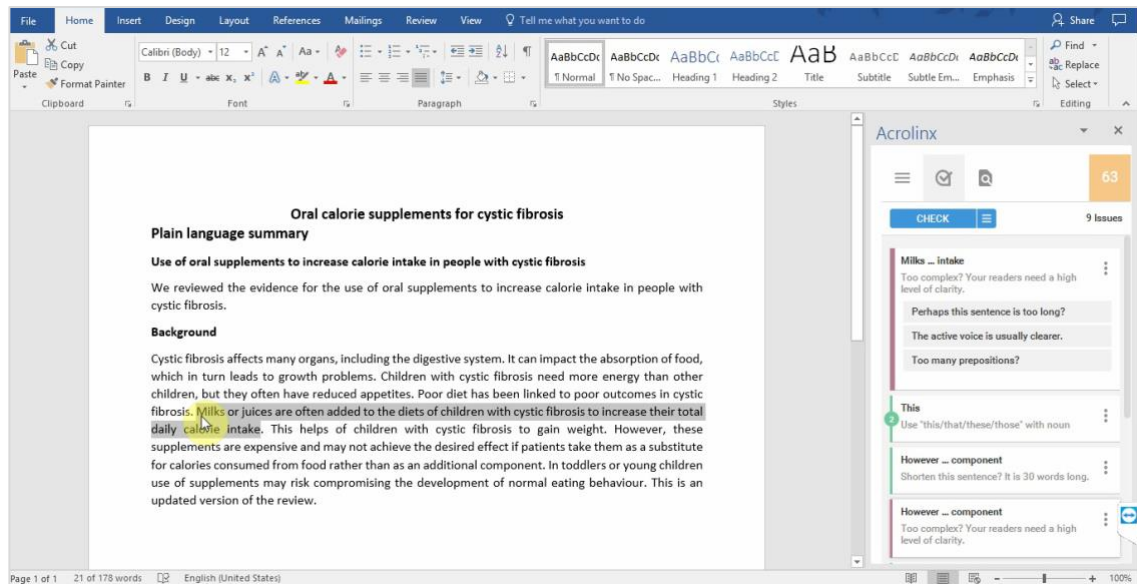


Figure 4.2: Acrolinx flagging readability/translatability issues in a sample text, and presenting suggestions and additional information in the sidebar in Microsoft Word

We adopted the set of CL rules *Standard_US*, which means that, in terms of spelling, the texts were checked against US (rather than UK) English spelling. We selected US spelling due to the high number of US-based review authors (about 1,500) and users (over 1.8 million visits in 2009) of the Cochrane Library (Tovey and Dellavalle 2010). In the *Cochrane Style Manual* (The Cochrane Collaboration 2016, p. 37) it is specified that “Cochrane Review Groups support both British (i.e. UK) and American (i.e. US) English. [...] Cochrane Reviews can use either spelling but the choice should be applied consistently within a single Cochrane Review or document”.

We used a local server which allowed for the (de)selection of CL rules based on the requirements of the study. More precisely, prior to the beginning of the experiment,

we deactivated Acrolinx CL rules which contradicted Cochrane PL guidelines. For instance, the Acrolinx rule on avoiding modal verbs was deactivated since Cochrane PL guidelines include the following:

If your assessment of the quality / certainty of the evidence is anything other than high, then you should avoid strong statements such as “[intervention] leads to [“outcome”]. You should rather indicate to the reader that there is some degree of uncertainty by adding modifying terms such as “probably”, “may” (Cochrane Norway 2017, pp. 4-5).

Other Acrolinx CL rules that were deactivated are: (i) the rule on avoiding future tense except for definite future events — it contravened Cochrane PL guidelines, which encourage authors to use the future tense even for planned events (The Cochrane Collaboration 2016); (ii) the rule on avoiding the hyphen after prefixes *re* and *sub*, even when the former is followed by a word starting with *e* — this rule contradicted Cochrane PL guidelines, which instruct authors to use a hyphen after *re* if the following word starts with *e*, and after *sub* if the word that follows starts with *b* (ibid.); (iii) the rule on using the en dash (–) to separate range of numbers — it contravened Cochrane PL guidelines, which encourage authors to use *from* and *to* when indicating ranges of numbers (ibid.); (iv) the rule on reporting both the imperial (e.g. feet and inches) and the metric units of measurement — it contradicted Cochrane PL guidelines, in which only the use of the metric system is recommended (ibid.); and (v) the rule on avoiding possessives, which contradicted Cochrane PL guidelines, where it is specified that apostrophe and *s* can be used to indicate possession (ibid.).

As reported in Section 3.2, PL could be regarded as a type of CL, i.e. “a constructed language that is based on a certain natural language, being more restrictive concerning lexicon, syntax, and/or semantics, while preserving most of its natural properties” (Kuhn 2014, p. 123). The discrepancies existing between Acrolinx CL rules and Cochrane PL guidelines originate from the fact that the Acrolinx software has been originally developed for the optimisation of technical/enterprise content (rather than medical texts). We did request that the Acrolinx team tailor the CL checker to Cochrane medical content, but this was not possible. Therefore, the deselection of specific CL

rules was the only available strategy to ensure that the use of the CL checker would not lead authors to implement irrelevant or contradictory rules.

Another difference between Acrolinx CL rules and Cochrane PL guidelines lies in the fact that the former aims to increase both readability and translatability of texts, while translation is barely mentioned or acknowledged in the manuals that form Cochrane's non-automated simplification approach – only in the checklist *How to Write a Plain Language Summary of a Cochrane Intervention Review* (Cochrane Norway 2017), is it reported that using modifying terms (such as *may* or *probably*) might lead to a decrease in the translatability of the PLS, as the meaning of these terms varies across languages. We did not deactivate Acrolinx translatability-oriented rules as they did not contravene Cochrane PL guidelines.

Cochrane provides authors of PLS with guidelines not only on language/style, but also on content — it should be remembered that PLS are the result of both a simplification and a summarisation process of systematic reviews (Section 1.1). Examples of guidelines on content are: “[i]nclude population details such as severity of condition, age, gender and comparators. Not all details of the included studies need to be reported fully” (The Cochrane Collaboration 2013, p. 6), or “[i]f the review explicitly considers how funding sources may affect the quality of the evidence then include a statement indicating the impact in the PLS” (ibid., p. 6). Noncompliance with guidelines on content cannot be described using a specific formalism and, in turn, cannot be integrated into Acrolinx's error-type oriented approach to checking (Bredenkamp, Crysmann and Petrea 2000). Therefore, even though the Acrolinx CL checker can be integrated into Cochrane's non-automated simplification approach to scan texts for language-related issues that can be formalised, it might not replace it completely since authors also require indications on the content that should be included in PLS. The inability to automatically check for compliance with all rules is also mentioned in Schwitter (2015). Describing ASD-STE, the author (ibid., p. 453) points out that the implementation of some rules requires human knowledge and experience, as in the case of “[p]resent new and complex information slowly”.

Finally, despite the differences between Cochrane PL guidelines and the Acrolinx CL checker, they both imply a structural approach to text simplification, which is characterised by lists of words and structures that should be used to enhance readability and comprehensibility (Crossley, Allen and McNamara 2012). Because of this characteristic, both the non-automated and the semi-automated simplification approach differ from an intuitive approach, in which authors can only rely on their intuition of which edits are required.

4.4 Recruitment of Cochrane Authors

The recruitment of participants among Cochrane authors was conducted between April and May 2017, after receiving ethical approval from the Research Ethics Committee at Dublin City University (DCUREC/2016/155) (Letter of Approval in Appendix A), and after preparing the call for participation (CFP) (Appendix B). In line with recommendations from the Research Ethics Committee, the CFP specified: (i) the names and affiliations of the researchers involved in the study, that is, the author of this thesis, her supervisor (Dr Sharon O'Brien), and Dr Silvia Rodríguez Vázquez (who was also a member of INTERACT [Section 1.3] and provided technical assistance with the set-up of the Acrolinx software); (ii) the requirement for participation (i.e. having experience in producing PLS of Cochrane Reviews); (iii) the description of the tasks and the expected time commitment (i.e. around two hours); (iv) the indication that participation was on a voluntary basis, that participants could withdraw from the study at any point without repercussion, and that data would be treated confidentially; and (v) an invitation for interested participants to contact the researchers via email.

As reported in Section 1.3, this study was conducted within the context of INTERACT, in which Cochrane was also involved. In particular, this experiment on authors' satisfaction was conducted when the author of this thesis was on secondment at Cochrane UK in Oxford (in May and June 2017). We could therefore avail of the assistance from Cochrane UK at the recruitment stage. At first, a random sampling recruitment technique was adopted — the CFP was advertised through the websites and social media of Cochrane UK and the broader Cochrane community, as well as on

TaskExchange²³ (an online platform used by Cochrane contributors to post or respond to tasks). In addition, specific Cochrane Review Groups (e.g. Cochrane Wounds and Cochrane Hepato-Biliary Group) saw the CFP and shared it on their websites and social media. Due to the low number of responses obtained with this recruitment technique — only three people emailed us to inform us that they were willing to participate — and after consulting with our collaborator at Cochrane UK, we decided to also adopt a snowball recruitment technique, which is very common in the areas of humanities and social sciences (Saldanha and O’Brien 2013). In particular, we sent the CFP to the Co-ordinating and Managing Editors of Cochrane Review Groups for email distribution to the authors in their Groups. The work of Co-ordinating and Managing Editors consists in supporting authors in the preparation, maintenance, and update of Cochrane Reviews (*Editorial Team* 2018). We assumed that this technique would allow us to reach out to those authors who had not seen the CFP online. Finally, we also adopted a purposive sampling technique by searching the Cochrane Database for systematic reviews that had been published in the three months prior to the beginning of the recruitment. We then sent the CFP to the corresponding authors by email. Due to the lack of responses, we then repeated the same recruitment strategy with the corresponding authors of reviews published within one year prior to the beginning of our recruitment.

To summarise, similar to the recruitment technique reported in Gaspari, Almaghout and Doherty (2015), we combined direct emails with online advertisement of our study with the aim of maximising the number of potential participants. Since the CFP was circulated with a variety of methods that were only partially controlled by the author of this thesis, it was not possible to calculate the response rate.

4.5 Experimental Environment, Procedure and Tasks

The Cochrane authors that replied to our CFP via email and agreed to take part in the experiment conducted the following tasks:

Task 1. After reading the PL statement describing the study and the informed consent form, authors completed an online background questionnaire which contained six open-

²³ TaskExchange can be accessed here: <https://bit.ly/2A8Toe9> [Accessed 12 December 2018].

ended questions and two multiple-choice questions (Appendix C). The questionnaire aimed to collect data on the authors' background characteristics and on their eligibility (i.e. having produced at least one Cochrane PLS). In addition, authors were asked to report the title(s) of the systematic reviews for which they had produced a PLS (if any). Answers to this question informed the selection of the experimental materials (Section 4.6). Moreover, in the background questionnaire, authors indicated in which editing environment they usually worked (e.g. Microsoft Word or Google Drive) and the month and year of their most recent PLS. Finally, authors were asked to insert additional comments (if any) and their email addresses, so that they could be contacted for the following tasks (Section 4.8.1). After completing the background questionnaire, each eligible author was assigned a participant ID that was then used as the unique identifier in the following tasks;

Task 2. Authors completed an online questionnaire (Appendix D) on their typical interaction with Cochrane PLS guidelines²⁴ (i.e. the non-automated simplification approach) and the level of satisfaction associated with it. Furthermore, authors were given the possibility to add comments (Sections 4.8.2 and 4.8.4). This questionnaire contained 14 multiple-choice questions, one checkbox question, and two open-ended questions;

Task 3. Each author remotely accessed our computer, where they could find the Acrolinx software installed as a plugin in Microsoft Word (Section 4.3), one sample PLS for a warm-up task with Acrolinx, and one of their old PLS, on which they were asked to run Acrolinx for a readability/translatibility check in Microsoft Word and to make simplification-oriented edits, as appropriate (main editing task). It should be noted that, prior to Task 3, we conducted a small pilot study with two health professionals (recruited with a convenience sampling technique) in order to: refine instructions for

²⁴ In the sets of Cochrane PLS guidelines, recommendations on simplification are provided alongside recommendations on summarisation. Accordingly, it might have been complicated for authors to report on their interaction/satisfaction with simplification guidelines only. We therefore specified that, by *Cochrane PLS guidance*, we meant any instructions, recommendations or guidelines that authors had been provided regarding the authoring of PLS. This decision entailed the limitation of not being able to isolate authors' interaction/satisfaction with simplification guidelines from their interaction/satisfaction with summarisation guidelines.

authors; test that there would be no technical issues when using Acrolinx remotely; and identify any other potential problems that authors might have encountered during the editing task. As a result of the pilot study, we made the following changes to the instructions: (i) we specified that participants could run as many Acrolinx checks as they deemed necessary (Appendix F); and (ii) we warned participants against changing the Acrolinx settings that had been selected, as this would have had a negative impact on the validity of the experiment (Appendix G). Task 3 can be further segmented into six different sub-tasks:

(i) Authors were required to install a free version of TeamViewer²⁵ on their machines by following the instructions available in Appendix E. TeamViewer is a piece of software that allows for remote support, access and collaboration by sharing an ID and a password. All authors took part in the study remotely since Cochrane contributors are located in more than 130 different countries and conduct most of their work online (*About Us* 2018). Because of the remote involvement of authors, the experimental environment was only partially controlled by the researcher. In other words, unlike the reading comprehension study, which was conducted in a laboratory setup (Chapter 6), we were not able to control environmental factors that might have influenced the editing work of authors, such as background noise, or support from external individuals. To partially reduce the impact of external factors, authors were instructed to avoid any interruptions (such as phone calls or email checks) once they began the editing tasks (Appendix F). Despite these limitations, allowing participants to take part in an experiment in their usual working environment had the benefit of increasing the ecological validity of the study, namely the extent to which the experimental environment and the tasks mimic real world situations (MacKenzie 2013, p. 143);

²⁵ A description of TeamViewer can be found at: <https://bit.ly/2xKclgT> [Accessed 12 December 2018].

(ii) Authors were also asked to complete a Doodle Poll to indicate the time and the day that suited them the most for the editing task. The day before the task, authors were sent a reminder via email;

(iii) On the agreed day, and ten minutes before the editing task started, each author was sent an email containing the ID and the password required to access our computer via TeamViewer. In the email, the authors were also informed that they would find all the necessary materials in two folders on the desktop: *WarmUp_Task* folder and *Main_PLS_Task* folder;

(iv) Authors were instructed to open the *WarmUp_Task* folder first, as this would give them a chance to familiarise themselves with Acrolinx. They were informed that they could spend as much time as they needed on the warm-up task (all the instructions on the warm-up task are available in Appendix G);

(v) After the warm-up task, authors were instructed to open the *Main_PLS_Task* folder containing the materials for the main editing task, which involved running the Acrolinx CL checks on their PLS and editing it in Microsoft Word. Prior to submitting the PLS to authors, we ensured that their formatting was consistent in terms of font (size), justification, and spacing between paragraphs. Moreover, since sometimes the title of the PLS differs from the title of the entire systematic review, we also provided authors with the titles of the systematic reviews. Authors were informed that they did not have any time limit, that they could run as many checks as they deemed necessary, and that they could use their common sense in deciding whether to apply a change recommended by Acrolinx or not (the entire list of instructions on the main editing task is available in Appendix F). As reported in Section 4.3, Acrolinx rules tackle both readability and translatability. However, in the instructions (Appendix F), we did not mention translatability and authors were asked to check the PLS for readability only. This decision was taken because, whilst authors had experience in simplifying texts to increase their readability (as emerged from their answers to the background questionnaire in Appendix C), they might have lacked similar experience in translation. Therefore, we assumed that introducing an instruction on

translatability might have resulted in an unfamiliar and confusing working scenario for the authors. Nonetheless, authors were not prevented from using any of the Acrolinx rules, regardless of whether they addressed readability or translatability. Moreover, the same Acrolinx rule often dealt with both issues (e.g. long sentences were flagged by Acrolinx for being difficult to read and to translate);

(vi) A Notepad file was open throughout the entire session — authors were informed that they could open the file and ask for assistance if they encountered a technical problem and/or if they had any questions. One of the benefits of using TeamViewer consists in the fact that, even though the computer is being remotely controlled, it is still responsive. Therefore, it was possible for the author of this thesis to answer authors' questions and comments that were typed in the Notepad file;

Task 4. Finally, authors completed a post-session online questionnaire (Appendix H) on the level of satisfaction experienced when running the Acrolinx CL checker on their old PLS produced with the non-automated simplification approach. In the questionnaire, authors were also asked which type of authoring support (if any) they would use in the future to ensure text readability, and what were the reasons for their answers. Finally, they were given the opportunity to add comments (Sections 4.8.3 and 4.8.4). This questionnaire contained 11 multiple-choice questions and three open-ended questions.

In summary, authors completed a warm-up editing task and a main editing task with Acrolinx remotely, as well as three online questionnaires (a background questionnaire; a questionnaire on their interaction and level of satisfaction with the non-automated simplification approach; and a questionnaire on their future preferences and their satisfaction when using Acrolinx). The three questionnaires were presented to authors on Google Forms. Not all tasks were conducted in the same session — task 1 was conducted in one session; task 2 was conducted in a different session, and finally, tasks 3 and 4 were carried out in the same session. The time span between sessions varied across authors, and not all authors carried out all the tasks (Section 4.8.3). For the

sake of clarity, the main tasks of this experiment (along with the data collected from each of them) are summarised in Figure 4.3.

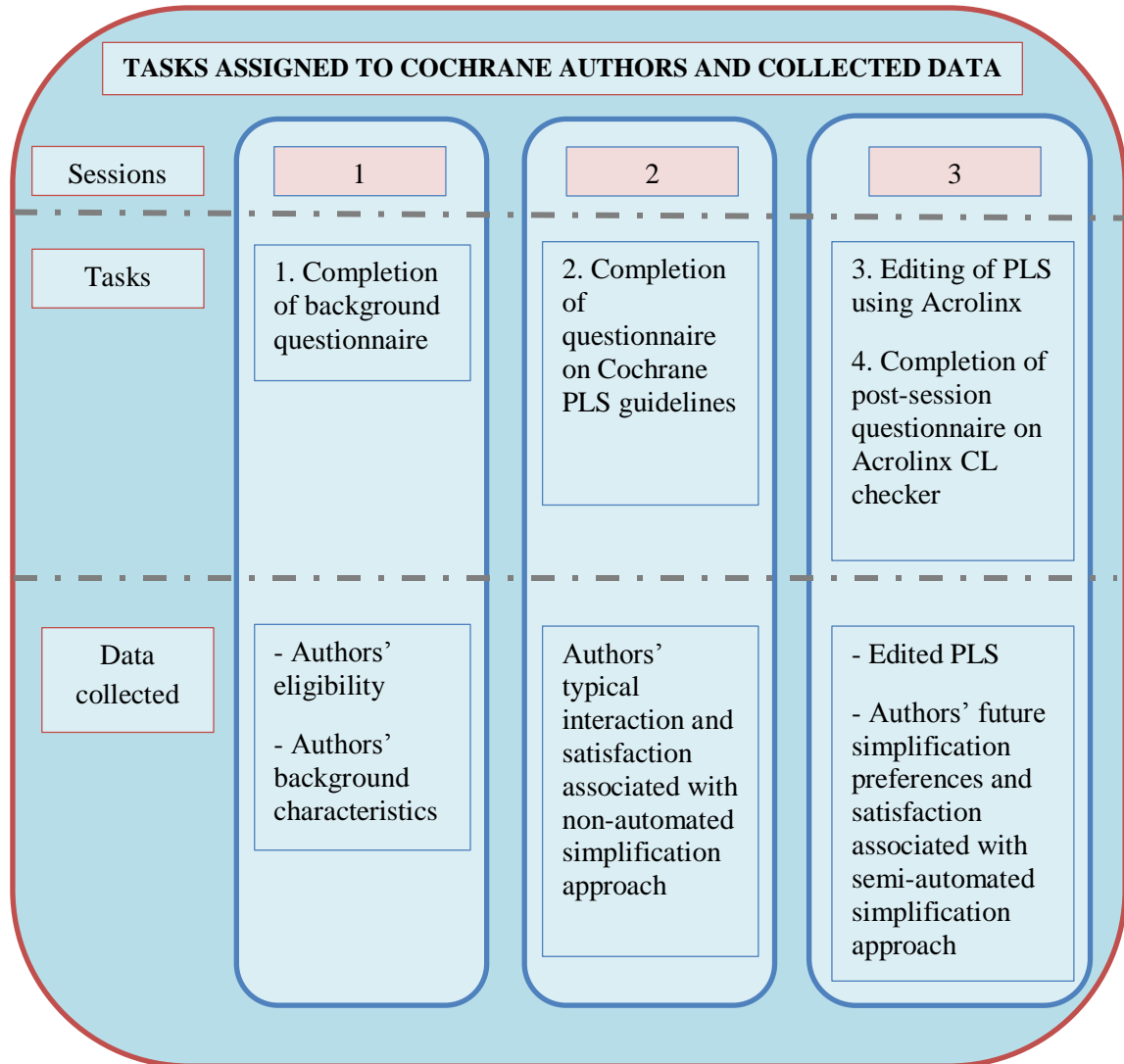


Figure 4.3: Tasks assigned to Cochrane authors per session and collected evidence

As reported in Section 4.3, the set of CL rules selected in Acrolinx was *Standard_US*. In order to ensure that there would be no contradictions between the spelling recommended by Acrolinx and the spelling recommended in Microsoft Word (where the editing tasks were conducted), we selected US English as the language for spelling and grammar proofing in Microsoft Word. The display language of all the Microsoft Word functionalities was also American English. In addition, we deactivated the option *Show readability statistics* in Microsoft Word proofing settings and we only used Microsoft Word grammar proofing. This decision was taken to ensure that: (i) style recommendations from Microsoft Word would not contradict Acrolinx

recommendations; and (ii) the impact of Acrolinx only (rather than Acrolinx and Microsoft Word settings) on text readability, comprehension and machine translatability would be evaluated.

4.6 Experimental Design and Selection of Experimental Materials

When originally designing this experiment, we considered the possibility of creating two comparable simplification scenarios: one scenario with authors producing a PLS from scratch by following Cochrane guidelines only, and another scenario with authors producing a PLS from scratch by using Acrolinx only. However, we abandoned this original design because Acrolinx could not provide authors with indications on the content to be included in PLS — as reported in Section 4.3, Cochrane PLS guidelines deal with content (in addition to language). Guidelines on content are necessary because, to produce PLS from entire systematic reviews, authors need indications on what information to include and what information to leave out. In other words, PLS are the result of a process of both text summarisation and simplification, as shown by their very name, which contains a reference to both *PL* and *summaries* (Section 1.1).

Moreover, asking authors to produce a PLS from scratch from an entire systematic review (rather than only checking its readability and editing it accordingly) would have resulted in a higher time commitment and, possibly, in a lower number of participants. Furthermore, we were advised against this design by our collaborator at Cochrane UK (T. Docherty 2017, personal communication, 31 March), who informed us that around 2-3 PLS are produced per month. Accordingly, the number of authors who could be recruited would have been limited. For the same reason, we could not ask authors to produce a PLS from scratch using intuitive simplification, namely simplification without any type of assistance (neither from the set of Cochrane guidelines nor from the Acrolinx CL checker).

Based on these considerations, we decided to adopt an alternative within-subject design — each eligible author was assigned one PLS that they had produced in the past (using the non-automated simplification approach, i.e. Cochrane PLS guidelines) and asked to check its readability using Acrolinx as a CL checker (Appendix F). In other

words, the PLS produced with the non-automated simplification approach in the past served as the starting point on which the CL checks were run and the simplification-oriented edits were implemented. We regarded this design as more in line with a realistic scenario of text simplification, which involves the iterative shaping and editing of content through feedback (Schrivver 2017).

With this design, it was not possible to include efficiency as a DV in our investigation — efficiency is traditionally measured in terms of the temporal and physical/cognitive effort required by a task (Tullis and Albert 2013). Since the simplification task with the non-automated approach was conducted by our participants in the months/years prior to our study, we could not measure and compare these aspects.

Adopting a within-subject design has several advantages: (i) compared to between-subject designs, a smaller sample size is needed; and (ii) the impact of individual differences is isolated since the same participant is exposed to different experimental conditions (Lazar, Feng and Hochheiser 2010, pp. 48-49). For these reasons, Lazar, Feng and Hochheiser (2010) recommend within-subject designs when: (i) the target participant pool might be small, such as in the case of users with disabilities or highly educated professionals; and (ii) participants are asked to conduct complicated tasks such as reading or writing, wherein individual differences (e.g. in cognitive skills) are likely to represent confounding variables. In our study, the target participants were health professionals and academics in the health field with busy schedules (Section 4.8.1). Moreover, the task of writing/editing falls into the category of complex tasks. Therefore, following recommendations in Lazar, Feng and Hochheiser (2010), a within-subject design was regarded as more appropriate. Furthermore, by adopting a within-subject design, we were able to isolate additional aspects that might have represented confounding variables, such as authors' different language backgrounds and PL writing skills, as well as different text characteristics (e.g. in terms of length, content or complexity of the PLS) (Sections 4.8.1 and 4.8.2).

Lazar, Feng and Hochheiser (2010) also discuss the limitations of within-subject design, namely, the impact of learning effect and fatigue. In our study, these two factors were unlikely to influence the findings, since authors had adopted the non-automated

simplification to produce a PLS between one month and three years prior to the beginning of the study (Section 4.8.1). It should be noted that, whilst this time gap between the non-automated text simplification task and the adoption of Acrolinx allowed for the avoidance of learning and fatigue effect, it proved problematic when authors had to answer questions on their interaction and satisfaction with the non-automated simplification approach adopted in the past (Section 8.4).

4.7 Methods Adopted for the Assessment of Authors' Satisfaction

In Section 4.3, we specified that *satisfaction* is defined here as “extent to which the user’s physical, cognitive and emotional responses that result from the use of a system, product or service meet the user’s needs and expectations” (ISO 9241-11:2018, 3.1.14). In turn, *product* is an umbrella term adopted to indicate any “item that is made or created by a person or machine” (ibid., 3.1.2). This definition of *product* seems to coincide with that of *artefact*, namely any object produced or employed by users for a specific purpose, which can range from calendars to information technology (Risku 2004, quoted in Krüger 2016). In this chapter, we will therefore use *product* as an umbrella term to refer to both the documents that characterise Cochrane PLS guidelines and the Acrolinx software.

Since users’ satisfaction is the result of their subjective assessment of the interaction with a product, in some studies, terms such as *perceived usability* and *subjective usability* have been used when describing satisfaction (Brooke 2013; Kortum and Acemyan 2013; Orfanou, Tselios and Katsanos 2015). Tullis and Albert (2013) assign satisfaction to the category of self-reported or subjective data, and argue that asking participants to report on their experience with a product might be the most straightforward way to collect evidence on its usability.

This section will be divided into two further sections. Section 4.7.1 will present the reasons for selecting a specific questionnaire for the assessment of Cochrane authors’ satisfaction. Then, Section 4.7.2 will describe the features of the selected questionnaire and how it was employed in the present study.

4.7.1 Rationale behind the Selection of the SUS

For the purpose of this study, the satisfaction of Cochrane authors (who were the users of both Cochrane PLS guidelines and Acrolinx) was measured by means of the System Usability Scale (SUS), which can be seen in Appendix D (question 5) and in Appendix H (question 2). Developed by Brooke (1996), the SUS has become one of the most widely used questionnaires to measure users' satisfaction (Sauro and Lewis 2009). This questionnaire is composed of ten statements. Each statement is accompanied by a 5-point Likert scale through which users of a product can indicate how strongly they agree (or disagree) with each statement.

In the present study, the SUS was adopted for several reasons. Firstly, it belongs to the category of post-session (or post-study) questionnaires, which allow users to indicate their overall level of satisfaction after completing all interactions with a product (Berkman and Karahoca 2016). Post-session questionnaires are different from post-task questionnaires, which contain a maximum of three items and are used when the interaction with a product can be segmented into smaller sub-tasks (as in the case of a purchase from an online store) (Christophersen and Konradt 2010). Simplifying or editing medical texts is a complex process that cannot be segmented into a series of different, separate tasks. In other words, we assumed that there would not be a phase in which authors read guidelines or rules, a phase in which they edited/simplified, and one in which they revised their work, but rather that these activities would alternate and overlap. Therefore, we regarded post-session questionnaires like the SUS as more appropriate than post-task questionnaires.

Secondly, the SUS was used in the present study because it is a technology-agnostic instrument, i.e. it can be adapted to various products (Bangor, Kortum and Miller 2008). In relation to this, Lewis and Sauro (2009) and Bangor, Kortum and Miller (2008) show that minor changes to item wording (such as replacing the original term *system* with other terms such as *application* or *website*) do not influence the findings and do not result in detectable differences in terms of reliability or factor structure. Numerous scholars have therefore replaced *system* with more appropriate terms, depending on the goals of their studies. For example, Chaparro et al. (2014) used

keyboard, and Sauro and Lewis (2011) used *website*. This characteristic of the SUS was particularly relevant to the present work, in which the satisfaction associated with products as diverse as a set of written guidelines and a piece of software was investigated.

The SUS was selected for this study also because previous works have shown that it is a reliable, valid and sensitive tool, which are the three main qualities of a scale (Berkman and Karahoca 2016). Cairns (2013) defines reliability as the extent to which the results of a questionnaire are consistent (e.g. the extent to which the items in a questionnaire measure the same dimension). Typically, questionnaires are required to have a minimum coefficient of reliability of .70 (Landauer 1997). Regarding the SUS, Kirakowski (1994) showed that its coefficient of reliability is 0.85. Tullis and Stetson (2004) compared the SUS with other questionnaires for assessing website usability (i.e. the Questionnaire for User Interface Satisfaction, the Computer System Usability Questionnaire, a variant of Microsoft's Product Reaction Cards and a questionnaire developed at their usability laboratory). The authors (ibid.) found that the SUS results were among the most reliable across sample sizes. Bangor, Kortum and Miller (2008) analysed the reliability of the SUS based on the results of 206 usability studies and found a value of 0.91. Similarly, in a subsequent analysis of 324 SUS questionnaires, Lewis and Sauro (2009) found a coefficient of reliability of 0.92.

As far as validity is concerned, Cairns (2013) describes it as the extent to which questionnaires are able to measure a concept. Evidence of the validity of the SUS, and more precisely of its concurrent validity (defined by Cairns [2013] as the ability of a questionnaire to provide results that match the findings of different instruments for measuring the same concept) can be found in Bangor, Kortum and Miller (2008). Based on the results of a pilot study, the authors (ibid.) suggest that different adjective descriptors (namely, *worst imaginable*, *poor*, *OK*, *good*, *excellent* and *best imaginable*) can be associated with different spans of SUS scores. Similarly, Sauro (2011) found a correlation of 0.79 between the scores of SUS and those of the Software Usability Measurement Inventory (SUMI) (Kirakowski and Corbett 1993). The author (ibid.) also observed that correlations between SUS scores and the Website Analysis

and MeasureMent Inventory scores (Kirakowski, Claridge and Whitehand 1998) were very high. Lewis and Sauro (2009) argue that the study conducted by Bangor, Kortum and Miller (2008) also provides evidence of the sensitivity of the SUS, thus showing that the questionnaire is able to discern different interfaces and different versions of the same product.

A further advantage of the SUS is that it contains fewer items than other questionnaires (e.g. the SUMI contains 50 items) and can therefore be filled out in a short time, thus reducing the time commitment of the participants (Brooke 2013). It should be noted that even shorter versions of the SUS have been developed. For instance, Finstad (2010) developed the Usability Metric for User Experience (UMUX), which only contains four items. The rationale behind its development was the assumption that presenting participants with fewer questions would encourage them to take part in usability studies. Even though the correlations between SUS scores and UMUX scores are very high ($r=0.96$), Cairns (2013) highlights some limitations of the UMUX. In particular, the author (ibid.) argues that the UMUX has very high internal consistency, which may indicate that the four items are redundant, thus making this questionnaire too specific. As for face validity (namely the extent to which questions appear appropriate for measuring satisfaction), Cairns (2013) reports that all the questions in the UMUX contain a clear reference to usability, which may have the disadvantage of making the goal of the questionnaire apparent and, in turn, leading respondents to provide socially desirable answers (Kline 2000). Lewis, Utesch and Maher (2013) developed an even shorter version of the UMUX, called UMUX-LITE, which only contains two items. Berkman and Karahoca (2016) found that this instrument is a reliable and valid scale, but is less sensitive than the SUS in detecting differences among software applications. Therefore, for our study, the SUS was preferred to both its shorter versions (i.e. the UMUX and the UMUX-LITE) for the measurement of satisfaction. The SUS was also preferred to a homegrown questionnaire, i.e. a questionnaire specifically developed by researchers for the purposes of their studies, since it has been shown that standardised questionnaires are more reliable than homegrown ones (Hornbæk and Law 2007).

A further reason for using the SUS is that it provides an overall single score (ranging from zero to 100) that can be easily comprehended even by non-experts (Orfanou, Tselios and Katsanos 2015). Furthermore, several studies are available that guide usability practitioners in the interpretation of SUS scores. Bangor, Kortum and Miller (2008) argue that a product needs to have a SUS score above 70 to be passable, and that products receiving a score lower than 50 should be the object of substantial concern.

Two independent studies, one conducted by Lewis and Sauro (2009) and the other by Borsci, Federici and Lauriola (2009), also showed that the SUS measures not only the satisfaction/perceived usability of a product, but also its learnability (by means of statements 4 and 10). Krüger (2016, p. 132) specifies that learnability “is concerned with how easily new users can familiarise themselves with a given software system”, and suggests the integration of learnability into ISO 9241-11:1998 (later revised into ISO 9241-11:2018), alongside the three usability dimensions of satisfaction, effectiveness (or goal completion) and efficiency (Section 2.2). The ability of the SUS to indicate the learnability of a specific product is particularly important since learning new skills has been shown to boost volunteers’ motivation (Section 4.2). In our study, the learnability dimension was especially relevant for the Acrolinx software, as Cochrane authors already had experience in adopting the non-automated simplification approach (i.e. Cochrane PL guidelines). Moreover, Guillardau (2009) and Krüger (2016) remark that the learnability of a product is influenced by the availability of product documentation (or lack thereof). In our study, to facilitate authors’ familiarisation with Acrolinx, we provided them with instructions on how to use the software (Appendix G).

The SUS is also non-proprietary and has been made freely available, as long as researchers acknowledge the source of the scale (Bangor, Kortum and Miller 2008; Tullis and Albert 2013). Finally, the SUS has already been successfully adopted to measure the satisfaction of non-professional writers using a CL checker on Japanese texts (Miyata et al. 2017).

4.7.2 Characteristics and Adoption of the SUS

As reported in Section 4.7.1, the SUS is a coarse-grained questionnaire, i.e. it can be used for a variety of different products and its wording can be slightly tailored (Krüger 2016). For the purposes of our study, we replaced the original term *system* with *Cochrane PLS guidance* (Appendix D) and *Acrolinx* (Appendix H), respectively, and we used those terms consistently in all the ten statements, as recommended by Lewis and Sauro (2009). In addition, in statement 5 of the SUS on Cochrane PLS guidelines (Appendix D), we replaced the term *functions* with *documents*, since the former was deemed more inappropriate for the non-automated simplification approach, in which no technological functions are available.

In the version of the SUS originally developed by Brooke (1996), the five odd-numbered statements are positively worded, while the five even-numbered ones are negatively worded. Sauro (2011) specifies that alternating items are often employed in questionnaires with the aim of reducing extreme response bias (i.e. the tendency of respondents to select the extremes of a rating scale) and acquiescence bias (namely, the tendency of respondents to agree with the statements). In addition, Brooke (2013) explains that the adoption of alternating items would encourage participants to read and think more carefully about whether they agree or disagree with a statement. The scoring of the SUS can be divided into two stages:

1. Subtract one from the odd numbered items and subtract the even numbered responses from 5. This scales all values from 0 to 4 (with four being the positive response).
2. Add up the scaled items and multiply by 2.5 (to convert the range of possible values from 0 to 100 instead of from 0 to 40) (Sauro and Lewis 2011, p. 2).

Sauro and Lewis (ibid.) point out that using a questionnaire with an alternation of positively- and negatively-worded statements might result in some problems. In particular, they identify three disadvantages: misinterpret (participants may agree or disagree more with negatively worded statements than they would do if the statements were positively worded); mistake (participants might accidentally overlook item alternation, thus not reversing their scores); and miscode (researchers themselves may fail to take into account the need to reverse the scales when scoring). The authors (ibid.)

compared an original SUS questionnaire with a version containing all positively worded items. They found that both versions were characterised by a high level of reliability. Sauro and Lewis (2011) also demonstrated that, when using the SUS with negatively and positively worded items, the possibility that either respondents or researchers make mistakes is real — for instance, their analysis of 158 SUS questionnaires suggested that 13.3% of SUS questionnaires submitted remotely presented mistakes on the part of the respondents. Therefore, following recommendations from Sauro and Lewis (2011), the present study adopted an all positive version of the SUS.

The use of the SUS entails the limitation of not allowing researchers to collect diagnostic information that could explain why a user rated a specific product very highly or very poorly on satisfaction (Brooke 2013). In order to compensate for this lack of diagnostic information and to collect more fine-grained data on the aspects of a product that affect users' satisfaction, usability practitioners have adopted field notes, video/audio recordings (Perrier, Kealey and Straus 2014), and post-hoc preference questions (Chaparro et al. 2014). Similarly, in our study, we complemented SUS scores with follow-up questions to Cochrane authors. In particular, we asked them to indicate which type of support (if any) they would use to check text readability in the future (i.e. whether Cochrane PLS guidance only, Acrolinx only, both Cochrane PLS guidance and Acrolinx, another type of support, or none). Authors were also asked to explain the reasons for their preferences and were encouraged to leave additional comments, if they had any (Appendix H).

Several studies (e.g. Chaparro et al. 2014) have used the SUS along with the NASA Task Load Index, a tool which can be adopted to collect data on participants' subjective assessment of their workload (Hoonakker et al. 2011). This tool contains six scales (from very low to very high) which ask participants to rate their mental demand, physical demand, temporal demand, performance, effort, and frustration, respectively. By using this tool in the present study, it would have been possible to identify potential (lack of) correlations between satisfaction and perceived workload ratings. However, submitting additional follow-up questions to Cochrane authors would have increased their time commitment and their fatigue, which could have potentially resulted in a

higher dropout rate. Therefore, priority was given to those questions which were strictly necessary to assess participants' level of satisfaction (i.e. the SUS and the follow-up questions).

Based on their analysis of the mean scores of 206 studies involving a total of 2,324 SUS questionnaires, Bangor, Kortum and Miller (2008) suggest that another limitation of the SUS might be that its scores are restricted in terms of study means. In particular, the authors (*ibid.*) observed that no group score fell below 30, and that fewer than 6% of the mean scores of these studies fell below 50. Kortum and Acemyan (2013) conducted a study with different paper voting interfaces to determine if SUS scores are characterised by a range limitation. Their results did not confirm the limitation identified by Bangor, Kortum and Miller (2008), as the study mean SUS scores for 57% of the voting interfaces fell below 40. Based on Kortum and Acemyan's (2013) results, we assumed that we would not observe a range limitation in the mean scores provided by the Cochrane authors in our study.

Finally, Tullis and Albert (2013) point out that the SUS is a questionnaire that should be filled out right after completing all interactions with a product. While this proved possible for authors' interaction with the Acrolinx CL checker, it was not possible for their interaction with Cochrane PLS guidelines, since authors had produced their previous PLS in the months prior to the beginning of this study — as discussed in Section 4.6, it was not possible to conduct this study as part of the live production of PLS at Cochrane. Therefore, the results regarding authors' satisfaction with Cochrane PLS guidance (Section 4.8.4) should be treated with caution.

4.8 Data Analysis and Results

This section describes the analysis of the data collected from Cochrane authors through: (i) a background questionnaire on their characteristics and eligibility; (ii) a questionnaire on their typical interaction and satisfaction with Cochrane PLS guidance (i.e. in the non-automated simplification approach); and (iii) a questionnaire on their satisfaction with the Acrolinx CL checker (i.e. the semi-automated simplification approach) and on their future simplification preferences (Section 4.5). The main goal of this analysis was to

determine if introducing Acrolinx into Cochrane's current non-automated simplification approach would increase authors' satisfaction (DV1).

This section will be further divided into four sections. Section 4.8.1 will present the findings on the background characteristics of the Cochrane authors who took part in our study. Section 4.8.2 will report results on authors' typical interaction with and opinions on Cochrane PLS guidelines, while Section 4.8.3 will describe findings on authors' future simplification preferences and on the reasons for their preferences. Finally, Section 4.8.4 will present the SUS scores assigned by authors to both Cochrane PLS guidelines and the Acrolinx CL checker.

4.8.1 Cochrane Authors' Background Characteristics

Twenty-six respondents showed interest in this study and completed the background questionnaire (Appendix C). Eight of these respondents were discarded because they did not meet the requirement of having produced a Cochrane PLS in the past. Therefore, the following analysis is based on data collected from 18 authors.

The first question of the background questionnaire focused on the authors' native languages. The majority of them (13 out of 18) indicated English as their native language, while five reported a native language other than English, namely Dutch (n=2), German (n=1), Portuguese (n=1), and Russian (n=1). Authors were included in the study regardless of their native language. Cochrane relies on a global network of contributors from more than 130 countries (Section 4.5). Therefore, this decision was meant to increase the external validity of the study, which is determined (among other factors) by the extent to which the sample of participants is representative of a larger population (MacKenzie 2013, p. 141). Moreover, we adopted a within-subject design (Section 4.6), which allowed us to isolate the impact of individual differences in English ability (Lazar, Feng and Hochheiser 2010).

The second question of the background questionnaire dealt with the authors' jobs. Thirteen of them reported having an academic job in the health field (e.g. senior lecturer, associate professor, postdoctoral research fellow, etc.). Four authors reported being health professionals (such as epidemiologist or radiation oncologist). Finally, one

author reported being a freelance medical publications consultant. In three cases, it was not clear from the authors' responses whether they worked in the health field. In those cases, the answers that they provided were complemented with details from their email signatures or online profiles. For instance, P14 only wrote "Statistician", but from her online profile it emerged that she was a research assistant in statistics working in the medical department of a university, so she was assigned to the category of academic.

In the third question, authors were asked if they had produced a Cochrane PLS in the past. Only the 18 participants who answered *yes* to this question were eligible to participate in the present study. The fourth question asked participants to either provide the URLs or write the titles of the systematic reviews for which they had produced a PLS in the past. This question was particularly important since, as reported in Section 4.5, each author was then assigned one of their old PLS to edit using Acrolinx. Several authors reported having produced PLS for multiple systematic reviews. In these cases, authors were assigned their most recent PLS (as per year and month of first online publication of the systematic review). With the exception of two participants collaborating with the Cochrane Vascular Group, the rest of the respondents were contributors to various Cochrane Review Groups (Section 1.1), such as Cochrane Work Group, Cochrane Stroke Group, Cochrane Dementia and Cognitive Improvement Group, or Cochrane Breast Cancer Group. As a result, there was high variability in the content of the PLS employed for this study. In addition, the PLS that were then assigned to authors for the main editing task also varied in terms of length/word count, from 302 to 694 words. Since we adopted a within-subject design (Section 4.6), these differences in text content and length were not assumed to characterise confounding variables.

The fifth question asked authors to indicate the editing environment in which they usually produce Cochrane PLS. Out of 18 authors, 11 answered that they use Microsoft Word, six indicated Review Manager (RevMan)²⁶, and one reported using both Microsoft Word and RevMan. RevMan is the Cochrane authoring infrastructure,

²⁶ Details on Review Manager can be found at: <https://bit.ly/2NQrgxM> [Accessed 12 December 2018].

i.e. the software adopted for the writing and maintenance of systematic reviews. Figure 4.4 shows the RevMan 5.3 interface.

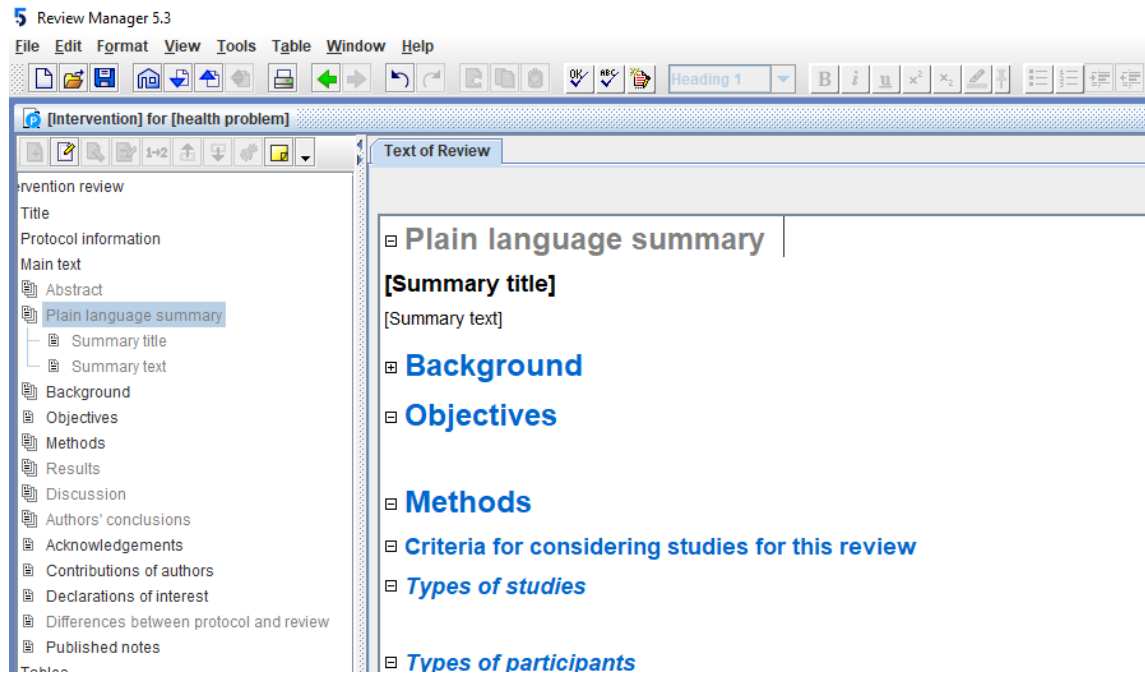


Figure 4.4: RevMan 5.3 interface

To increase the ecological validity of the study, it would have been preferable to allow the six participants who reported normally using RevMan to conduct the Acrolinx editing task in this environment. Nonetheless, the Acrolinx server which was made available to us could only be installed as a Microsoft Word plugin (Section 4.3). Despite the inability to provide these six participants with an Acrolinx plugin for RevMan, it can be assumed that the ecological validity of the study was only slightly affected since: (i) Microsoft Word is a widely adopted program (Badarudeen and Sabharwal 2010) and authors were likely to be familiar with it even though they did not usually employ it to produce PLS; (ii) in Microsoft Word, authors could perform the same editing tasks that were available in the RevMan text editor, such as copy and paste, or add and delete words.

The sixth question of the background questionnaire aimed to collect data on the month and year in which Cochrane authors produced their latest PLS. The vast majority of the participants (17 out of 18) reported working on a PLS between 2016 and 2017. More precisely, seven participants had last worked on their PLS in 2016, while ten

reported having produced a PLS in 2017 (six of whom in March). The oldest PLS production task was conducted by one participant in November 2014. The most recent PLS production task reported was in April 2017, only one month before the beginning of our study. Question 7 asked respondents to report their email addresses (which were treated confidentially). Finally, question 8 of the background questionnaire asked participants if they had any comments. No relevant comments were provided by the participants.

4.8.2 Authors' Interaction with Cochrane PLS Guidance

In the questionnaire on Cochrane PLS guidance (Appendix D), the first question asked authors to report the participant ID that had been assigned to them (e.g. P03). The second question focused on the type of guidance document that authors had received from Cochrane in order to write PLS of Cochrane Systematic Reviews. This question was asked because Cochrane guidelines on PLS can be found in several documents (Section 4.3). Results for this question are reported in Figure 4.5 below.

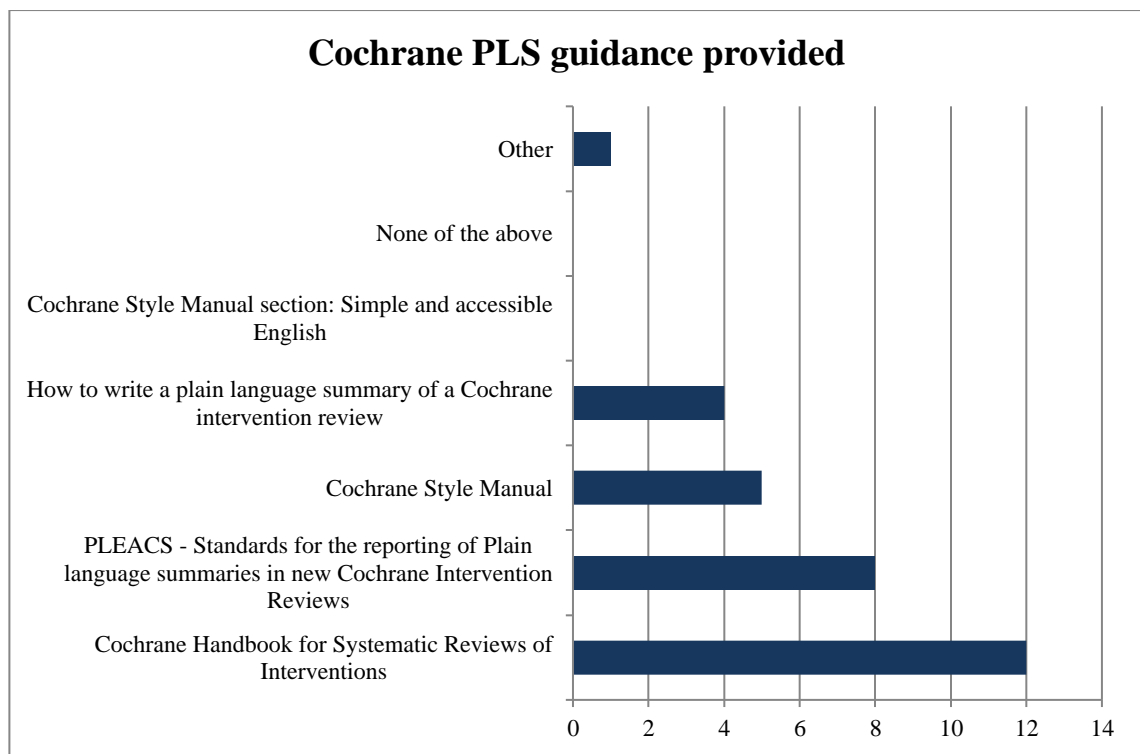


Figure 4.5: Cochrane PLS guidance provided to authors

Each author could select more than one option. Therefore, the number of responses exceeds the number of participants. The horizontal axis shows the number of selections per document. The documents that received the highest number of mentions were the *Cochrane Handbook for Systematic Reviews of Interventions* (Higgins and Green 2011) (n=12), and the *PLEACS* (The Cochrane Collaboration 2013) (n=8). It is interesting to note that not all Cochrane authors were provided with the same guidelines on the production of PLS.

The third question dealt with the authors’ workflow for PLS production — authors were asked to select the statement which best described their PLS authoring procedure in terms of use of Cochrane PLS guidance. Findings are shown in Figure 4.6. One response in the *Other* option was discarded because it was irrelevant, i.e. it did not deal with the workflow of PLS production. Therefore, the total number of responses reported is 17.

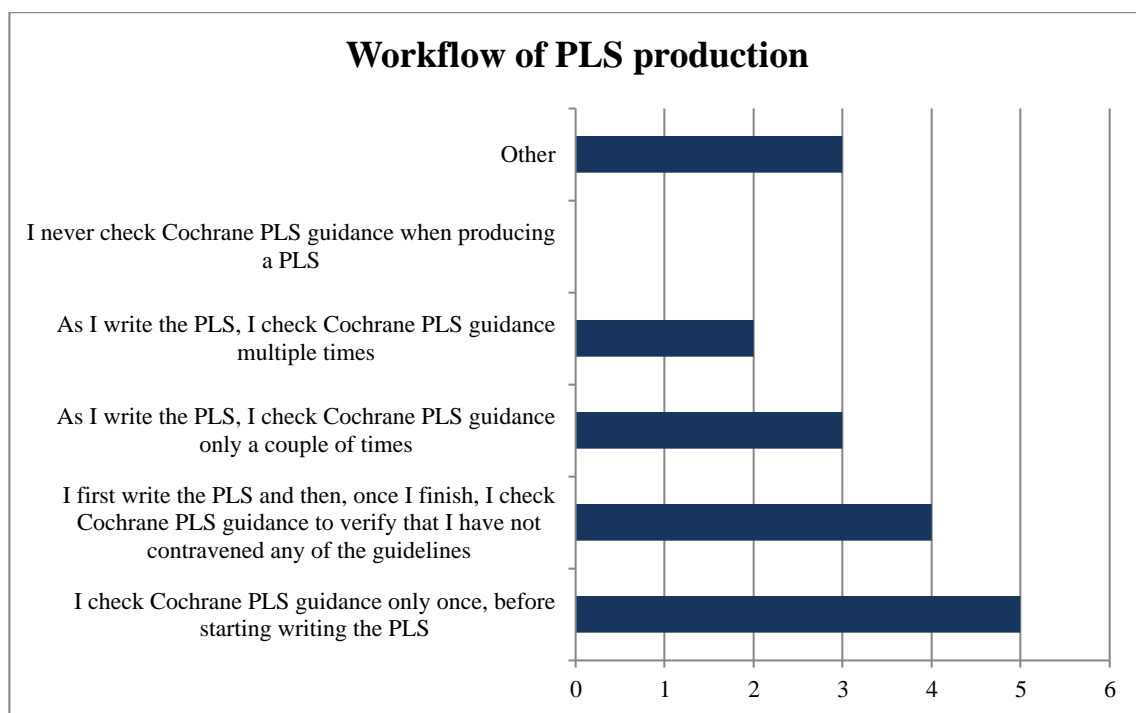


Figure 4.6: Workflow of PLS production

The horizontal axis represents the number of respondents who selected each option. Most authors reported consulting Cochrane guidance either before starting the

summarisation/simplification process that would lead to the production of PLS (n=5), or at the end of it (n=4), to check for their compliance with Cochrane recommendations. This finding is not surprising as it can be assumed that interrupting the writing process to check the guidelines is both time- and energy-consuming. This assumption is supported by the fact that only two participants (out of 18) reported checking the PLS guidance multiple times while writing. The category *Other* includes responses that describe additional types of workflow. P05 replied that they read and followed Cochrane guidance in the past, but not routinely. P18 reported asking what the new guidance is rather than checking it, because Cochrane recommendations change often. P20 replied that, in addition to checking Cochrane guidelines at the beginning and at the end of the writing process, they also consulted previous PLS. Regardless of the preferred workflow, all authors seem to include (or to have included) the consultation of Cochrane guidance in the process of PLS production.

The fourth question on Cochrane PLS guidance asked authors to indicate how often they consulted it, i.e. whether for all, some or none of the PLS that they have produced. Responses are reported in Figure 4.7. One response in the *Other* option was discarded because it was irrelevant, i.e. it did not deal with the frequency of consultation of PLS guidance. Therefore, the overall number of responses analysed is 17.

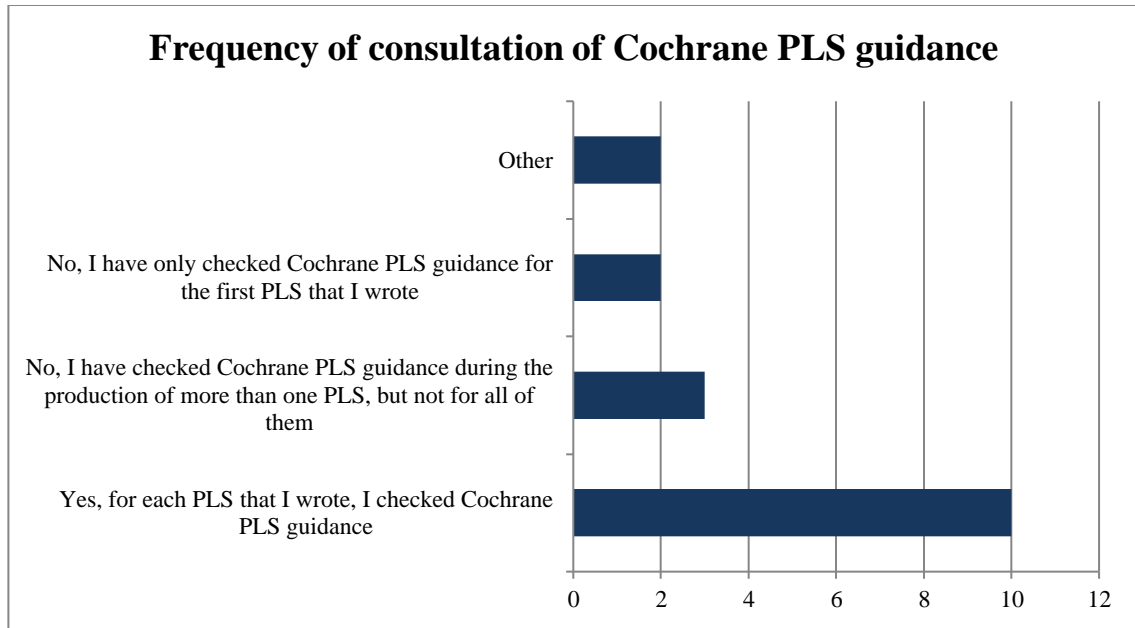


Figure 4.7: Frequency of consultation of Cochrane PLS guidance

The horizontal axis in Figure 4.7 represents the number of respondents who selected each option. It can be observed that the majority of participants (n=10) reported checking Cochrane guidelines on how to write PLS each time that they produced one. In addition, the participants who selected the *Other* option provided responses that indicate a frequent consultation of Cochrane guidance. More precisely, P20 replied that they checked Cochrane guidance both when producing a PLS from scratch, and when updating it, in case Cochrane recommendations had changed. P22 specified that they had written only one PLS, but they intended to recheck the guidance for future PLS. Only five participants (out of 18) reported not checking Cochrane recommendations for all the PLS they wrote.

The fifth question asked participants to complete the SUS on their interaction with this non-automated simplification approach. Findings for this question will be presented in Section 4.8.4, where they will be compared with the SUS scores assigned to the Acrolinx CL checker.

In the sixth question, authors were asked to indicate their level of agreement with the following statements:

a. Cochrane PLS guidance provides enough indications on the type of content to be included in PLS

b. Cochrane PLS guidance provides enough indications on the writing style to be followed in PLS

Two separate statements were used because, as reported in Section 4.3, the documents that compose Cochrane PLS guidance provide recommendations on both content (for summarisation purposes) and language/style (for simplification purposes). A Likert scale (from *1-Strongly disagree* to *5-Strongly agree*) was adopted. Results for the content-related question are shown in Figure 4.8, while Figure 4.9 presents results for the style-related question. In both figures, the vertical axis represents the number of respondents who selected a specific ranking.

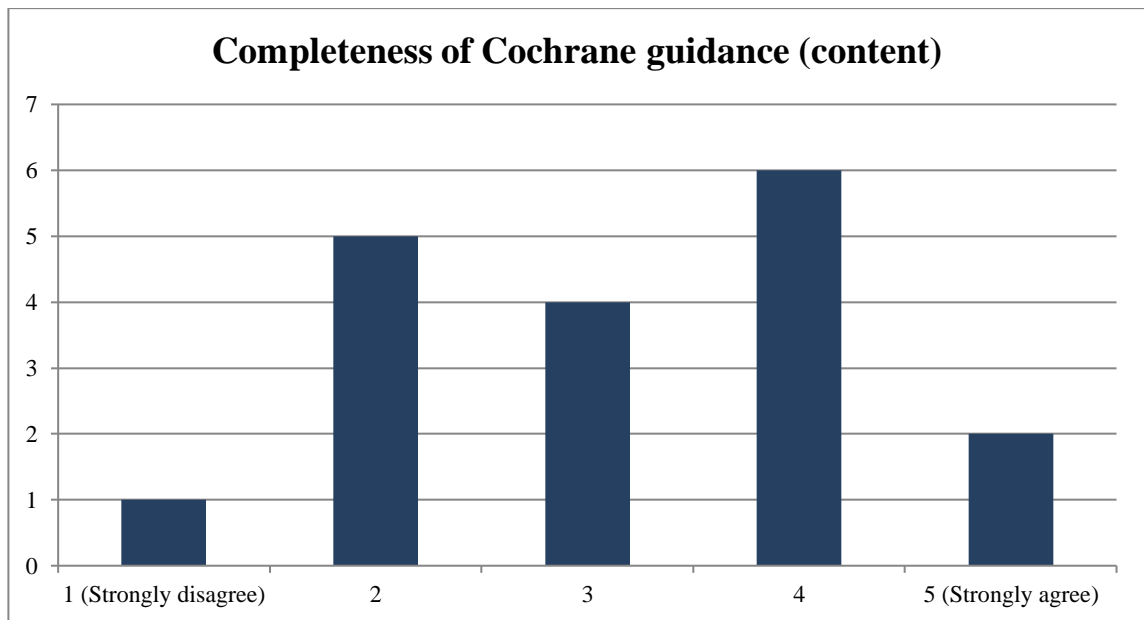


Figure 4.8: Rankings of completeness of Cochrane PLS guidance in terms of content

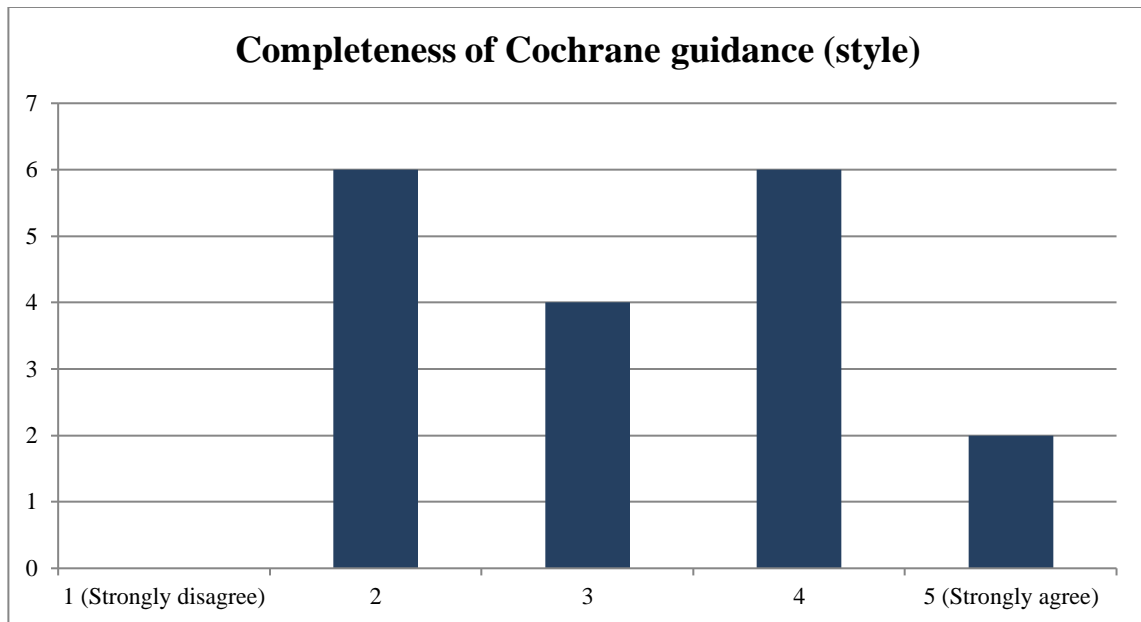


Figure 4.9: Rankings of completeness of Cochrane PLS guidance in terms of style

To more easily compare the rankings assigned by Cochrane authors to the completeness of Cochrane PLS guidelines in terms of content and style, the median and the mode of the rankings are reported in Table 4.1. Likert scales provide ordinal data for which it is not possible to calculate the distance among the different responses. Accordingly, the mean would not provide meaningful results. For instance, the mean value of *1-Strongly disagree* and *5-Strongly agree* could not be quantified (Sullivan and Artino 2013). In contrast, the median and the mode are regarded as appropriate measures of central tendency for ordinal data (Boone and Boone 2012).

Rankings of completeness of Cochrane PLS guidance	Median	Mode	N
<i>Content</i>	3	4	18
<i>Style</i>	3	2 and 4	18

Table 4.1: Descriptive statistics for Cochrane PLS guidance rankings

The median rates show that half of the scores assigned by the participants were greater than (or equal to) 3, and half of the scores were lower, both for content- and style-related guidelines. The mode rates show that, when ranking the completeness of Cochrane PLS guidance in terms of content, the score most frequently assigned by participants was 4 (n=6). When ranking the completeness of Cochrane PLS guidance in terms of style, the

scores most frequently assigned by respondents were 4 (n=6) and 2 (n=6). Overall, these results indicate that there is high variability in Cochrane authors' opinions on the completeness of Cochrane PLS guidance, with one third of the participants agreeing that the guidance contains enough recommendations on content and style, while another third disagreeing with these statements.

Finally, at the end of the questionnaire on Cochrane PLS guidance, authors were asked if they had any comments. The final comments provided qualitative data that complemented the quantitative data collected through this questionnaire. In particular, it emerged that the level of experience in PL writing can influence authors' opinions of Cochrane PLS guidance. P27 (who agreed with both statements on the completeness of Cochrane PLS guidance) specified that they had years of experience in PL writing for patients and that, as a result, they found Cochrane PLS guidance easy to follow. Similarly, P01, who assigned the highest SUS score to Cochrane PLS guidance (i.e. 100 out of 100) (Section 4.8.4), commented that PL writing represented an area of expertise for them.

Authors also used the comment section to highlight some of the limitations of Cochrane PLS guidance. P18's remarks on Cochrane PLS guidance were: "some of the sections required are a bit odd" and "I really wondered how much PPI [Patient and Public Involvement] input had been sought in terms of identifying relevant headings/content". P15 commented: "The guidance for PLS writing is too vague and, generally, not helpful". Unsurprisingly, P15 is also the participant who assigned the lowest SUS score to Cochrane PLS guidance (i.e. 7.5 out of 100) (Section 4.8.4).

Finally, a comment from P05 ("I have found the Cochrane Norway template more helpful than PLEACS") shed light on the fact that, while we had been asking questions on Cochrane PLS guidance in general, there might have been differences in terms of completeness and accuracy among the various documents that characterise the non-automated simplification approach adopted at Cochrane (Section 4.3).

4.8.3 Authors' Future Simplification Preferences

Of the 18 Cochrane authors who met the requirement for participating in this study, only 12 volunteered to conduct the Acrolinx editing task (Section 4.5) and to complete the post-session questionnaire on their interaction with this CL checker (Appendix H). Therefore, in this section, we will present the analysis of data collected from 12 participants only. The first question of the post-session questionnaire asked authors to report the ID number that had been assigned to them (e.g. P01). The second question asked authors to complete the SUS on their interaction with Acrolinx. Findings for this question will be reported in Section 4.8.4, where they will be compared with the SUS scores assigned to Cochrane PLS guidance.

Question 3 asked authors to indicate which type of authoring support they would use to check text readability when producing a PLS in the future, i.e.: (i) both Cochrane PLS guidance and Acrolinx; (ii) Acrolinx only; (iii) Cochrane PLS guidance only; (iv) other types of authoring support; or (v) no support at all. Since SUS results do not provide diagnostic information (Brooke 2013), this question (and the following two) were asked in order to complement SUS scores and to facilitate their interpretation (Section 4.7.2). Findings are reported in Figure 4.10, where the horizontal axis indicates the number of participants who selected each option.

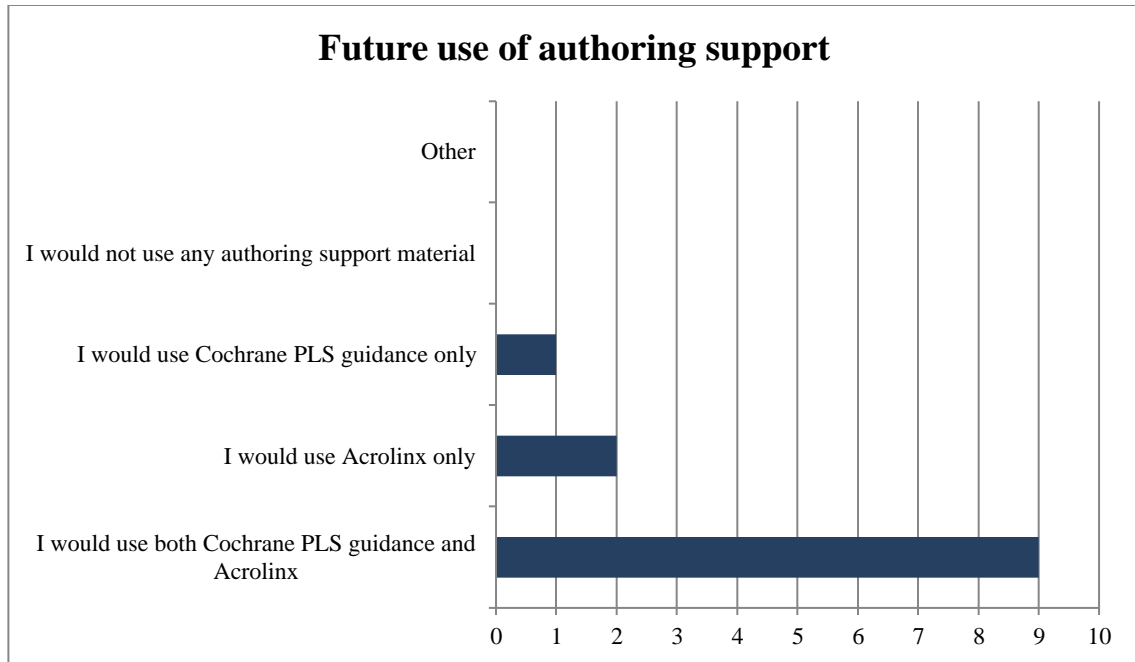


Figure 4.10: Future use of authoring support for the production of PLS

The majority of participants reported that, in the future, they would use both Cochrane PLS guidance and Acrolinx to produce a PLS, if they had the opportunity to do so. Only a small percentage of participants reported that they would use either Cochrane guidance only ($n=1$) or Acrolinx only ($n=2$). This result is not surprising when considering that, by using Acrolinx alone, authors would not have any indications on which content should be included in the PLS during the summarisation process. On the other hand, by using Cochrane PLS guidelines only, authors had to rely on their memory of language/style-related rules and could not receive feedback on the level of text readability achieved (Section 4.3). In other words, from these results, Cochrane authors show a future preference for complementing Cochrane PLS guidelines with the Acrolinx CL checker, i.e. for introducing a form of semi-automation into Cochrane's non-automated simplification approach. It should also be noted that no participant reported that they would produce a PLS without using any support. This result is not surprising considering that all the Cochrane authors who conducted the Acrolinx task were either health professionals or academics in the health field who were likely to be more familiar with the production of specialised medical language than with PL.

Question 4 focused on the rationale behind the future preferences shown by authors in question three. Finally, the fifth question asked authors if they had any comments. Responses to both questions were analysed jointly.

Among the authors stating that they would use both Cochrane PLS guidance and Acrolinx for the production of future PLS, there seems to be agreement about the fact that these tools have complementing features. P07 replied that Cochrane PLS guidance seems to be more tailored to the terms and abbreviations that are common in the specific context of evidence synthesis, while Acrolinx suggestions are more general since they can be applied to different types of texts. P28 replied that Acrolinx could be used as a supplement to Cochrane PLS guidance. Similarly, P04 replied that both a good structure and a simple language are needed, possibly referring to, on the one hand, Cochrane templates for PLS (Cochrane Norway 2017) and, on the other hand, to Acrolinx feedback on readability. P22 answered that they would check Cochrane guidance, but they would also use Acrolinx because it is more targeted. In addition, P20 commented on the possibility of using Acrolinx not just to check the readability of a PLS that has already been written using Cochrane guidance, but also to receive real-time feedback during the summarisation process:

I found that Acrolinx was useful to edit the existing PLS. If the software was enabled when creating the PLS then it would be very helpful to make is [sic] more readable.

P20 goes on to comment that “Acrolinx is good because it makes you think about simplifying the text and using shorter sentences”. This remark seems in line with previous studies showing that volunteers welcome feedback that can improve their skills (Section 4.3).

There seems to be some agreement also on the need to practise with Acrolinx for a while before becoming familiar with the tool. P19, who assigned the lowest SUS score to Acrolinx (i.e. 50 out of 100) (Section 4.8.4), specified that “it was tricky at first”. For this study, participants were assigned a warm-up task with Acrolinx before conducting the main editing task on their PLS (Section 4.5). Nonetheless, the time that authors dedicated to the warm-up varied greatly. In addition, it might be the case that the very

fact of being involved in a study (even though remotely and in their own environment) might have resulted in participants feeling under pressure. In turn, the stress might have hindered their familiarisation with the CL checker. In relation to this, P01 wrote:

I think until I am very confident to [sic] using Acrolinx I would need both. I would hope to move to just Acrolinx very quickly. I found the task quite stressful - almost like an exam but doing it in my own time would help. I thought it was intuitive and I enjoyed using it.

The impact that practice might have on authors' perceptions of Acrolinx was highlighted by P14, who commented that, after using Acrolinx several times, "you will know and probably you can do without". P14 was possibly referring to the fact that authors might be able to memorise the readability rules and then apply them without the need to see the issues flagged by the CL checker. However, each PLS is likely to contain different readability issues and the corresponding Acrolinx flags are likely to vary accordingly.

Some of the participants reporting that they would like to use both Cochrane PLS guidance and Acrolinx for the production of future PLS also identified several issues with the CL checker. The issues reported were:

- a. Flags can be ignored by accident and, once they are ignored, the software gives users the possibility to undo the action only for a few seconds. Afterwards, it is not possible to undo the action;
- b. The word *very* should not be flagged, as it should be used in PLS to signal when the quality of evidence is *very low*, among others (Cochrane Norway 2017) — Acrolinx flags *very* as a result of the rule on avoiding needless words.

This latter issue was caused by the fact that the Acrolinx software used for our study had not been tailored to Cochrane content (Section 4.3). Possibly due to this lack of tailoring, P14 reported that they would use the combination of Cochrane guidance and Acrolinx, but would not necessarily apply all the changes recommended by the CL checker. Two participants also reported that they would have preferred to use the British spelling rather than the American one. The decision to adopt the US English set of rules had been taken prior to the beginning of the PLS study, i.e. when the geographical location of the future participants was unknown (Section 4.3).

Two authors (P05 and P15) answered that they would only use Acrolinx in the future (Figure 4.10), if they had that choice. P05 provided as a reason the fact that Acrolinx readability suggestions were specific. However, in the comment section, this participant specified that they struggled to increase the Acrolinx readability score of their PLS because of the inability of the software to recognise medical terminology. They added that this issue could be solved by creating an Acrolinx term set for the medical field, thus again shedding light on the need to tailor the CL checker. P15 reported that they found the software very useful to improve both the readability and the style of the text. In addition, this participant seemed to appreciate Acrolinx suggestions on translatability and added that those suggestions could also be used for Cochrane podcasts:

I would also mention that this software could be useful to improve the scripts of the podcasts produced for the Cochrane reviews. Podcasts are also translated and this software could improve the ability of people understanding the podcast's scripts and translating them.

Unsurprisingly, this participant was the one who assigned the highest SUS score to Acrolinx (i.e. 97.5 out of 100) (Section 4.8.4). P15 also indicated that they had already been using an authoring support tool called Grammarly, thus providing an additional indication that Cochrane PLS guidance alone might not be enough.

In the comment section, P15 specified that, when applying Acrolinx suggestions for splitting long sentences, the result is an increase in the number of words. According to this participant, this might represent an issue since Cochrane PLS are supposed to contain no more than 400 words. However, our analysis of Cochrane guidance on PLS has shown that there is variability in the recommendations regarding the length of PLS (Section 4.3).

Finally, the only participant (P06) who reported that, in the future, they would only use Cochrane PLS guidance provided as a reason the fact that they did not find Acrolinx very easy to use. Unsurprisingly, this participant assigned the second lowest SUS score to the tool (i.e. 60 out of 100) (Section 4.8.4).

It should be noted that a technical issue arose when conducting the Acrolinx editing task remotely. Prior to this study, the language in the programs that authors would use (i.e. Microsoft Word for the editing task and Adobe Reader for the instructions) had been changed to English. However, the default language of the researcher's computer was Italian. Accordingly, when P07 opened multiple Acrolinx help files and then tried to close them, the message asking if they wanted to close all tabs or just one tab appeared in Italian. In the post-session comments, this participant wrote that Acrolinx should not display help files in another language. However, it can be easily assumed that P07 referred to the message about closing the tabs rather than to the Acrolinx help files, where all the content was always provided in English.

4.8.4 SUS Scores

SUS scores range from 0 to 100. The higher the SUS score, the higher the satisfaction of the users of a product. Eighteen authors completed the SUS on their satisfaction with Cochrane PLS guidance (Appendix D). Of these, only 12 carried out the Acrolinx editing task and then completed the SUS on Acrolinx (Appendix H). Table 4.2 contains descriptive statistics for the SUS scores assigned by all the Cochrane authors involved in this study, while Table 4.3 only shows results from the 12 authors who also conducted the Acrolinx task, in order to allow a more direct comparison between the SUS scores assigned to Cochrane PLS guidance and to the Acrolinx software. As reported in Section 4.7.2, SUS scores assigned to Cochrane PLS guidance should be interpreted with caution, since authors had to rely on their memory of the interaction with the guidelines.

It is important to specify that, even though the SUS scores are collected by means of Likert scales (similarly to the questions on the completeness of Cochrane PLS guidance, described in Section 4.8.2), results from the ten Likert-type statements that compose the SUS were combined into a single score for each participant prior to being analysed. Scores from the ten SUS Likert-type statements need to be combined because, taken individually, they do not relate to any specific characteristic of a system (Brooke 2013). As specified in Boone and Boone (2012), composite scores should be treated as interval data for which the descriptive statistics recommended are the mean and the

standard deviation (SD). However, we also included the median because it is less influenced by extreme values than the mean and can therefore provide a comparison that is less affected by individual differences (Leys et al. 2013).

Product	SUS score (mean)	SUS score (median)	SUS score (SD)	N
<i>Cochrane guidance</i>	60.97	63.75	21.91	18
<i>Acrolinx</i>	75.41	78.75	14.49	12

Table 4.2: Descriptive statistics for the SUS scores assigned by the total sample of participants

Product	SUS score (mean)	SUS score (median)	SUS score (SD)	N
<i>Cochrane guidance</i>	62.29	70	26.53	12
<i>Acrolinx</i>	75.41	78.75	14.49	12

Table 4.3: Descriptive statistics for the SUS scores assigned by the 12 participants who conducted the Acrolinx editing task

The mean and the median rates in Table 4.2 indicate that the use of Acrolinx resulted in a higher level of authors' satisfaction, compared with the adoption of Cochrane PLS guidelines. This difference remains but becomes less pronounced when only considering the data collected from the 12 authors who also conducted the Acrolinx editing task, as shown in Table 4.3. If we take the means in Table 4.3 and apply the adjective rating scale developed by Bangor, Kortum and Miller (2008, pp. 586-588), where each range of SUS scores is associated with an adjective descriptor, we observe that the Acrolinx CL checker was rated as *good*, while Cochrane PLS guidance was rated as *OK*.

To determine if the difference between the means of the SUS scores was statistically significant, a paired t-test was conducted. The paired t-test is a parametric statistical test employed to compare matched pairs of data (Helsel and Hirsch 2002), i.e. means that are contributed by the same group of participants (Lazar, Feng and Hochheiser 2010, p. 76). Therefore, only the data from the 12 authors who completed both SUS questionnaires were included in this statistical analysis (Table 4.3). The paired t-test is based on the assumption that the distribution of the paired differences is normal (Helsel and Hirsch 2002). To check for this assumption, we adopted the Shapiro-Wilk

normality test, which showed that the data set was normally distributed ($p=0.51$)²⁷. Since the assumption of normality was met, it was possible to conduct the paired t-test. The paired t-test revealed that the difference between the mean SUS score assigned to Cochrane PLS guidance and the mean SUS score assigned to Acrolinx was not statistically significant: $t(11)=1.25$, $p=0.23$. However, due to the small sample size ($n=12$), it is not possible to exclude a Type II error, namely the inability to observe a true effect (Lieberman and Cunningham 2009).

Overall, even though the difference in satisfaction scores was not statistically significant, the descriptive statistics (in Tables 4.2 and 4.3) and the adjective descriptors seem to confirm the data collected through the open-ended questions submitted to authors (Sections 4.8.2 and 4.8.3). In particular, it emerged that: (i) Cochrane authors did not show a high level of satisfaction with Cochrane PLS guidance, possibly as a result of their vagueness, contradictions, and incompleteness; and (ii) Cochrane authors seemed more satisfied with the Acrolinx CL checker, even though this was their first encounter with this product. As discussed in Section 4.6, the adoption of Cochrane PLS guidelines and the use of Acrolinx did not represent two comparable simplification scenarios — Cochrane PLS guidelines were adopted in the process of summarisation/simplification that led to the production of PLS from scratch, while Acrolinx was used to check the readability and translatability of the already produced PLS, and to edit them accordingly. Therefore, it is not possible to conclude that authors would be more satisfied if the Acrolinx CL checker replaced Cochrane PLS guidelines. However, results seem to indicate that introducing a CL checker might, over time, turn the simplification approach into a more satisfactory experience for authors.

4.9 Discussion and Summary of the Study on Cochrane Authors' Satisfaction

In this chapter we have described an experimental study aiming to answer RQ1, namely whether introducing Acrolinx as a CL checker into Cochrane's non-automated simplification approach (thus rendering it semi-automated) would boost authors' satisfaction (DV1). To the best of our knowledge, no previous studies have been

²⁷ We have used an alpha level of 0.05 for all the statistical tests reported in this thesis.

conducted on Cochrane authors' satisfaction with the standard non-automated approach to PLS production, nor on the impact of introducing a CL checker to edit the PLS. In this final section, we will delve into the main findings, discuss their relevance for Cochrane, and briefly introduce the next chapters of this thesis.

The evidence gathered has allowed us to gain a broader understanding of the standard PLS production workflow at Cochrane. Our previous analysis of Cochrane PLS guidance had shown that recommendations for authors can be found in a variety of documents, and that they often show contradictions and vagueness (Section 4.3). Data collected from Cochrane authors of PLS with a health background have complemented our previous analysis. In particular, based on authors' rankings of the completeness of Cochrane PLS guidelines, and on their comments regarding vagueness and lack of input from the public, it emerged that a revision of Cochrane PLS guidelines in terms of characteristics considered, level of detail, and soundness might be beneficial. Furthermore, we found that not all authors are provided with the same set of guidelines. A similar result also emerged from the report produced by Glenton (2017, p. 5):

[T]he CRGs [Cochrane Review Groups] gave different levels of direction to their review authors regarding how to write PLSs. One CRG gave no specific information, but assumed that review authors would see reference to the Plain Language Expectations for Authors of Cochrane Summaries (PLEACS) in RevMan; while two CRGs directed review authors to the MECIR [Methodological Expectations of Cochrane Intervention Reviews] standards, which include reference to the PLEACS. Five CRGs had developed some sort of PLS guidance, including standard headings and checklists that they expected review authors to use.

This finding could explain the inconsistencies and variations that have been identified across PLS (Glenton et al. 2010). Anecdotally, we also observed that there might be differences in Cochrane authors' opinions of the various documents that characterise the non-automated simplification approach (Section 4.3). Similar differences in opinions, specifically between the *PLEACS* and the PLS template developed at Cochrane Norway (2017), were also observed by Glenton (2017).

Therefore, there seems to be a need to combine the entire body of Cochrane guidelines, removing contradictions, and ensuring that all PLS authors are assigned the same set of recommendations, especially considering that most authors check the guidelines for each PLS they write (mainly at the beginning or at the end of the authoring task) (Section 4.8.2). For language/style-related guidelines, the introduction of a CL checker which automatically and consistently flags readability and translatability issues in a text might offer a solution to these needs by ensuring a higher level of consistency, providing examples to follow, and reducing authors' need to manually check PL guidelines.

Unsurprisingly then, we found that most authors would welcome the introduction of the Acrolinx CL checker, but in combination with Cochrane PLS guidelines, mainly because these two products are seen as having complementing features. Here we suggest that, to maximise the benefits of this potential integration, Cochrane PLS guidelines might focus on content only and be used during the process of summarisation (i.e. when authors need to summarise an entire systematic review from scratch); on the other hand, with its focus on language and style, the Acrolinx CL checker, or similar, might be used for the subsequent simplification phase (i.e. to ensure that the summary is written in PL), either in Microsoft Word or in RevMan. With this integration, authors could save time and reduce effort since they would not have to check language-related guidelines during or before/after the simplification task. However, as our findings showed, the CL checker should be tailored to suit the needs of Cochrane authors (e.g. in terms of preferred spelling) and the high specificity of Cochrane medical content.

In terms of satisfaction as assessed by the SUS, Cochrane authors seem slightly more satisfied with the Acrolinx CL checker than with Cochrane PLS guidelines, which might explain why they feel the need to introduce the CL checker into the non-automated simplification approach. Overall, the quantitative and qualitative data analysed in this chapter seem to point to the acceptance of our alternative hypothesis (Section 4.3), according to which rendering Cochrane's current non-automated simplification approach semi-automated would be beneficial in terms of authors'

satisfaction. This beneficial effect might encourage authors to continue contributing simplified health content for Cochrane, in line with the organisation's accessibility goals (Section 1.1).

We found that Cochrane authors need some form of guidance or (technological) assistance while simplifying — none of the authors in our study reported that they would write a PLS without any form of support. Anecdotally, we also observed that the feedback provided by Acrolinx on text readability/translatability might be welcomed from authors as a way to develop their PL writing skills further. The importance of receiving feedback on PL writing, e.g. from authors more familiar with PLS, also emerged in Glenton's (2017) pilot study. These findings are not surprising considering that Cochrane authors tend to be experts in the health field and, therefore, more familiar with specialised medical language than PL.

In Section 4.3, we specified that Acrolinx is commercial product whose cost might be prohibitive for non-profit organisations like Cochrane. Nyberg, Mitamura and Huijsen (2003, p. 249) discuss some of the costs associated with the introduction of a CL checker into an existing document production process:

An organization can either license and customize an existing CL product, or bear the expense of designing, developing, and maintaining their own CL. Designing a new CL involves several phases of linguistic analysis and terminology development. In addition, development may include the in-house construction or purchase of a CL checker [...]. The CL must also be maintained: it must continuously adapt to changing needs and wishes, new terminology and new standards, etc.

Some tools are freely available online to assist authors. Examples are Simplish²⁸, Rewordify²⁹, or Article Simplifier³⁰, which allow users to upload or copy a text, and provide alternatives in simple language, although exclusively at the vocabulary level. Moreover, since these tools are not tailored to (Cochrane) medical content (as in the case of the Acrolinx CL checker adopted in this study), authors would need to use their common sense and intuition in deciding which of the recommended changes are

²⁸ Simplish is available at: <https://bit.ly/2CSxAUC> [Accessed 12 December 2018].

²⁹ Rewordify is available at: <https://bit.ly/2yovIV3> [Accessed 12 December 2018].

³⁰ Article Simplifier is available at: <https://bit.ly/2SrUmWW> [Accessed 12 December 2018].

appropriate. A freely available online tool that is tailored to Cochrane content and sets of PLS guidelines is the Plain Language Summary Tool³¹, which, for each section of a PLS, lists the guidelines to be followed and their explanations (Harniss et al. 2013). However, differently from a CL checker, this tool is not able to automatically verify compliance with the guidelines. It should also be considered that, regardless of the tool adopted as a CL checker or authoring support, authors are likely to need some time to familiarise themselves with it.

Finally, in this chapter we have focused on authors' preferences and satisfaction, i.e. on the subjective component of usability (Lindgaard and Kirakowski 2013). The level of satisfaction experienced by users may not be an accurate reflection of how successful the use of a product is (Brooke 2013). This statement is in line with previous studies demonstrating that the three components of usability (i.e. satisfaction, effectiveness and efficiency) are either not correlated or weakly correlated (Frøkjær, Hertzum and Hornæk 2000). In Section 4.6, we explained the reason why efficiency was not included in this investigation. Instead, we will focus on effectiveness/goal completion (DV2), and in particular on the impact of introducing a CL checker on the readability (Chapter 5), comprehensibility (Chapter 6), and machine translatability (Chapter 7) of Cochrane PLS. The next chapter will deal with the readability of Cochrane PLS.

³¹ The Plain Language Summary Tool is available at: <https://bit.ly/2DzyzbO> [Accessed 12 December 2018].

CHAPTER 5

ASSESSING THE READABILITY OF COCHRANE PLS AND ABSTRACTS

5.1 Aims of the Study on Text Readability and Overview of the Chapter

This chapter presents an experimental study which was carried out with the main aim of determining whether introducing the Acrolinx CL checker into Cochrane's non-automated simplification approach could increase text readability. A secondary goal was to provide empirical evidence of the benefits of simplification (regardless of the approach adopted) on text readability. As reported in Section 1.2, readability (or DV2.1) is one of the three components of effectiveness (or goal completion), which is the DV2 of the empirical investigation described in this thesis. In Figure 5.1, we highlighted readability to show how the experiment described in this chapter relates to our broader investigation.

This chapter will begin with a summary of related work. Subsequently, we will present the motivation for examining the readability of Cochrane PLS, the RQ, and the research hypotheses of this experiment. We will then describe the experimental materials and the method adopted for measuring text readability. Finally, we will present the analysis of the data and discuss the findings.

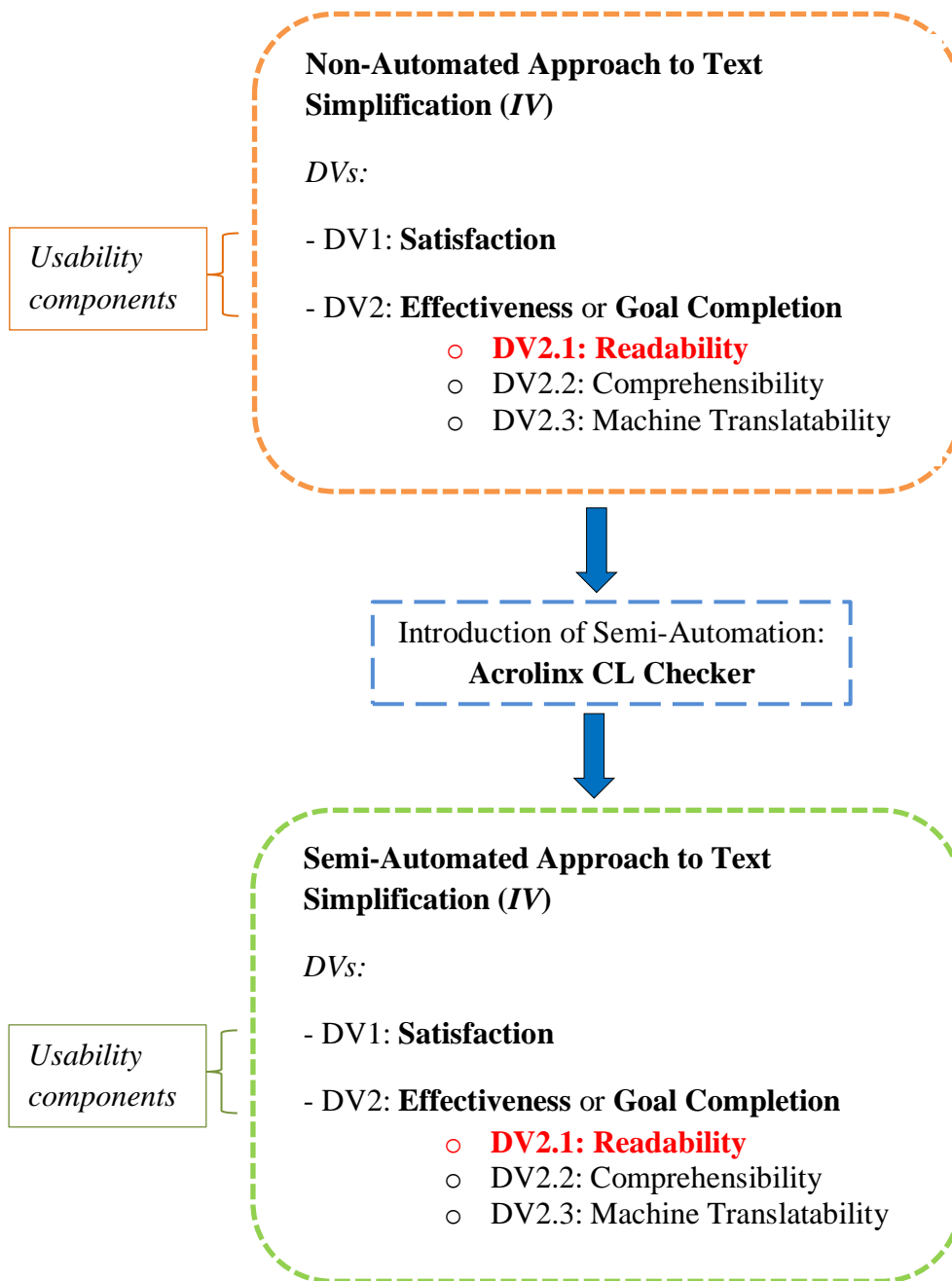


Figure 5.1: Text readability as DV2.1

5.2 Related Work on Readability

Text readability is one of the goals of text simplification (Shardlow 2014), and is defined as “the ease of understanding or comprehension due to the style of writing” (Klare 1963, quoted in DuBay 2007, p. 5). In other words, readability is determined by

those characteristics of the text (ranging from vocabulary, to syntax, to cohesion) which are expected to make it easy to comprehend (Collins-Thompson 2014). Readability is linked with *reading ease* or *comprehension easability*, which are often presented as synonyms (Flesch 1948; McNamara et al. 2014). On the other hand, the terms *text difficulty* or *text complexity* are adopted as antonyms of *readability* (Hiebert 2002; Crossley, Greenfield and McNamara 2008; Temnikova 2012).

Readability differs from legibility, which is determined by typeface and layout (DuBay 2007). Readability also differs from actual comprehension (or understanding), which is a function of the text, the reader's characteristics (such as native language, prior knowledge, motivation, or disabilities), the purpose of reading, and the broader sociocultural context (Snow 2002; Doherty 2012; Shardlow 2014). As Rello et al. (2012) point out, sometimes readability and reading comprehension are used interchangeably because the former is expected to affect the latter. However, the need to distinguish between the two concepts emerges when considering, for example, that the adoption of shorter words has been found to increase comprehension of texts on different topics amongst readers with dyslexia, but not amongst readers without dyslexia (Rello et al. 2013). Another example of the difference between readability and comprehension is reported in McNamara et al. (2014), who describe the *reverse cohesion effect*, according to which readers with low prior knowledge of a topic tend to benefit from the introduction of cohesive devices (e.g. connectives that clarify the relationship between sentences). On the other hand, more knowledgeable readers seem to benefit from cohesive gaps, which lead them to use their prior knowledge to fill the gaps by making inferences about the text.

The difference between readability and comprehension is described in Leroy, Kauchak and Mouradi (2013) as difference between perceived difficulty and actual difficulty of texts. More precisely, the authors (ibid.) state that modifying specific text characteristics with the goal of increasing readability (e.g. by reducing sentence length) might make texts look easier, but actual comprehension might not always benefit from the simplification edits. The two-fold focus on readability, on the one hand, and on comprehension, on the other hand, emerges as important when evaluating the

effectiveness of both manual and automatic text simplification — Štajner, Mitkov and Saggion (2014, p. 1) argue that automatic text simplification systems “are either evaluated for: (1) the quality of the generated output, or (2) the effectiveness/usefulness of such simplification on reading speed and comprehension of the target population”. In line with these studies, in this thesis the analysis of readability (described in this chapter) has been conducted separately from the analysis of actual reading comprehension (Chapter 6).

Studies on readability started in the United States at the beginning of the last century and focused on English texts (Klare 1974; DuBay 2004). Subsequently, work on readability assessment began to be conducted for languages other than English, such as French (Tharp 1939), Dutch (Douma 1960, quoted in Klare 1974), Spanish (Spaulding 1951), and Hindi (Bhagoliwal 1961). The main goal of these studies was to identify texts that would match the abilities of readers with varying levels of reading proficiency (Feng, Elhadad and Huenerfauth 2009), since a mismatch results in the reader failing and/or refusing to use the text (Zamanian and Heydari 2012). Readability studies have traditionally focused on educational settings (Duran et al. 2007), where the adoption of texts that match the reading abilities of students is pivotal for the development of their literacy skills and, ultimately, for learning (Gallagher, Fazio and Gunning 2012; Fisher and Frey 2015). Science textbooks have been shown to be particularly challenging for students due to their difficult vocabulary (usually of Latin and Greek origin even in English texts) and complex syntax, as well as their higher reliance on inferential thinking and prior knowledge (Best et al. 2005; Gallagher, Fazio and Gunning 2012). It is usually recommended that teachers avoid texts that are either too easy to read (because students would be bored) or overly difficult (as students might find the reading experience frustrating) (Beinborn, Zesch and Gurevych 2012). Vygotsky’s (1978) zone of proximal development (characterised by reading skills which are not still possessed but are beginning to mature) has often guided teachers in the selection of texts with readability levels that can ensure students’ optimal growth (STAR Reading 2008).

Readability also plays an important role in the field of health communication, where texts need to be written in a way that allows lay people/patients to understand

them and, ultimately, to apply their content (Leroy and Endicott 2011). Soergel, Tse and Slaughter (2004) list the benefits of increasing the readability of medical texts for health information consumers, including more informed decision-making and higher compliance with the instructions provided by health practitioners. However, in the United States alone, more than 300 studies have found that the reading level of health-related materials (from informed consent forms to medication package inserts) exceeds the reading ability of the intended audience (Kindig, Panzer and Nielsen-Bohlman 2004). In their systematic review of studies assessing the readability of both print and online nutrition-related materials produced in the United States, Carbone and Zoellner (2012) showed that, despite recommendations to write materials for readers at a fourth- to eight-grade level, the majority of texts matched the reading skills associated with a ninth- or higher grade level. Moreover, online health-related texts were found to be more difficult than their print counterparts (*ibid.*). Similarly, Leroy, Eryilmaz and Laroya (2006) observed that health information on English websites tends to be written in a way that matches a 10th grade or even higher reading level. Individuals with low levels of health literacy represent a particularly vulnerable group (Weiss 2007).

Various features can increase text difficulty. Adnan, Warren and Orr (2010) observed that the section on clinical management in electronic discharge summaries contains text features that make it difficult to read for patients/lay people, such as abbreviations and specialised medical vocabulary. To estimate the difficulty of medical vocabulary for the average health information consumer, the authors (*ibid.*) used the scores provided in the Open Access Collaborative Consumer Health Vocabulary³². Interestingly, and in line with the hypotheses of our experiment (Section 5.3), Adnan, Warren and Orr (2010) suggest the introduction of computer-based support that provides authors of electronic discharge summaries with interactive feedback on the readability of their texts. In Rosemblat et al. (2006), medical vocabulary also emerged as a factor able to predict text difficulty for lay audience, according to the health communication experts surveyed. Kim et al. (2007) compared the semantic and syntactic characteristics

³² The Open Access Collaborative Consumer Health Vocabulary is available at: <https://bit.ly/2RYYApm> [Accessed 12 December 2018].

of electronic health records with those of difficult texts (i.e. abstracts of scientific journal papers) and easy-to-read texts (extracted from consumer health websites). The authors (ibid.) found electronic health records to be more similar to difficult texts (than to easy-to-read texts) in terms of semantic and syntactic features. Kandula and Zeng-Treitler (2008) asked health literacy and clinical experts to evaluate the readability of different sets of health-related documents, from education materials to journal articles. Their experts recognised vocabulary, sentence structure and voice as characteristics affecting readability.

Cohesion has also emerged as a text feature affecting readability. Interestingly, traditional readability formulas (based on word and sentence length) tend to predict a decrease in readability when cohesion-oriented edits are applied (Section 5.5.1) since these edits usually involve the addition of words (e.g. connectives) which increase sentence length, or the replacement of pronouns with long, unfamiliar words (to reduce ambiguity) (McNamara et al. 2014; Schriver 2017). However, cohesion-oriented edits have been shown to facilitate reading comprehension (particularly of low-knowledge readers), thus providing further indication of the difference between readability and comprehension. For instance, Liu and Rawl (2012) observed that cohesive colorectal cancer screening information is read faster and understood more (as shown by answers to yes-no questions) than its low-cohesion counterpart.

While readability studies mainly focused on English texts, it is worth mentioning that similar results were observed when assessing the readability of health-related texts in other languages. For instance, using human evaluation to assess the readability of a Japanese text on chronic suppurative otitis media, Sakai (2013) reported that medical vocabulary was perceived by her participants as hard to understand. The author (ibid.) combined participants' evaluations of text readability with their actual comprehension (assessed by means of a cloze test). In line with previous studies, she found that results on readability and comprehension did not always coincide. Regarding medical texts in Swedish, Abrahamsson et al. (2014) assessed the predicted difficulty of comprehending specialised medical vocabulary by measuring the frequency of both entire words and substrings of words (based on the compounding nature of Swedish). The authors (ibid.)

then replaced difficult words with PL synonyms available in the Medical Subject Headings thesaurus³³, and measured readability by means of the LIX readability formula (influenced by word length) and the OVIX (the word variation index). They found that, while readability increased according to OVIX, it decreased according to the LIX, possibly because PL synonyms were longer than their specialised Greek and Latin versions.

In summary, scholars seem to agree that, in order to reduce text difficulty (and, in turn, increase readability), authors should adhere to a series of recommendations, which mainly include the use of short sentences and simple vocabulary (Eltorai et al. 2015). Compliance with these recommendations is expected to decrease the reading skills required to understand texts in educational and health settings. Nonetheless, previous research has also shown that simplifying texts with a view to increasing readability is not guaranteed to be beneficial in terms of readers' comprehension (Terranova et al. 2012). Moreover, as will be discussed in Section 5.5.1, text characteristics that have been excluded from traditional readability analyses (e.g. cohesion) can play an important role in enhancing comprehension.

5.3 Motivation for Assessing the Readability of Cochrane PLS, Research Question, and Research Hypotheses

In recent years, there seems to be growing interest in the readability of Cochrane PLS, which are the most widely consulted components of systematic reviews (Gyte and Struthers 2015), and are supposed to be easy to read and comprehend, particularly for lay people/health information consumers (Kadic et al. 2016). For example, during the 22nd Cochrane Colloquium, health information consumers were involved in training on the evaluation of various characteristics of Cochrane PLS (including their readability) (Struthers, Lyddiatt and McIlwain 2014). Furthermore, as part of the Cochrane Colloquium held in Vienna the following year, a workshop was conducted to help authors implement Cochrane PL guidelines (Cochrane Colloquium Vienna 2015).

³³ The Medical Subject Headings thesaurus is available at: <https://bit.ly/1S6n23S> [Accessed 12 December 2018].

Despite these efforts, several studies have shown that the text characteristics of Cochrane PLS might make them difficult to read. Flodgren (2016) used an online readability calculator and Microsoft Word proof-reading tool to assess the readability of a set of 143 PLS produced by 50 different Cochrane Review Groups. Her findings showed that Cochrane PLS were characterised by textual features that increased their perceived difficulty, such as passive voice, long sentences, and a high percentage of hard words (namely words containing more than three syllables). Karačić et al. (2017) used the SMOG index — a readability formula which calculates the number of words with three or more syllables and provides an indication of the education level required to read a text (Mc Laughlin 1969) — to assess the readability of Cochrane PLS and of other formats of Cochrane summaries (namely press releases, abstracts, and Cochrane clinical answers). Their results showed that, even though PLS had a lower SMOG Index than the other summary formats, they still required more than 14 years of education to be comprehended. Smith (2018) describes the integration of Cochrane PLS into PubMed Health, and reports that reading them would require at least college level literacy skills, as shown by the results of Readability Analyzer³⁴ (an online readability calculator).

In addition to showing low readability levels, Cochrane PLS have been found to present inconsistencies. Kadic et al. (2016) conducted a systematic analysis on the extent to which 1,738 PLS published between 2013 and 2015 adhered to Cochrane guidelines (and, more precisely, to the *PLEACS*) implemented through the standard non-automated simplification approach (Section 4.3). The authors (ibid.) observed that Cochrane PLS differed widely in terms of both length and adherence to the *PLEACS*. More precisely, it emerged that not a single PLS adhered to the entire list of guidelines, and that certain review groups had their own standards for producing PLS, which differed from the *PLEACS*. Kadic et al. (2016) argue that one possible explanation for this low level of adherence might be the fact that the Cochrane PL guidelines are neither ideal nor evidence-based.

³⁴ Readability Analyzer is available at: <https://bit.ly/2Bnthws> [Accessed 12 December 2018].

In summary, studies on Cochrane PLS shed light on the need to increase both their readability and the consistency with which simplification edits are implemented. Addressing this need would be in line with Cochrane’s goal of making evidence accessible/readable (*Strategy to 2020* 2013). Leroy, Kauchak and Mouradi (2013) argue that, by semi-automating text simplification with the introduction of a writing support tool (e.g. as a Microsoft Word plugin), health professionals with no linguistics background (such as the Cochrane authors of our investigation) might be able to conduct the simplification tasks more effectively, i.e. to produce more readable texts. In Section 4.3 we described the characteristics of the Acrolinx CL checker. In particular, we discussed the benefits of adopting this tool, such as: (i) automatic flagging of readability issues; (ii) checking of documents against one consistent set of CL rules; and (iii) availability of suggestions and examples on how to solve readability issues. The usage of the CL checker is expected to increase the level of readability achieved since: (i) authors can avail of simplification examples to follow; (ii) authors can avoid the inconsistent application of rules; (iii) authors do not run the risk of forgetting to implement simplification recommendations.

In Chapter 4, we hypothesised and showed that introducing the Acrolinx CL checker into the Cochrane’s non-automated simplification approach might boost authors’ satisfaction. Here, in line with Leroy, Kauchak and Mouradi (2013), we tested the hypothesis that the introduction of the CL checker would also increase authors’ effectiveness (DV2) and, more specifically, the level of readability that they achieved (DV2.1) as a result of simplification. *Effectiveness* is defined as “the accuracy and completeness with which users achieve certain goals” (ISO 9241-11:2018, 3.1.12). As reported in Chapter 1 (Section 1.2), the RQ associated with effectiveness (or goal completion) is the following:

RQ2: Does semi-automating a non-automated simplification approach by introducing a CL checker increase authors’ effectiveness?

RQ2 was then further segmented into three other questions (one per each of the goals of readability, comprehensibility, and machine translatability) (Figure 5.1). The RQ associated with readability (RQ2.1) is the following:

RQ2.1: Does semi-automating a non-automated simplification approach by introducing a CL checker increase readability?

The corresponding research hypotheses are:

H0: Semi-automating a non-automated simplification approach by introducing a CL checker does not increase readability.

H1: Semi-automating a non-automated simplification approach by introducing a CL checker increases readability.

In this chapter, we will describe the experiment conducted to address RQ2.1.

5.4 Experimental Materials

For the purposes of our readability experiment, we used the texts produced by the 12 Cochrane authors who took part in the study described in Chapter 4. More precisely, our readability analysis was conducted on the following experimental materials:

- (i) a corpus of 12 PLS produced with Cochrane's non-automated simplification approach (henceforth *non-automated PLS*);
- (ii) a corpus of the same 12 PLS edited by using Acrolinx, i.e. after introducing semi-automation in the simplification approach (henceforth *semi-automated PLS*);
- (iii) a corpus of 12 abstracts extracted from the same systematic reviews as the PLS.

The analysis of the level of readability achieved in the three corpora can be assigned to the category of product-oriented research (Saldanha and O'Brien 2013). In this experiment, we adopted a corpus-based/deductive approach, in that we started with a precise null hypothesis (Section 5.3), and examined the corpora for evidence for or against it (Saldanha and O'Brien 2013, p. 62).

As stated in Section 5.1, the primary goal of the experiment described in this chapter was to assess and compare the level of readability achieved before and after the

introduction of Acrolinx. However, we expanded our readability analysis to also include abstracts. Similar to PLS, abstracts precede Cochrane Systematic Reviews and are produced with the intent of summarising their content (Higgins and Green 2011). However, differently from PLS, Cochrane abstracts mainly target health practitioners (ibid). In other words, the primary focus of abstracts is to summarise (rather than simplify for the lay public) the information contained in the systematic reviews. Abstracts can therefore be assigned to the category of authentic texts, namely texts that did not undergo a process of text simplification (Crossley et al. 2007b). Abstracts represented the baseline condition — they were used to determine whether text simplification (regardless of being non-automated or semi-automated) results in an increase in readability when compared with lack of simplification efforts. As argued in Saldanha (2009, p. 3), “the most rigorous counting of linguistic features is meaningless unless we can provide a relative norm of comparison”.

Prior to expanding our readability analysis to abstracts, we contacted the 12 Cochrane authors who had conducted the Acrolinx editing task (Section 4.8.3) and asked them if they had also authored the abstracts of the same systematic reviews, which was the case for all the authors. This check ensured that the three corpora had a comparative (or within or repeated) organisation (McNamara et al. 2014, p. 157), in line with the within-subject design adopted in the authoring study (Section 4.6). In other words, the texts in the three corpora were not independent since they were produced by the same authors under three different conditions: 1) lack of simplification efforts (abstracts); 2) simplification before the introduction of Acrolinx (non-automated PLS); 3) simplification after the introduction of Acrolinx (semi-automated PLS). This within or repeated organisation informed the statistical tests described in Section 5.6.

With regard to the length of the texts, descriptive statistics for number of words are reported in Table 5.1, divided by corpus. It can be observed that, on average, semi-automated PLS and non-automated PLS contained around the same amount of words. On the other hand, abstracts were longer than PLS in both corpora.

Corpora	Mean	SD	Min	Max
<i>Non-automated PLS</i>	437	120.95	302	694
<i>Semi-automated PLS</i>	436.41	132.85	295	712
<i>Abstracts</i>	636.75	146.34	361	867

Table 5.1: Descriptive statistics for text length (number of words), divided by corpus

A corpus has been defined as “a set of written, representative and balanced, computationally readable texts that form a reasonable point of departure as a thematically related language variety, register, genre, or text-type” (McNamara et al. 2014, p. 146). Here we break down this definition and show how it relates to the three corpora used in our experiment (i.e. abstracts, non-automated PLS, and semi-automated PLS). Firstly, all our texts were *written* and *computationally readable* (as .doc documents). Secondly, texts in the three corpora were *thematically related* to each other since they fell under the common theme of Cochrane health content on the impact of treatments and surgical interventions (Section 1.1). Texts also belonged to the same expository *genre* since the goal of Cochrane is to increase health knowledge and facilitate healthcare decision making by means of systematic reviews and their summaries (About Cochrane 2018).

As McNamara et al. (2014) point out, the terms *representative* and *balanced* are closely related to the notion of *thematically related*. With regard to representativeness, the corpora analysed contained texts that fell strictly under the category of Cochrane PLS or Cochrane abstracts. In other words, summaries from other health-related websites were excluded from the analysis since assessing the level of readability of online health content in general was beyond the scope of our investigation. In terms of balance, the number of texts in each of our three corpora (i.e. 12) was very limited, particularly when considering that, as of 2018, a total of 7,572 systematic reviews (and corresponding PLS and abstracts) are available in the Cochrane Database (Cochrane Database of Systematic Reviews 2018). The number of texts available to us was determined by the number of Cochrane authors who volunteered to use Acrolinx to edit

their previously produced PLS (Section 4.8.3). This limited number of texts per corpus did not allow us to generalise our conclusions (Section 8.4).

Representativeness and balance also apply to differences within each text, for instance between introduction and conclusion (McNamara et al. 2014). Therefore, following common practice in the field of corpus linguistics (Saldanha and O'Brien 2013), we adopted PLS and abstracts in their entirety as units of analysis. Finally, our three corpora can be considered as *a reasonable point of departure* in that they were not created to act as reference points (such as the British National Corpus) (McNamara et al. 2014), but rather to achieve the practical and contained goal of answering RQ2.1 (Section 5.3).

5.5 Method Adopted for the Measurement of Text Readability

In this section, we will present the method (or tool) adopted to analyse text readability, namely Coh-Metrix. It should be noted that our intent is not to provide a thorough description of Coh-Metrix components (which is available in McNamara et al. [2014]), but rather to highlight the most important features of the tool, and explain how it was applied to our experiment. We will also discuss the differences between Coh-Metrix and traditional readability formulas, thus providing a rationale for our selection of Coh-Metrix. Subsequently, we will discuss the Coh-Metrix variables (indices and measures³⁵) adopted in our readability analysis.

5.5.1 Characteristics of Coh-Metrix and Rationale behind its Use

Coh-Metrix is a theoretically grounded computational tool which provides multiple measures for the automatic analysis of texts (Dowell, Graesser and Cai 2016). The Coh-Metrix output allows for the scaling of texts on readability (versus difficulty), with the ultimate goal of matching texts to readers (McNamara et al. 2014). In other words, Coh-Metrix can be used to predict the level of text readability (or difficulty) for the intended readers, although actual reading comprehension will also be determined by the reader's

³⁵ In line with McNamara et al. (2014), we adopted the term *measure* to refer to a theoretical construct, such as referential cohesion. The terms *index* and *indices* were used to describe the ways in which Coh-Metrix assessed measures (as in the case of the index *argument overlap*, used to measure referential cohesion). Finally, the term *variable* was used to refer to both indices and measures.

characteristics, among others (as discussed in Section 5.2). Coh-Metrix has been mainly developed for and used on English texts. However, versions of this tool for Chinese³⁶, Spanish and Brazilian Portuguese³⁷ texts are also available (Quispesaravia et al. 2016). Moreover, Tonelli, Manh and Pianta (2012) describe an adaptation of Coh-Metrix for Italian (called *Coease*³⁸).

Coh-Metrix was originally developed in 2002 with the aim of measuring text cohesion, namely the presence of text elements (such as connectives) that “guide the reader in interpreting the substantive ideas in the text, in connecting ideas with other ideas, and in connecting ideas to higher level global units (e.g., topics and themes)” (Graesser et al. 2004, p. 193). Cohesion interacts with the reader’s characteristics (such as reading skills and prior knowledge) in the formation of a coherent mental representation of the text (ibid.), hence the prominence originally attributed to this text characteristic.

However, Coh-Metrix soon expanded its analysis to also include: 1) words/vocabulary; 2) syntax; 3) textbase/propositions; 4) situation (or mental) model; and 5) genre. Each of these levels has the potential to hinder (or facilitate) comprehension (Dowell, Graesser and Cai 2016). For instance, unfamiliar, rare words (such as medical vocabulary) can make entire sentences difficult to comprehend (Graesser, McNamara and Kulikowich 2011). Various lexicons are therefore incorporated in Coh-Metrix for word analysis, such as the MRC Psycholinguistic Database, CELEX Word Frequency, WordNet, Parts of Speech, Special-Purpose Word Categories (McNamara et al. 2014). Similarly, it is difficult to generate meaning from long sentences, often containing numerous subordinate clauses (ibid.). Two syntactic parsers have therefore been implemented in Coh-Metrix: the Apple Pie parser and the Charniak parser (ibid.).

³⁶ The version of Coh-Metrix for Chinese texts is available at: <https://bit.ly/2I3RYag> [Accessed 12 December 2018].

³⁷ The version of Coh-Metrix for Brazilian Portuguese texts is available at: <https://bit.ly/2JuMwMm> [Accessed 12 December 2018].

³⁸ Coease is available at: <https://bit.ly/2yKHgMf> [Accessed 12 December 2018].

The textbase level includes the explicit ideas conveyed by the entire propositions, and the level of referential cohesion among propositions, namely the extent to which nouns, pronouns and noun phrases make reference to other elements of the text (Graesser, McNamara and Kulikowich 2011). Cohesion gaps have been shown to increase reading times and hinder comprehension (ibid.). Coh-Metrix measures referential cohesion by computing overlap in nouns, pronouns, arguments, stems, and content words (McNamara et al. 2014). The tool also measures lexical diversity by means of the type-token ratio, the vocabulary diversity algorithm, and the Measure of Textual Lexical Diversity (ibid.).

The situation or mental model “is the subject matter content or the narrative world that the text is describing” (Graesser, McNamara and Kulikowich 2011, p. 227). The forming of a situation/mental model involves inferences based both on the meaning of the propositions (i.e. the textbase) and on the reader’s prior knowledge (McNamara et al. 2014). As reported in Patel and Kaufman (2006, p. 145): “[i]n medicine, the situation model would enable a physician to draw inferences from a patient’s history leading to a diagnosis, therapeutic plan, or prognosis”. This mental presentation of the text has been defined as the outcome of reading comprehension (McNamara and Magliano 2009, p. 302). Zwaan and Radvansky (1998) identified five dimensions of the situation model, namely causation, intentionality, time, space, and protagonists. For example, in describing the situational model of a biology text on the circulatory system, Graesser, McNamara and Kulikowich (2011, p. 227) specify that it would include:

- a) causal networks of the events, processes, and enabling states that transpire over time, (b) properties of components, (c) the spatial composition of the anatomy, and (d) goal-oriented actions of doctors, patients, and family who try to improve the functioning of someone’s circulatory system.

Coh-Metrix allows for the analysis of the situation model dimensions of causation, intentionality, space, and time, by computing, among others, the number of causal and intentional particles, the frequency of spatial signals, or temporal adverbials and connectives (Graesser, McNamara and Kulikowich 2011; McNamara et al. 2014).

Finally, with regard to genre (or text category), Coh-Metrix is able to differentiate between narrative and informational texts, even though texts often display features of both text categories (Graesser, McNamara and Kulikowich 2011). Zwaan (1994) observed that expectations regarding the genre of a text can influence the way in which a text is processed.

This multilevel text analysis (from words to genre) provided by Coh-Metrix is supported by the theoretical framework of the Construction-Integration model (Kintsch 1998). The Construction-Integration model is regarded as “the most complete and well-formulated model of text comprehension” (McNamara and Magliano 2009, pp. 306-307). Within this model, the term *construction* refers to the process of activation of both relevant and irrelevant knowledge, originating from different sources: the current sentence or proposition (the textbase); the previous sentences or propositions; prior/related knowledge; and previous texts (McNamara and Magliano 2009). At the construction stage, “instead of precise inference rules, sloppy ones are used, resulting in an incoherent, potentially contradictory output” (Kintsch 1988, p. 164). The term *integration* refers to the subsequent stage, when specific concepts receive greater activation than others, as a result of their connections with other concepts in the mental representation of the text (hence the importance of cohesion). At the integration stage, irrelevant or peripheral information is therefore inhibited (and not recalled) (McNamara and Magliano 2009). As reading proceeds and new textual information is integrated, the reader continues to update the mental representation of the text (Beck et al. 1991).

By drawing upon the multilevel theoretical framework of the Construction-Integration model, Coh-Metrix aims to overcome the limitations of traditional readability formulas. Similar to Coh-Metrix, readability formulas have been developed to predict text difficulty (DuBay 2007). By the 1980s, more than 200 readability formulas had been developed, and over 1,000 studies have confirmed their validity (ibid.). Examples of traditional readability formulas are: the Flesch Reading Ease, which counts average sentence length, the number of affixed morphemes, and the number of personal references (Klare 1974); the Flesch–Kincaid Grade Level, which converts the Flesch scores into corresponding grades in the American education system (Doherty

2012); or the SMOG (Section 5.3). It is worth noting that these formulas have been commonly used to assess the readability of print and online cancer information, among others (Friedman and Hoffman-Goetz 2006).

Unlike Coh-Metrix, readability formulas provide a unidimensional metric of text difficulty based on shallow text characteristics (Graesser, McNamara and Kulikowich 2011) — they only calculate (or correlate with) word length³⁹ and sentence length (McNamara et al. 2010), thus excluding other text characteristics that have been shown to be predictors of reading comprehension, such as cohesion (Graesser et al. 2014). Kintsch (2004, p. 1274) argues: “[w]hat makes reading difficult is determined not only by sentence length and familiarity of the words used but also by the number of ideas expressed, their coherence and their structure”.

Several studies have compared the efficacy of Coh-Metrix with that of traditional readability formulas in predicting text difficulty, with slightly varying results. For instance, using the three Coh-Metrix indices of syntactic complexity, co-referentiality and word frequency to assess the difficulty of academic texts, Crossley et al. (2007a) found that their combination could predict 91% of variance in terms of text difficulty, as assessed via cloze tests. The authors (*ibid.*) also observed that readability formulas based on Chall and Dale (1995) and Bormuth (1969) yielded similar results. Dufty et al. (2006) analysed the efficacy of Coh-Metrix automated indices of cohesion and the Flesch-Kincaid Grade Level in estimating the appropriate grade level of textbooks. The authors (*ibid.*) found that, by including cohesion indices provided by Coh-Metrix, the prediction of grade level was significantly enhanced. Crossley, Greenfield and McNamara (2008) tested the hypothesis that differences in lexical frequency, syntactic similarity and content word overlap (as assessed by Coh-Metrix) would allow a more accurate prediction of text difficulty for readers of English as L2 than traditional readability formulas. The authors (*ibid.*) selected those three indices because they correspond to three operations identified in numerous psycholinguistic models of reading comprehension, namely decoding, syntactic parsing and meaning construction.

³⁹ According to Zipf (1949), words which are rarely used in a language tend to be longer. Therefore, word length can be used as a proxy for word familiarity.

Results showed that the three Coh-Metrix indices provided a more accurate prediction of text difficulty compared to traditional readability formulas. The formula resulting from Crossley, Greenfield and McNamara's (2008) study was the Coh-Metrix L2 Reading Index, which was also used for the analysis of texts in our experiment (Section 5.5.2).

Despite their limitations, readability formulas are still widely used in different fields, from education to healthcare (DuBay 2007; McNamara et al. 2010; Benjamin 2012), as emerges from the studies reviewed in Section 5.2 (whose main focus was on the simplification of vocabulary and syntax, rather than on the enhancement of cohesion). However, particularly in healthcare, relying on the unidimensional metric of text difficulty provided by traditional readability formulas might lead to misleading results, as discussed in the following excerpt:

[A] re-validation of the formulas with modern text covering healthcare-relevant topics is needed. For example, many formulas equate long words with difficult words. However, in medicine, this relationship may not hold true, e.g., "apnea" would be considered more difficult than "diabetes" or "obesity" by most readers (Leroy, Kauchak and Mouradi 2013, p. 719).

In summary, based on previous studies showing that, compared with traditional readability formulas, Coh-Metrix indices can improve the prediction of text difficulty by accounting for multiple, theoretically grounded text levels (rather than merely word and syntax), we regarded Coh-Metrix as a more appropriate tool for assessing the readability of our health-related texts (Section 5.4). Finally, our decision to use Coh-Metrix was also supported by the fact that previous studies have already successfully used this tool to identify differences between texts simplified for different groups of readers (from beginner to advanced) (Crossley, Allen and McNamara 2012), as well as differences between authentic texts and texts simplified for L2 readers of English (Crossley and McNamara 2008).

5.5.2 Coh-Metrix Measures

We included the following measures in our readability analysis: narrativity; syntactic simplicity; word concreteness; referential cohesion; deep cohesion; Flesch-Kincaid

Grade Level; and Coh-Metrix L2 Reading Index. A more detailed description of each of these seven variables will be reported in Section 5.6. In this section we describe: how the readability scores were obtained, why we selected these variables, and how the readability scores should be interpreted.

Readability scores for narrativity, syntactic simplicity, word concreteness, referential cohesion, deep cohesion, and Flesch-Kincaid Grade Level were obtained by means of Coh-Metrix Common Core Text Ease and Readability Assessor (T.E.R.A.)⁴⁰, while the Coh-Metrix L2 Reading Index was calculated by Coh-Metrix 3.0⁴¹ (since this variable was not available in T.E.R.A.). Both Coh-Metrix T.E.R.A. and Coh-Metrix 3.0 have been designed to be quick to consult and user-friendly (Dowell, Graesser and Cai 2016).

Narrativity, syntactic simplicity, word concreteness, referential cohesion, and deep cohesion are defined as “easability components” (McNamara et al. 2014, p. 84), and were selected because they have been found to account for 54% of variance in text difficulty (or readability) (ibid.). For each text, Coh-Metrix T.E.R.A. assigns a percentile score on narrativity, syntactic simplicity, word concreteness, referential cohesion, and deep cohesion. Percentile scores are based on a comparison of the text being assessed with thousands of other texts in the Touchstone Applied Science Associates (TASA) corpus (Jackson, Allen and McNamara 2016). The higher the percentile scores on these measures, the higher the readability of the texts. For instance, a percentile score of 70% in syntactic simplicity means that 70% of the texts in the TASA corpus present higher level of syntactic complexity than the text being assessed (McNamara et al. 2014).

Coh-Metrix T.E.R.A. also uses the Flesch-Kincaid Grade Level readability formula, which provides an indication of the reading ability (in terms of US grade-school level) required to be able to read a text (Section 5.5.1). The higher the grade level, the higher the difficulty of the text (Graesser et al. 2004). Drawing upon Graesser et al. (2004), the Flesch-Kincaid Grade Level formula was included in the analysis for comparison with the other multidimensional readability scores provided by Coh-Metrix.

⁴⁰ Coh-Metrix T.E.R.A. is available at: <https://bit.ly/2Aj6Vhk> [Accessed 12 December 2018].

⁴¹ Coh-Metrix 3.0 is available at: <https://bit.ly/2ybzUV6> [Accessed 12 December 2018].

Finally, the Coh-Metrix L2 Reading Index obtained with Coh-Metrix 3.0 provides a prediction of text readability, particularly for L2 readers (McNamara et al. 2014). This variable was included in the analysis since we then tested comprehension of our experimental materials among both native and non-native speakers of English (Chapter 6). The higher the Coh-Metrix L2 Reading Index score, the lower the predicted text difficulty for non-native speakers of English (Crossley, Allen and McNamara 2011).

5.6 Data Analysis and Results

This section presents the analysis and interpretation of the readability scores assigned by Coh-Metrix to: (i) 12 abstracts (i.e. summaries of Cochrane Systematic Reviews for which no simplification approach was adopted); (ii) 12 non-automated PLS (produced by following Cochrane guidelines on simplification); and (iii) 12 semi-automated PLS (checked for readability with Acrolinx and edited accordingly) (Section 5.4). The main objective of this analysis was to determine whether introducing semi-automation into Cochrane's standard simplification approach led to an increase in readability (DV2.1). A secondary goal was to provide empirical evidence on the impact of simplification on readability, as opposed to lack of simplification efforts (as observed in the abstracts).

Drawing upon Crossley, Allen and McNamara (2011), we started by calculating descriptive statistics for the selected measures (Section 5.5.2) for each of the three corpora (Section 5.4). Subsequently, a series of repeated measures analyses of variance (ANOVA) was conducted using the selected measure as the DV, and the three corpora of texts as related groups of the same IV (i.e. non-automated simplification approach, semi-automated simplification approach, or lack of simplification efforts). Repeated measures ANOVA (or within-subjects ANOVA) is an extension of the paired t-test that can be conducted when the IV consists of categorical related groups and the DV is measured at the interval level (Pevalin and Robson 2009). The null hypothesis of this statistical test is that the means of all the related groups of the IV are equal (Kao and Green 2008). Repeated measures ANOVA is adopted to compare three or more group means when the participants are the same in each group, i.e. when there are three or

more observations per each participant (ibid). We could use this test for our experiment because each Cochrane author had produced not only the two PLS, but also the corresponding abstract (Section 5.4).

For the results of a repeated measures ANOVA to be valid, the data need to meet the three following assumptions: (i) absence of outliers (namely, scores that fall beyond the lower and upper quartiles [Leys et al. 2013]) in each of the related groups; (ii) normal distribution of the DV in the related groups; and (iii) sphericity, i.e. the variances of the differences between the treatment levels need to be equal (Park, Cho and Ki 2009). To ensure the validity of the repeated measures ANOVA against these assumptions, we took the following actions:

- when outliers or extreme values were identified (i.e. for narrativity, word concreteness, referential cohesion, deep cohesion, and L2 readability), these scores were removed and a repeated measures ANOVA was conducted again in order to determine if the original results were confirmed. Since a repeated measures ANOVA is based on the assumption that the same participants are present in all the conditions, if a participant had produced an extreme value for a measure in one corpus (e.g. abstracts), his/her scores were removed for all three corpora before re-running the repeated measures ANOVA;
- when the Shapiro-Wilk test of normality showed that the assumption of normality was not met (which was the case for the distribution of the syntactic simplicity scores of the semi-automated PLS, $z=2.137$, $p=0.01$), the Friedman test (i.e. a non-parametric version of the repeated measures ANOVA) was conducted instead. The Friedman test is traditionally used when the DV is measured at the ordinal level. Nonetheless, this test can also be conducted with interval and ratio variables (Ferguson 1976, quoted in Sheldon, Fillyaw and Thompson 1996). The null hypothesis of the Friedman test is that the distribution of the DV is the same in each of the related groups (Pevalin and Robson 2009);
- in order to account for potential violations of the sphericity assumption, the Greenhouse-Geisser (G-G) correction was adopted after each repeated measures ANOVA (Cardinal and Aitken 2013). The G-G correction involves an adjustment of the degree of freedom and is indicated with ϵ (epsilon). More precisely, in addition to the p-

value of the F statistic produced by the repeated measures ANOVA, the p-value of the G-G correction was consulted to check significance at the 0.05 level (Vasey and Thayer 1987; Park, Cho and Ki 2009).

Repeated measures ANOVA is an omnibus test, i.e. it only indicates if at least two means differ significantly (Kao and Green 2008). To find out which means differ significantly from each other, post hoc tests are used (Pevalin and Robson 2009). In the present experiment, we used a Tukey Honestly Significant Difference (HSD) post hoc test when repeated measures ANOVA indicated the presence of at least one statistically significant difference among means. The null hypothesis of the Tukey HSD post hoc test is that there is no difference between a pair of means (Sato 1996).

Similar to the repeated measures ANOVA, a Friedman test can only show if there are statistically significant differences among related groups, but does not show where differences lie. Therefore, after running a Friedman test on syntactic simplicity scores, three separate Wilcoxon signed-rank tests with Bonferroni adjustment were conducted (Lund Research Ltd 2013). The null hypothesis of the Wilcoxon signed-rank test is that the median difference between matched or paired observations is zero (McDonald 2009). The Bonferroni test involves multiple comparisons (one for each combination of related groups) and the adjustment of the significance level by dividing 0.05 by the number of comparisons (Park, Cho and Ki 2009). For instance, if three comparisons are made, the significance level will be set at 0.017 (i.e. at $0.05/3$). The adjustment of the significance level is needed in order to reduce the probability of a Type I error (namely, the incorrect rejection of a true null hypothesis), which would otherwise increase as the number of comparisons increases (Kao and Green 2008).

For each of the measures analysed (Section 5.5.2), a separate table is presented which contains descriptive statistics divided per corpus, each of which contained 12 texts (see Tables 5.2-5.8). For all the measures, the descriptive statistics reported are the means and the SD. The only exception is syntactic simplicity, for which the median rates are reported instead since data in one of the three related groups were not normally distributed — when data show a skewed distribution, the median rates can provide more

meaningful information than the means since they are less influenced by extreme values (Leys et al. 2013).

Measure	Non-automated PLS	Semi-automated PLS	Abstracts
	<i>Mean (SD)</i>	<i>Mean (SD)</i>	<i>Mean (SD)</i>
<i>Narrativity</i>	21.75 (14.15)	20.66 (10.33)	8.25 (2.45)

Table 5.2: Descriptive statistics for narrativity scores divided by corpus

Narrativity is determined by a variety of text characteristics, including pronoun incidence and word frequency (McNamara et al. 2014). Even though narrative texts have been shown to have low referential and verbal cohesion, they are also characterised by the use of frequent words, as well as by high causal and temporal cohesion. Therefore, an increase in narrativity is expected to result in texts that are easier to read (McNamara et al. 2011).

The mean rates for narrativity in Table 5.2 show that: (i) regardless of the simplification approach adopted, PLS scored higher on narrativity than their authentic counterparts (i.e. the abstracts); (ii) compared to non-automated PLS, semi-automated PLS were characterised by a slight decrease in narrativity. A repeated measures ANOVA showed that at least two means differed significantly in narrativity scores, $F(2,22)=12.71$, $p=0.0002$. This result was also confirmed when using the G-G correction, $\epsilon=0.5417$, $p=0.0034$. After removing the outliers, a repeated measures ANOVA and its G-G correction confirmed the presence of at least one statistically significant difference in mean narrativity scores, $F(2,20)=19.76$, $p=0.0000$, $\epsilon=0.5905$, $p=0.0006$.

A Tukey post hoc test showed that the narrativity of the abstracts was significantly lower than the narrativity of both the non-automated PLS ($t=-3.24$, $p=0.008$) and the semi-automated PLS ($t=-2.98$, $p=0.015$). Tukey post hoc results also showed no statistically significant difference between the mean narrativity of non-automated PLS and semi-automated PLS ($t=-0.26$, $p=0.964$). In summary, by applying these statistical tests to narrativity scores, we could observe that: (i) abstracts were less readable than both types of PLS in terms of narrativity aspects such as pronoun

incidence and word frequency; and (ii) semi-automated PLS and non-automated PLS showed a similar level of readability in terms of narrativity.

Measure	Non-automated PLS	Semi-automated PLS	Abstracts
	<i>Median</i>	<i>Median</i>	<i>Median</i>
<i>Syntactic simplicity</i>	61.5	78.5	70

Table 5.3: Descriptive statistics for syntactic simplicity scores divided by corpus

Syntactic simplicity is influenced by the number of words in a sentence and by the complexity of its syntactic structures (McNamara et al. 2011). Texts which score high on syntactic simplicity contain sentences characterised by fewer words and simple syntactic structures. These characteristics are assumed to facilitate the reading process (McNamara et al. 2014).

The median rates for syntactic simplicity in Table 5.3 show that: (i) semi-automated PLS scored higher on syntactic simplicity than both abstracts and non-automated PLS; (ii) unexpectedly, abstracts scored higher on syntactic simplicity than non-automated PLS. A Friedman test showed that there was a statistically significant difference in syntactic simplicity scores, $\chi^2(2)=13.0417$, $p=0.0015$.

For each of the three possible combinations of approaches, a Wilcoxon signed-rank test with Bonferroni adjustment ($p<0.017$ significance level) was carried out. The Wilcoxon signed-rank test comparing the syntactic simplicity scores assigned by Coh-Metrix to the non-automated PLS and the semi-automated PLS showed that the increase in syntactic simplicity observed in the semi-automated PLS corpus was statistically significant ($z=-3.072$, $p=0.0021$). In contrast, no statistically significant differences in syntactic simplicity were identified between the non-automated PLS and the abstracts ($z=-0.903$, $p=0.3666$), or between the semi-automated PLS and the abstracts ($z=1.846$, $p=0.0649$). To sum up, these statistical tests showed us that: (i) semi-automated PLS had a higher level of readability in terms of syntactic simplicity when compared with non-automated PLS; and (ii) abstracts showed a similar level of readability in terms of syntactic simplicity compared with both corpora of PLS.

Measure	Non-automated PLS	Semi-automated PLS	Abstracts
	<i>Mean (SD)</i>	<i>Mean (SD)</i>	<i>Mean (SD)</i>
Word concreteness	25.41 (15.10)	21.91 (14.38)	22.75 (13.83)

Table 5.4: Descriptive statistics for word concreteness scores divided by corpus

Word concreteness is associated with the presence of concrete words that (unlike abstract ones) are expected to facilitate the evoking of mental images and, in turn, the processing of the text (McNamara et al. 2011). The mean rates for word concreteness in Table 5.4 show that: (i) semi-automated PLS scored lower on word concreteness than both abstracts and non-automated PLS; (ii) non-automated PLS were assigned the highest score on word concreteness.

A repeated measures ANOVA showed that there were no statistically significant differences in mean word concreteness scores, $F(2,22)=0.82$, $p=0.4533$. This result was confirmed when using the G-G correction, $\epsilon=0.6055$, $p=0.4039$. After removing the outliers, a repeated measures ANOVA and its G-G correction confirmed the absence of statistically significant differences in mean word concreteness, $F(2,20)=0.77$, $p=0.4757$, $\epsilon=0.6162$, $p=0.4233$. In other words, the three corpora of texts had a similar level of readability in terms of word concreteness. Since no significant differences were identified, the Tukey post hoc test was not conducted.

Measure	Non-automated PLS	Semi-automated PLS	Abstracts
	<i>Mean (SD)</i>	<i>Mean (SD)</i>	<i>Mean (SD)</i>
Referential cohesion	44.08 (15.58)	41.08 (16.92)	20.33 (7.86)

Table 5.5: Descriptive statistics for referential cohesion scores divided by corpus

Referential cohesion is determined by the overlap of content words between adjacent sentences and across all sentences in a text (McNamara et al. 2011). A high level of referential cohesion is expected to facilitate reading comprehension by helping the reader identify the connections within a text (ibid).

The mean rates for referential cohesion in Table 5.5 show that: (i) both non-automated PLS and semi-automated PLS scored higher on referential cohesion than abstracts; (ii) compared to non-automated PLS, semi-automated PLS showed lower

referential cohesion. A repeated measures ANOVA showed that differences in mean referential cohesion scores were statistically significant, $F(2,22)=16.98$, $p=0.0000$. This result was also confirmed when using the G-G correction, $\epsilon=0.7718$, $p=0.0002$. After removing the outliers, a repeated measures ANOVA and its G-G correction confirmed the statistically significant difference in mean referential cohesion scores, $F(2,18)=18.20$, $p=0.0000$, $\epsilon=0.7386$, $p=0.0003$.

A Tukey post hoc test showed that the difference in referential cohesion between the non-automated PLS and the semi-automated PLS was not statistically significant ($t=-0.52$, $p=0.860$). Nonetheless, a statistically significant increase in referential cohesion was observed in both non-automated PLS ($t=-4.14$, $p=0.001$) and semi-automated PLS ($t=-3.62$, $p=0.003$) when comparing these two corpora with the abstracts. In other words, these statistical tests showed that: (i) non-automated PLS and semi-automated PLS had a similar level of readability in terms of referential cohesion (e.g. repetition of content words); and (ii) abstracts had a lower level of readability in terms of referential cohesion when compared with both corpora of PLS.

Measure	Non-automated PLS	Semi-automated PLS	Abstracts
	<i>Mean (SD)</i>	<i>Mean (SD)</i>	<i>Mean (SD)</i>
Deep cohesion	54.83 (24.65)	48.58 (23)	34.33 (16)

Table 5.6: Descriptive statistics for deep cohesion scores divided by corpus

Deep cohesion is determined by the extent to which causal and intentional relationships within the text are signalled by connectives (McNamara et al. 2014). The presence of these connectives is expected to facilitate the comprehension of the text by helping the reader form a coherent mental image of the content (McNamara et al. 2011).

The mean rates for deep cohesion in Table 5.6 show that: (i) both non-automated PLS and semi-automated PLS scored higher on deep cohesion than their authentic counterparts (i.e. the abstracts); (ii) using Acrolinx resulted in a decrease in deep cohesion. A repeated measures ANOVA indicated the presence of a statistically significant difference in mean deep cohesion scores, $F(2,22)=7.78$, $p=0.0028$. This result was also confirmed when using the G-G correction, $\epsilon=0.6052$, $p=0.0119$. After

removing the outliers, a repeated measures ANOVA and its G-G correction confirmed the statistically significant difference in mean deep cohesion scores, $F(2,20)=8.03$, $p=0.0028$, $\epsilon=0.5969$, $p=0.0123$.

Nonetheless, a Tukey post hoc test did not confirm this finding. More precisely, for none of the three possible combinations of corpora (i.e. non-automated PLS and semi-automated PLS, non-automated PLS and abstracts, semi-automated PLS and abstracts) were the differences in mean deep cohesion scores found to be statistically significant ($p>0.05$).

Due to the discrepancy in findings between the repeated measures ANOVA and the Tukey post hoc test, multiple comparisons by means of paired t-tests with Bonferroni adjustment ($p<0.017$ significance level) were also conducted. The null hypothesis of the paired t-test is that the mean difference between matched observations equals zero (McDonald 2009). It was possible to conduct the paired t-tests because the data in the three related groups met the assumption of normality, as measured by the Shapiro-Wilk test ($p>0.05$). The paired t-test comparing the deep cohesion scores assigned by Coh-Metrix to non-automated PLS and semi-automated PLS showed that the increase in deep cohesion observed in the non-automated PLS corpus was not statistically significant ($t=-2.6216$, $p=0.0238$). Similarly, the paired t-test comparing the deep cohesion scores assigned by Coh-Metrix to semi-automated PLS and abstracts showed that the increase in deep cohesion observed in the semi-automated PLS corpus was not statistically significant ($t=2.3445$, $p=0.0389$). On the other hand, the paired t-test comparing the deep cohesion scores assigned by Coh-Metrix to the non-automated PLS and the abstracts showed that the increase in deep cohesion observed in the non-automated PLS corpus was statistically significant ($t=3.1443$, $p=0.0093$). In other words, the non-automated PLS were more readable than the abstracts in terms of deep cohesion (i.e. the presence of connectives in the text). On the other hand, levels of readability in terms of deep cohesion were similar when comparing the two corpora of PLS, and semi-automated PLS and abstracts.

The results of the paired t-tests with Bonferroni corrections seemed to confirm the repeated measures ANOVA result, i.e. that there was a statistically significant

difference between the means of at least two related groups (Abdi and Williams 2010). The discrepancy with the results of the Tukey post hoc test might be due to a Type II error in the Tukey post hoc test, leading to the incorrect acceptance of a false null hypothesis. In relation to this, Sato (1996) reported that the Tukey HSD test decreases the Type I error rate while increasing the Type II error level.

Measure	Non-automated PLS	Semi-automated PLS	Abstracts
	<i>Mean (SD)</i>	<i>Mean (SD)</i>	<i>Mean (SD)</i>
<i>Coh-Metrix L2 Reading Index</i>	12.25 (3.5)	14.01 (3.72)	6.2 (2.17)

Table 5.7: Descriptive statistics for L2 readability scores divided by corpus

The Coh-Metrix L2 Reading Index includes three variables (i.e. lexical coreferentiality, syntactic similarity, and word frequency) which have been shown to predict L2 reading difficulty better than traditional readability formulas (Crossley, Greenfield and McNamara 2008). The Coh-Metrix L2 Reading Index provides a score that considers text challenges at the sentence, word, and cohesion levels (McNamara et al. 2014).

The mean rates in Table 5.7 show that: (i) regardless of the simplification approach adopted, PLS scored higher on L2 readability than the abstracts; (ii) semi-automated PLS were assigned the highest L2 readability scores. A repeated measures ANOVA showed a statistically significant difference in mean L2 readability, $F(2,22)=40.08$, $p=0.0000$. This result was confirmed when using the G-G correction, $\epsilon=0.9903$, $p=0.0000$. After removing the outliers, a repeated measures ANOVA and its G-G correction again confirmed the presence of a statistically significant difference in L2 readability scores, $F(2,20)=38.02$, $p=0.0000$, $\epsilon=0.8781$, $p=0.0000$.

A Tukey post hoc test showed that: (i) when comparing non-automated PLS and semi-automated PLS, the increase in L2 readability observed in the latter was not statistically significant ($t=1.35$, $p=0.38$); (ii) the decrease in L2 readability observed in the abstracts was statistically significant when compared with the L2 readability scores of both non-automated PLS ($t=-4.62$, $p=0.000$) and semi-automated PLS ($t=-5.97$, $p=0.000$). To summarise, results of these statistical tests indicated that: (i) compared with both corpora of PLS, abstracts had a lower level of readability in terms of lexical

coreferentiality, syntactic similarity, and word frequency; and (ii) semi-automated PLS and non-automated PLS had a similar level of readability in terms of lexical coreferentiality, syntactic similarity, and word frequency.

Measure	Non-automated PLS	Semi-automated PLS	Abstracts
	<i>Mean (SD)</i>	<i>Mean (SD)</i>	<i>Mean (SD)</i>
<i>Flesch-Kincaid Grade Level</i>	12 (1.59)	10.58 (0.9)	13.83 (1.11)

Table 5.8: Descriptive statistics for Flesch-Kincaid Grade Level scores divided by corpus

Finally, the Flesch-Kincaid Grade Level is obtained by calculating word length and sentence length (D’Alessandro, Kingsley and Johnson-West 2001; McNamara and Graesser 2012; McNamara et al. 2014). The mean rates for the Flesch-Kincaid Grade Level in Table 5.8 show that: (i) regardless of the simplification approach adopted, PLS were assigned a lower grade level than their authentic counterparts (i.e. the abstracts); (ii) semi-automated PLS were assigned the lowest grade level. A repeated measures ANOVA showed that there were statistically significant differences in mean Flesch-Kincaid Grade Level, $F(2,22)=24.79$, $p=0.0000$, as was also confirmed when using the G-G correction, $\epsilon=0.7702$, $p=0.0000$.

A Tukey post hoc test showed that: (i) when comparing non-automated PLS and semi-automated PLS, the decrease in the Flesch-Kincaid Grade Level observed in the latter was statistically significant ($t=-2.80$, $p=0.022$); (ii) when comparing non-automated PLS and abstracts, the decrease in the Flesch-Kincaid Grade Level observed in the former was statistically significant ($t=3.63$, $p=0.003$); and (iii) when comparing semi-automated PLS and abstracts, the decrease in the Flesch-Kincaid Grade Level observed in the former was statistically significant ($t=6.43$, $p=0.000$). In summary, the statistical tests showed that all the three corpora had different levels of readability as assessed through the Flesch-Kincaid Grade Level, with semi-automated PLS showing the highest level of readability in terms of word length and sentence length.

5.7 Discussion and Summary of the Study on the Readability of PLS and Abstracts

In this chapter, we have described an experimental study that sought to answer RQ2.1 (Section 5.3), namely whether introducing the Acrolinx CL checker into Cochrane’s

non-automated simplification approach (thus rendering it semi-automated) could increase text readability (DV2.1). An additional goal of this experiment was to provide empirical evidence of the impact of simplification on readability, as opposed to lack of simplification efforts. While numerous studies have been conducted on readability (Sections 5.2 and 5.3), to the best of our knowledge, none of them has used Coh-Metrix to investigate the impact of a CL checker on the readability of Cochrane PLS.

In this final section, we will discuss the main findings and we will briefly introduce the next chapter of this thesis. Similar to the procedure described in Yaneva (2015), findings will be explained by referring to: (i) the potential impact of specific Acrolinx rules and Cochrane guidelines on the readability scores provided by Coh-Metrix; and (ii) edits implemented as a result of authors' intuition. However, as pointed out in Nyberg, Mitamura and Huijsen (2003, p. 257), it is often difficult to determine the impact of each individual CL rule.

In the interests of clarity, before discussing the findings, the Coh-Metrix scores for all the measures analysed and for all the three corpora are reported in Table 5.9 below. For each measure, the descriptive statistics reported are the means. The only exception is syntactic simplicity, for which the median rates are reported instead (Section 5.6). The asterisks indicate which differences were found to be statistically significant — different numbers of asterisks are used to indicate where the statistically significant differences lie. For instance, in the case of referential cohesion, a significant difference was found between non-automated PLS and abstracts (signalled with one asterisk), and between semi-automated PLS and abstracts (signalled with double asterisk) (Table 5.9).

Measures	Non-automated PLS	Semi-automated PLS	Abstracts
	<i>Means</i>	<i>Means</i>	<i>Means</i>
<i>Narrativity</i>	21.75 ^(*)	20.66 ^(**)	8.25 ^{(*)(**)}
<i>Word concreteness</i>	25.41	21.91	22.75
<i>Referential cohesion</i>	44.08 ^(*)	41.08 ^(**)	20.33 ^{(*)(**)}
<i>Deep cohesion</i>	54.83 ^(*)	48.58	34.33 ^(*)
<i>Coh-Matrix L2 Reading Index</i>	12.25 ^(*)	14.01 ^(**)	6.2 ^{(*)(**)}
<i>Flesch-Kincaid Grade Level</i>	12 ^{(*)(**)}	10.58 ^{(*)(***)}	13.83 ^{(*)(**)(***)}
	<i>Median</i>	<i>Median</i>	<i>Median</i>
<i>Syntactic simplicity</i>	61.5 ^(*)	78.5 ^(*)	70

Table 5.9: Descriptive and inferential statistics for measures analysed, per corpus

Compared to the non-simplified texts (namely the abstracts), both non-automated PLS and semi-automated PLS scored significantly higher on narrativity. This result is not surprising when considering that both Cochrane guidelines and Acrolinx rules addressed the narrativity of texts (and in particular, word frequency), although not explicitly. For instance, the *PLEACS* included: “Avoid technical terms and jargon or explain them clearly if they are unavoidable. Examples of jargon are clinical terminology [...] as well as terms that may have slightly different meanings in medicine than in common usage” (The Cochrane Collaboration 2013, p. 4); or “[i]f the original title of the review is difficult to understand, for instance if it includes technical terms or jargon, the PLS authors are advised to consider re-writing it in plain language” (Cochrane Norway 2017, p. 2). Similarly, Acrolinx rules dealt with the need to: (i) avoid obsolete words/expressions; and (ii) make words simpler. For instance, after using Acrolinx on their PLS, P01 replaced *vice versa* with *the opposite way*.

Results also showed a decrease in the narrativity of semi-automated PLS when comparing them with non-automated PLS. Even though this decrease was slight and not statistically significant, it is surprising. A possible explanation for the slight decrease in narrativity is that, while increasing some readability scores, the use of Acrolinx might have led to a decrease in others. For instance, the Acrolinx rule on shortening long sentences might have led participants to split a sentence in two and to repeat the same specialised terms in the newly formed sentence, thus increasing the overall number of unfamiliar words in the text. This edit was observed, for example, with P14, who split a sentence into two and repeated the subject, i.e. the word *evidence*.

Regarding syntactic simplicity, semi-automated PLS received a higher score compared with both non-automated PLS and abstracts. The difference between the syntactic simplicity scores assigned to non-automated PLS and to semi-automated PLS was found to be statistically significant. The significantly lower syntactic simplicity of non-automated PLS might be due to the vagueness and scarcity of Cochrane guidelines dealing with sentence length and sentence structure. An example of these guidelines was: “[l]imit sentences to one key point” (The Cochrane Collaboration 2013, p. 4). In contrast, Acrolinx rules tackled a higher number of syntactic issues and were more specific. For example, an exact indication of the maximum number of words allowed in a sentence was provided with Acrolinx. In addition, even when the same syntactic issue (such as the avoidance of passive voice) was specifically tackled by both Cochrane guidelines and Acrolinx rules, participants might have found it easier to apply the recommended modification in the semi-automated approach because the readability issue was automatically flagged in the text. In other words, differently from the non-automated approach, when using Acrolinx, Cochrane authors did not have to rely on their memory of the style-related guidelines.

As far as word concreteness is concerned, the type of simplification approach adopted (or lack thereof) did not result in statistically significant differences for this measure. This result might be explained by the fact that neither Cochrane guidelines nor Acrolinx rules dealt with the concreteness of words. The slight differences in word concreteness emerging from the mean rates (and in particular, the difference between

the abstracts and the non-automated PLS) might be due to the authors' intuitive simplification. In relation to this, Crossley, Allen and McNamara (2012) reported an increase in word concreteness mean rates in texts intuitively simplified for beginner readers, when comparing them with texts intuitively simplified for advanced readers, thus showing that authors tend to regard word concreteness as a text characteristic that could improve readability.

Regarding referential cohesion, non-automated PLS were characterised by a statistically significant increase in this measure, compared with abstracts. Cochrane guidelines did not explicitly address (referential) cohesion. Therefore, as was the case with word concreteness, the increase in this measure might be the result of the simplification strategies intuitively adopted by the authors in order to render a text more readable. This explanation seems supported by the results reported in Crossley, Allen and McNamara (2012), who found that intuitively simplified texts are characterised by a high level of noun overlap (one of the indices of referential cohesion), especially when they are simplified for beginner readers. Similar to non-automated PLS, semi-automated PLS scored significantly higher on referential cohesion when compared to the abstracts. Nonetheless, the corpus of semi-automated PLS was characterised by a slight (not significant) decrease in referential cohesion when compared with non-automated PLS. This decrease in referential cohesion is surprising when considering that one specific Acrolinx rule — asking authors to repeat the noun — was expected to increase referential cohesion by encouraging authors to repeat the same content word instead of using a pronoun. However, when observing the edits made by the participants, it emerged that, when they were presented with this Acrolinx rule, they tended to insert hypernyms or other semantically related words, rather than to repeat the same content word. For instance, P05 added the word *injection* to summarise the type of intervention that had been described in the previous sentences. Similarly, P19 used the word *issues* to describe both changes in treatment and referrals that had been previously mentioned. These additions of new words might have been the cause of the slight decrease in referential cohesion observed in semi-automated PLS.

As far as deep cohesion is concerned, non-automated PLS scored significantly higher on this measure than abstracts, even though Cochrane guidelines did not address the need to explicitly signal causal and intentional relationships in the text. Even in the case of deep cohesion, the intuition of authors might have led them to clarify connections and links within the text in order to make it more readable. In particular, authors might have been led to mainly use causal connectives, whose incidence was found to statistically increase in texts intuitively simplified for beginner readers (Crossley, Allen and McNamara 2012). These findings show that, even though Cochrane authors have a medical background, they are able to intuitively identify some of the characteristics that could make medical texts difficult to read. Therefore, should a CL checker be introduced into Cochrane's PLS production workflow, it would be important to train authors to also rely on their intuition.

As in the case of narrativity, word concreteness and referential cohesion, the adoption of Acrolinx also resulted in a decrease in deep cohesion, although the difference was not statistically significant. This slight decrease might be explained by the changes made by participants when using the CL checker. For instance, in an attempt to reduce the length of a sentence by splitting it (as recommended by Acrolinx), P06 deleted the connective *because* and started a new sentence. Another example is provided by P20, who deleted the intentional connective *to*.

As far as the Coh-Metrix L2 Reading Index is concerned, non-automated PLS scored significantly higher on this measure than abstracts. As reported in Section 5.6, the Coh-Metrix L2 Reading Index provides a score of L2 readability based on three variables: lexical coreferentiality, syntactic similarity and word frequency. Cochrane guidelines did not contain recommendations on lexical coreferentiality. However, they addressed syntactic similarity (e.g. by recommending the use of standardised sentences) and word frequency (e.g. by recommending the avoidance of hard, technical words) (The Cochrane Collaboration 2013; Cochrane Norway 2017). Therefore, the significantly higher L2 readability observed in non-automated PLS (when compared to abstracts) is likely to be the result of the authors' adherence to these guidelines. The adoption of Acrolinx resulted in a further (although not statistically significant) increase

in L2 readability as indicated by the Coh-Metrix L2 Reading Index. This increase is likely to be due to Acrolinx rules further underlining the need for frequent words and syntactic similarity (e.g. by using the active voice).

Regarding the Flesch-Kincaid Grade Level, non-automated PLS scored significantly lower on this traditional readability formula than abstracts, indicating that they could be read by individuals with reading skills associated with a lower US grade level. It should be noted that Cochrane guidelines included the following recommendation for authors: “The SMOG Calculated Index might be useful in implementing the standards for all PLS. This free online tool [...] will calculate sentence length and recommend text to be revised for improved readability” (The Cochrane Collaboration 2013, p. 4). Similar to the Flesch-Kincaid Grade Level, the SMOG is a traditional readability formula which provides a score based on shallow text characteristics (Section 5.3). Therefore, the increase in readability (as measured via the Flesch-Kincaid Grade Level) observed in non-automated PLS is likely to be the result of the authors’ use of the SMOG formula. In addition, after introducing Acrolinx, a further statistically significant increase in readability (as indicated by the Flesch-Kincaid Grade Level) was observed. This result might be due to the Acrolinx rules encouraging participants to decrease both sentence and word length, e.g. by signalling the presence of needless words.

Overall, readability findings showed that, compared with the non-automated approach, the introduction of semi-automation (i.e. the Acrolinx CL checker) resulted in a statistically significant increase in the syntactic simplicity of Cochrane PLS, as well as in a statistically significant decrease in their Flesch-Kincaid Grade Level (i.e. word length and sentence length). In other words, the evidence collected points to the partial acceptance of our alternative hypothesis (Section 5.3) — introducing the Acrolinx CL checker into Cochrane’s non-automated simplification approach for the production of PLS resulted in an increase in readability, but only in terms of syntactic simplicity, sentence length, and word length. As reported in Section 5.3, readability is one of the goals of text simplification conducted at Cochrane and, accordingly, one of the components of authors’ effectiveness or goal completion (ISO 9241-11:2018, 3.1.12).

By reporting these results in terms of usability, we could conclude that the effectiveness of Cochrane authors in terms of the level of readability achieved in the PLS was partially enhanced by the adoption of Acrolinx. This result might be due to: (i) the higher specificity of Acrolinx rules (as opposed to the vagueness which sometimes characterised Cochrane guidelines) (Section 5.7); (ii) the way in which suggestions were presented by Acrolinx (i.e. automatically and consistently flagged in the text) (Section 5.3); or (iii) a combination of both these aspects. These results on how editing PLS with Acrolinx significantly changed some text characteristics of the PLS will also be reported in Sections 6.9 and 7.9, where they will be used to interpret findings on comprehensibility and machine translatability, respectively.

We also observed that, compared to the lack of any simplification attempt (in the abstracts), both before and after the introduction of Acrolinx, simplification significantly increased narrativity, referential cohesion and L2 readability, while significantly decreasing the Flesch-Kincaid Grade Level. In other words, regardless of being non-automated or semi-automated, simplification led to texts scoring higher on a variety of readability measures. These data seem to provide empirical evidence of the benefits of simplification on text readability.

Other interesting findings were the slight decreases in narrativity, word concreteness, referential cohesion, and deep cohesion which were observed after the introduction of Acrolinx. Despite being not significant, these decreases shed light on the fact that text edits aimed to increase readability can have unintended consequences. In line with this observation, Nyberg, Mitamura and Huijsen (2003) remark that CL rules can often contradict each other, and that complying with one rule might lead authors to contravene another. Running multiple checks with a CL checker might help Cochrane authors evaluate the impact of their edits.

It is important to remember that there was high variability among Cochrane authors in terms of time elapsed between the production of the two simplified versions (with and without Acrolinx) (Section 4.6). The difference in time elapsed between the two simplification tasks might have influenced the readability of semi-automated PLS. First of all, authors who had produced their PLS with the non-automated approach only

two or three months before the Acrolinx editing task might have remembered the content more easily than participants who had authored their PLS one or two years before the Acrolinx task. As a result, the former might have felt more comfortable implementing changes suggested by the CL checker as they were less concerned about altering the content without noticing. In contrast, authors who had a less strong memory of their PLS might have felt less comfortable editing their old PLS for fear of inadvertently altering the information contained in it. We tried to reduce the impact of the time elapsed between simplification tasks by not setting any time limit for the Acrolinx task (Appendix F), so that authors could read their PLS and familiarise themselves with its content again if they had forgotten it. Secondly, the same author might have had different PL writing skills when conducting the non-automated simplification task and when using Acrolinx, depending on the amount of PLS produced in the period of time between tasks. However, we were informed that Cochrane authors might be involved in the production of as few as one systematic review (and corresponding PLS) per year (T. Docherty 2017, personal communication, 31 March). This detail is also reported in Glenton (2017, p. 3), who specifies that “[m]any review authors never produce more than one Cochrane Review”. Therefore, it was possible to assume that authors did not have numerous occasions to develop their PL writing skills in-between simplification tasks.

Finally, as discussed in Section 5.2, while the analysis of text characteristics is conducted to predict text readability (versus difficulty), actual comprehension of texts also varies depending on characteristics of the readers. As discussed in Friedman and Hoffman-Goetz (2006, p. 353), “[b]oth readability and comprehension are important concepts in the study of health literacy”. Therefore, to complement the results of this readability experiment, in Chapter 6 we will describe a study aimed to test the comprehension of Cochrane PLS (before and after using the Acrolinx CL checker) and abstracts.

CHAPTER 6

ASSESSING THE COMPREHENSIBILITY OF COCHRANE PLS AND ABSTRACTS

6.1 Aims of the Study on Comprehensibility and Overview of the Chapter

This chapter describes an experimental study which was conducted to determine if introducing the Acrolinx CL checker into Cochrane's non-automated simplification approach for PLS would increase comprehensibility. In line with the readability study presented in Chapter 5, a secondary goal was to provide empirical evidence of the benefits of simplification (regardless of the approach adopted) on reading comprehension. As discussed in Section 1.2, comprehensibility (or DV2.2) is one of the three components of effectiveness, which is the DV2 of the empirical investigation described in this thesis. In Figure 6.1, we highlighted comprehensibility to show how the experiment described in this chapter relates to our broader investigation.

This chapter will begin with an overview of related work. Subsequently, we will present the rationale behind the assessment of the comprehensibility of Cochrane PLS, along with the RQ and the research hypotheses of this experiment. We will then focus on: the recruitment of participants; the experimental environment and procedure; the tasks assigned to the participants; the experimental design; the texts adopted; and the methods used for data collection and data analysis. Finally, we will present the analysis of the collected evidence and discuss the results.

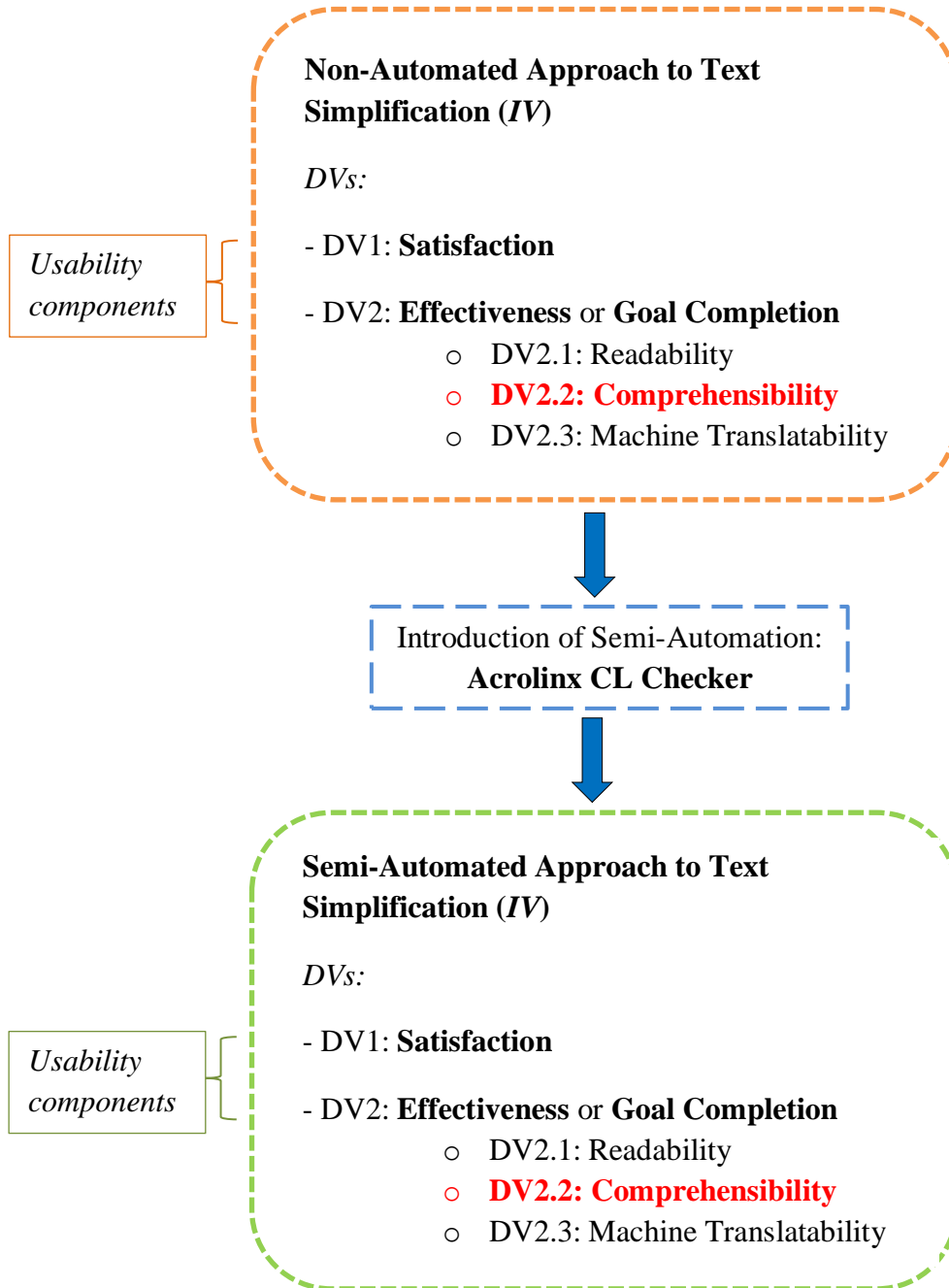


Figure 6.1: Comprehensibility as DV2.2

6.2 Related Work on Reading Comprehension of Health Content

Similar to text readability, comprehensibility (also called *understandability*) is one of the goals of simplification (Smith and Taffler 1992; Yano, Long and Ross 1994), and is defined as the extent to which texts allow readers to “construct a cognitive

representation of incoming information” (Beck et al. 1991). In other words, comprehensibility is determined by the extent to which readers can comprehend texts by integrating the linguistic information presented in a document with their prior knowledge (ibid.) (for a distinction between readability and comprehensibility, see Section 5.2). In line with our focus on health content, in this section, we will mainly present studies on the reading comprehension of medical/science texts, with special attention to the role played by simplification. We will also discuss factors, in addition to readers’ prior knowledge, that can hinder or facilitate comprehension, such as native language, level of English proficiency, reading skills, and level of health literacy.

Comprehension of health content (or lack thereof) might be a factor influencing health outcomes. For instance, Schillinger et al. (2002) point out that patients with low health literacy tend to be less aware of the recommended management of their conditions, and to provide low self-rating of their health. Clarke et al. (2005) found that comprehension of emergency department discharge instructions was correlated with compliance with the instructions, and that native language was moderately correlated with comprehension. In their analysis of how a sample of ethnically and linguistically diverse people living in the United States and taking warfarin for stroke prevention described the therapy and the condition, Fang et al. (2009) reported that LEP and low literacy levels resulted in inaccurate or vague knowledge of both warfarin and strokes. By means of a cloze test and the Short Test of Functional Health Literacy for Adults, Todd and Hoffman-Goetz (2011) analysed the comprehension of colon cancer prevention information among Chinese women in Canada, where immigrant groups have shown lower levels of health literacy and lower awareness of cancer-related health services. The authors (ibid.) observed that women who received the information in Chinese scored significantly higher than those who received the information in English. However, Todd and Hoffman-Goetz (ibid.) also found that comprehension scores were not satisfactory for either group, suggesting that presenting health information in readers’ L1 can reduce, but not eliminate comprehension barriers (Section 8.2).

Due to the inherent difficulty of medical content and the fact that “literacy skills, though necessary, are not sufficient to understand and use health information” (Kandula

and Zeng-Treitler 2008, p. 353), health-related texts are often simplified with a view to enhance their comprehensibility among patients/lay readers (Zarcadoolas 2010). Overall, it has been shown that simplification is beneficial in terms of comprehension (Wilson and Wolf 2009). Meppelink et al. (2015) investigated the impact, on recall, of manually simplifying Dutch texts dealing with colorectal cancer for people with low and high health literacy. Health literacy was assessed by means of the Short Assessment of Adult Literacy in Dutch, and data on recall were collected with open-ended questions, similarly to the method adopted in our experiment (Section 6.7.1). The authors (ibid.) also examined the impact of inserting images. Their results showed that simplification improved recall in both health literacy groups, and that images were particularly beneficial for low health literacy readers when texts were difficult (Section 8.2).

As discussed in Section 5.2, cohesion can influence the level of comprehensibility of a text, particularly for patients/lay people who do not have the health-related prior knowledge required to bridge cohesion gaps through inferences (McNamara et al. 2014). Ozuru, Dempsey and McNamara (2009) set out to investigate how text cohesion (Section 5.2) interacts with readers' prior knowledge and reading skills in the comprehension of science (biology) texts. To this end, the researchers (ibid.) simplified a set of texts by increasing their cohesion. The edits implemented included the use of connectives, or the inclusion of thematic sentences and topic headers. Reading comprehension was assessed through memory-based comprehension questions. Ozuru, Dempsey and McNamara's (2009) findings showed that comprehension of science texts is influenced both by prior knowledge and reading skills, and that (compared to reading skills) prior knowledge is a more significant predictor of comprehension. It was also found that the effect of text cohesion on the resulting comprehension is determined by participants' reading skills.

Liu and Rawl (2012) observed that English texts on colorectal cancer that scored high on referential and semantic cohesion (Section 5.5.1) were read faster and comprehended more than low-cohesion texts on the same topic by English-speaking readers. However, the authors (ibid.) did not observe a positive effect of high text cohesion on recall. It should be noted that, differently from the method adopted in our

experiment (Section 6.7.1), Liu and Rawl (2012) assessed comprehension via yes-no questions, and treated it as different from recall. Testing comprehension of health-related texts among older readers, Liu, Kemper and Bovaird (2009) found that increasing cohesion by repeating key words and ideas is beneficial for comprehension when also using shorter words and sentences.

It should be mentioned that results on the benefits of simplification for comprehensibility of health-related texts have often varied. For instance, Davis et al. (1998) compared the comprehension of two polio vaccine pamphlets, both simplified — one developed at the CDC while the other developed by the researchers (*ibid.*) to be more visually appealing — among parents with various reading skills. Results indicated that comprehension of both simplified texts was poor and possibly not adequate for clinical purposes. In general, achieving the goal of comprehensibility through text simplification has been recognised as a difficult and time-consuming task, particularly for individuals with a health background (like Cochrane authors) who might not be fully aware of the communication needs of the lay audience. In relation to this point, Smith et al. (2011, *n.p.n*) argue that “given the difficulty of engineering comprehensibility of clinical text, the most useful informatics tools will be those that can support the physicians, nurses, and patient educators tasked with making clinical information understandable to patients”.

In summary, increasing the comprehensibility of health content for patients and lay audience is a long-standing goal of text simplification. Overall, changes such as the adoption of short words, the insertion of cohesive devices, and the inclusion of images have been shown to have a beneficial effect. However, comprehension of health-related texts is only partially determined by text characteristics — the success of any simplification effort is likely to vary depending on the reader’s native language, reading skills, level of health literacy, and prior knowledge, among others. Simplifying texts with a view to increase comprehensibility has therefore been recognised as a complex task, particularly when conducted by authors with no linguistics background and without technological assistance.

6.3 Motivation for Assessing the Comprehensibility of Cochrane PLS, Research Question, and Research Hypotheses

In Section 5.3, we discussed the growing interest in the readability of Cochrane PLS. Here we focus on a parallel and complementary area of investigation, namely the comprehensibility of Cochrane PLS and abstracts for various types of readers/users. Due to the extensive length of Cochrane Systematic Reviews, research on comprehension has traditionally focused on the summary formats (e.g. abstracts and PLS), which precede the reviews and are the only content read by numerous readers (Maguire and Clarke 2014). In particular, special attention has been devoted to the level of comprehension achieved by the patients/lay public — making medical information accessible/comprehensible to a lay audience is part of Cochrane’s *Strategy to 2020*, and has been recognised as the first step towards informed health decision making (Section 1.1).

Glenton et al. (2010) identified a lack of consistency across PLS in the final section, where results and quality of the evidence are presented. Motivated by the need to address this issue, they first summarised previous evidence regarding the way in which findings of health research should be presented to patients and lay consumers:

- Patients’ perceptions of risk appear to be more accurate when they are presented with numbers rather than words;
- Patients’ perceptions of risk appear to be more accurate when they are presented with absolute rather than relative risk formats;
- Natural frequencies (e.g., 1 of 1000) are better understood than percentages or probabilities;
- Tables may be preferred and understood by consumers better than narratives; and
- Consistency in the numeric formats used and the avoidance of comparing “apples and oranges” is recommended to increase comprehension. (Glenton et al. 2010, p. 567)

The authors (ibid.) then applied this evidence to the development of a new format of PLS, in which the final section was accompanied by a Summary of Findings table, where key information on results and quality of evidence was presented consistently and

in a numerical format. Lay participants showed a preference for results presented using words and accompanied by numbers in a table. Moreover, a lack of comprehension emerged regarding, for example, the difference between a systematic review and a single study, or the fact that evidence can be of low or high quality. In a follow-up study, Santesso et al. (2015) compared the impact of the old and the new formats of PLS among the lay public in terms of: comprehension of benefits/harms of interventions and quality of the evidence; satisfaction; and preferences. The new PLS format had been further enhanced to include, for instance, headings that would present the text in a question-and-answer format. The authors (ibid.) found that the new format of the PLS was preferred by lay readers, and improved their comprehension and satisfaction. However, it was also observed that only up to 65% of readers answered the majority of comprehension questions correctly, which shed light on the need to further investigate best practices for presenting health content to the lay public. Differently from the text-retelling tasks adopted in our experiment (Section 6.7.1), Santesso et al. (2015) tested comprehension using multiple-choice questions (whose limitations will be discussed in Section 6.7.1). Moreover, differently from the English texts in our experiment (Section 6.6), the texts in Santesso et al. (2015) were translations of the PLS from English into Norwegian, Spanish, and Italian.

Buljan et al. (2018) recruited three groups of readers (university students, healthcare consumers, and health professionals) to measure recall (or knowledge obtained), reading experience, and perceived user-friendliness of three different summary formats of Cochrane Systematic Reviews, namely a PLS written by following Cochrane PLS guidance (Section 4.3), an abstract (Section 5.4), and an infographic. They found that the infographic and the PLS led to similar scores in terms of knowledge obtained, but participants preferred the infographic. The authors (ibid.) argue that, since producing infographics is likely to require higher financial costs than writing PLS, Cochrane should focus on making PLS more appealing. Buljan et al. (ibid.) also observed that abstracts received the lowest scores in terms of reading experience, user-friendliness, and resulting recall — these findings only partially coincide with the results of our study (Section 6.9). Despite the similarities with our experiment described in this

chapter, it should be noted that the texts used by Buljan et al. (2018) were in Croatian (rather than English), and that recall was measured by means of direct questions which required short answers, differently from the text-retelling tasks that we submitted to our participants (Section 6.7.1).

In their study on the impact of four different Cochrane summary formats (i.e. PLS, abstract, podcast, and podcast transcript) on comprehension of key messages of a systematic review, as assessed by means of multiple-choice questions, Maguire and Clarke (2014) reported that: (i) reading the abstract resulted in the lowest comprehension scores; (ii) listening to the podcast led participants to provide the highest percentage of correct answers; and (iii) readers of the PLS and of the podcast transcript provided the correct answers in only half of the questions. These findings led Maguire and Clarke (*ibid.*, p. 447) to ask whether academics are good at writing PLS, and whether systematic reviews should be tested with lay people before being published.

Several studies have also tested comprehension of Cochrane summary formats with readers having a health background (rather than the lay public). This trend is not surprising when considering that

[t]here has been a realization over the last few years that health professionals may not be at much more ease than consumers in understanding and interpreting statistical information, even when presented in a SoF [Summary of Findings] table (Langendam et al. 2013, p. 9).

For instance, Alderdice et al. (2016) examined the interpretation and comprehension of abstracts and PLS among midwifery students, along with the impact of presenting authors' conclusions at the end of the summary. Overall, findings showed poor understanding of results of systematic reviews among midwifery students. When review results were uncertain, providing authors' conclusions in the text appeared to be particularly beneficial, and educational experience seemed to predict a better comprehension outcome.

As these studies have shown, there is a need to enhance the comprehensibility of Cochrane PLS (as well as of other summary formats) for both the lay public — the target audience in our experiment (Section 6.4) — and healthcare practitioners. New

formats and improvements have been proposed for PLS, however, to the best of our knowledge, no previous studies have examined the impact of using a CL checker on the comprehensibility of Cochrane PLS. In Section 5.3, we hypothesised that the Acrolinx CL checker might be beneficial for the readability of PLS since: (i) it provides authors with examples to follow; (ii) it ensures that the same readability issues are consistently flagged across PLS; and (iii) it automatically flags readability issues, thus reducing the impact of authors' differences in terms of memory of simplification rules to be applied.

Unlike readability, comprehensibility is only partially determined by text characteristics (Section 5.2). Here we examined whether, by availing of the benefits that a CL checker entails, Cochrane authors with a health background would produce PLS more easily comprehended by lay readers. Concretely, we tested the hypothesis that the introduction of the Acrolinx CL checker would increase authors' effectiveness (DV2) and, more specifically, the level of comprehensibility that they achieved (DV2.2) as a result of simplification. It should be remembered that *effectiveness* is defined as "the accuracy and completeness with which users achieve certain goals" (ISO 9241-11:2018, 3.1.12). As reported in Section 1.2, the RQ associated with effectiveness (or goal completion) is the following:

RQ2: Does semi-automating a non-automated simplification approach by introducing a CL checker increase authors' effectiveness?

RQ2 was then further segmented into three other questions (one per each of the goals of readability, comprehensibility, and machine translatability) (Figure 6.1). The RQ associated with comprehensibility (RQ2.2) is the following:

RQ2.2: Does semi-automating a non-automated simplification approach by introducing a CL checker increase comprehensibility?

The corresponding research hypotheses are:

H0: Semi-automating a non-automated simplification approach by introducing a CL checker does not increase comprehensibility.

H1: Semi-automating a non-automated simplification approach by introducing a CL checker increases comprehensibility.

In this chapter, we will present the experiment conducted to answer RQ2.2.

6.4 Recruitment of Lay Readers

Participants/lay readers were recruited from the pool of ASU students between September and October 2017, since this study was conducted while the author of this thesis was on secondment at the Science of Learning and Educational Technology Lab at ASU (Phoenix) as part of the INTERACT project (Section 1.3). Prior to conducting our reading comprehension experiment, we sought and obtained ethical approval from the Research Ethics Committee at DCU (DCUREC/2017/066) and from the Institutional Review Board at ASU (both letters of approval are in Appendix I).

Recruitment was conducted online, through the ASU SONA system, which allows researchers to post a brief description of their study, along with the expected time commitment, the requirements for participation (if any), the credits obtained as a reward for participating, and the researcher's name, affiliation, and contact details. Interested and eligible students can then sign up for the study online, or contact the researcher for further information. Since experiments at ASU are traditionally conducted in a laboratory setting with limited capacity to accommodate participants, students also need to select a specific day and time slot among those available when signing up. The SONA system allowed us to adopt a random sampling recruitment technique, whereby every eligible undergraduate and postgraduate ASU student had an equal chance of reading the online post with the description of our study and signing up.

Since we aimed to test reading comprehension of Cochrane content among the lay public — namely individuals with “only common sense or everyday knowledge of a domain” (Patel and Kaufman 2006, p. 152) — the only requirement for participation in our experiment was not being enrolled in any health-related course or training at ASU. There were no requirements on the students' native languages, since we were interested in testing comprehension among both native and non-native speakers of English (Section 6.8.1). As in the case of the authoring study (Section 4.4) and the MT

evaluation study (Section 7.4), it was not possible to calculate the response rate. Overall, 77 native speakers of English and 38 non-native speakers of English took part in our experiment. As will be discussed in Section 6.8.1, data from several of these participants were excluded prior to the analysis. In line with ASU regulations, we assigned two credits (one per each hour of participation in the experiment) to all the students involved.

6.5 Experimental Environment, Procedure and Tasks

Prior to starting the main experiment, we ran a pilot study with six ASU students — one for each of the experimental groups (Section 6.6) — to ensure that there would be no technical issues, and that the online instructions and tasks were clear, which was the case. When the main experiment started, the students who had signed up for the study were asked to go to the laboratory at the agreed time and on the agreed day.

Upon arrival at the laboratory, each student was assigned an ID and a computer, since the tasks had to be conducted online, on the Qualtrics software⁴². The researcher asked the participants to put their phones on silent mode and briefed them about the experimental tasks. The researcher was in the laboratory for the duration of the experiment, but in a different room separated by glass doors. This setting was seen as a compromise between the need to ensure that participants would not be distracted by external factors (e.g. phone calls or Internet surfing) and the need to avoid making the participants feeling observed or under examination. MacKenzie (2013) recommends that the researcher shows a neutral attitude since an overly attentive researcher might make the participants feel nervous, while an indifferent researcher might lead participants to not devote adequate attention to the tasks assigned.

Participants/lay readers were asked to conduct the following tasks:

Task 1. Read the informed consent form (Appendix J);

Task 2. Insert the ID assigned to them and complete a short background questionnaire (Appendix J) containing: five multiple-choice questions, four open-ended questions, and one checkbox question (for non-native speakers of English); and three multiple-choice

⁴² Information on Qualtrics can be found at: <https://bit.ly/1TORi4e> [Accessed 12 December 2018].

questions, two open-ended questions, and one checkbox question (for native speakers of English). The aim of this background questionnaire was to gather evidence on participants' gender, age, ASU college/school being attended at the time of this study, year of college, native language, and types of texts generally read in English. The additional questions asked of the non-native speakers of English dealt with: language spoken at home; self-reported level of English proficiency; and years spent speaking English. Even though on the SONA system we had specified that students who were enrolled in health-related courses could not participate, the question on the ASU college/school that the participants were attending was asked as an extra check (Section 6.8.1);

Task 3. Complete the comprehension component of the online Gates-MacGinitie Reading Test (GMRT)⁴³ (MacGinitie and MacGinitie 1989). The GMRT is a standardised and timed questionnaire used to measure reading skills (Crossley, Yang and McNamara 2014). The test that we used contained 11 short and unrelated passages, each of them followed by 3-6 multiple-choice questions (48 multiple-choice questions in total) dealing with the content and interpretation of the passages. Participants were given 20 minutes to answer as many questions as they could. We submitted the GMRT to our participants because reading skills have been shown to have an impact on reading comprehension (Section 6.2);

Task 4. Read and answer free recall, cued recall, and rating questions on three texts, dealing with three different health topics and belonging to three different corpora (i.e. a non-automated PLS, a semi-automated PLS, and an abstract) (Section 6.6). Specifically, each of the three texts was followed by one free recall question, two cued recall questions, and a question in which participants were asked to indicate how strongly they agreed (or disagreed) with the fact that a text was easy to read (Appendix K).

In the instructions and guidelines preceding each text, participants were informed that the texts that they were about to read were summaries of Cochrane Systematic Reviews, and were provided with a short explanation of what a Cochrane Systematic Review is. This explanation was added since familiarity (or lack thereof)

⁴³ Since the GMRT is a commercial tool, we cannot make it freely available as an appendix.

with a specific genre has been shown to influence readers' expectations of a text and, in turn, their reading comprehension (Zwaan 1994). Glenton et al. (2010) observed a lack of awareness of the difference between a review of previous studies and a single study among readers with no (health) research background. The authors (ibid.) tried to solve this issue by explaining these concepts. However, they noticed that the additional explanation was either ignored or found to increase the difficulty of the content. As will be discussed in Section 6.7.2, similar issues emerged in our study with the description of Cochrane Systematic Reviews.

In the instructions preceding each text, we also specified that participants were not allowed to take notes while reading, and that they could spend as much time as they needed reading the text. Nonetheless, as was reported in Crossley and McNamara (2016), we set a time limit for the text-retelling (recall) tasks and instructed participants not to worry about spelling mistakes. More precisely, we set a four-minute time limit for each free recall question — in which participants were asked to type everything they could remember about the text they had just read — and one and a half minute for each cued recall question — in which participants were given directives, e.g. to type everything they could remember about a specific section of the text just read (McNamara, Ozuru and Floyd 2011). While the instructions and the free recall question were the same for all the texts, across all groups of participants (Section 6.6), the cued recall questions and the rating questions varied depending on the text to be read (Appendix K). Before moving on to the questions, participants were informed that they would not be able to go back to the text to reread it, and that they would not be allowed to consult the Internet or any other resource when answering the comprehension and rating questions (Appendix K);

Task 5. Answer nine multiple-choice questions (three per each of the texts read), aimed to assess participants' topic knowledge, namely their knowledge of the specific topics discussed in the texts (Appendix L). Both domain knowledge and topic knowledge belong to the broader category of prior knowledge (Alexander, Kulikowich and Schulze 1994). However, differently from domain knowledge, topic knowledge may vary even among non-experts of a field — as Alexander, Kulikowich and Schulze (ibid.) point out,

it is possible for individuals to have isolated knowledge about a topic, while at the same time lacking broader knowledge about a domain. In the case of this study, some participants might have undergone the same medical treatments described in the texts (or know someone who had), thus making them familiar with a specific health topic. The topic-knowledge questions were provided to the participants after the reading tasks — so as not to influence their reading behavior — and care was taken to ensure that the texts to be read did not contain answers to the multiple-choice questions despite sharing the same topics (Ozuru, Dempsey and McNamara 2009).

All these tasks were conducted in one session, with no breaks between tasks. For the sake of clarity, Figure 6.2 summarises tasks 2-5 assigned to participants/lay readers, along with the data collected through each task.

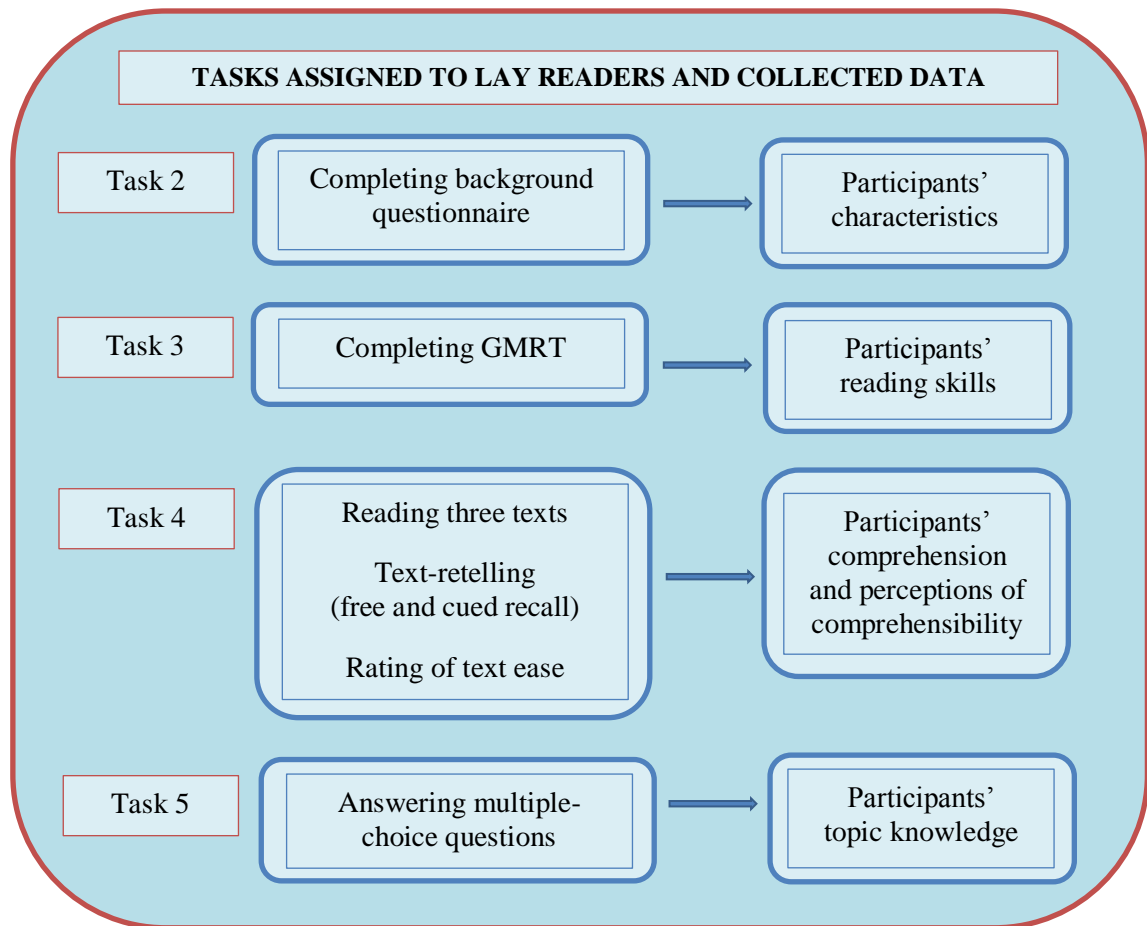


Figure 6.2: Tasks assigned to participants in reading comprehension study and collected evidence

As a final remark on the experimental environment, it should be noted that, in a similar study, Crossley and McNamara (2016) employed a moving windows self-paced reading task, whereby words in a text were shown one by one as the reader pressed a button. This technique does not allow participants to reread words, thus reducing the ecological validity of the reading task — Schotter, Tran and Rayner (2014) specify that rereadings (or regressions) occur for different reasons during the reading process, e.g. when a word is skipped or misinterpreted. The authors (*ibid.*) also showed that the inability to reread words was detrimental to comprehension. Therefore, in our experiment, we decided against the moving windows self-paced reading task, and presented the participants with all the words in the texts simultaneously.

6.6 Experimental Materials and Experimental Design

As specified in Section 6.1, the primary goal of our reading comprehension study was to determine if the introduction of semi-automation (in the form of the Acrolinx CL checker) into Cochrane’s non-automated simplification approach would increase text comprehensibility for lay readers. Buljan et al. (2018) did not include Cochrane abstracts in their study because abstracts are not intended for use by lay audience. However, as in the case of the readability experiment described in Chapter 5, we decided to include Cochrane abstracts as a baseline condition — these non-simplified summaries were used to examine if text simplification (regardless of being non-automated or semi-automated) resulted in an increase in comprehensibility when compared with lack of simplification efforts.

As reported in Section 5.4, the following experimental materials were available to us after the authoring experiment:

- (i) a corpus of 12 non-automated PLS, produced by following Cochrane guidelines/non-automated simplification approach;
- (ii) the corresponding corpus of 12 semi-automated PLS, edited by using Acrolinx;
- (iii) a corpus of 12 abstracts (i.e. non-simplified summaries), extracted from the same Cochrane Systematic Reviews as the PLS and written by the same authors.

Before submitting these texts to lay readers, we decided to run a short accuracy/completeness check with the aim to determine: (i) if there was an exact correspondence between the content in the abstracts and the content in the PLS of the same systematic reviews, namely whether the PLS were complete; and (ii) if, when editing the PLS with Acrolinx, authors had inadvertently altered their content, thus reducing their accuracy. This accuracy/completeness check was run by members of the editorial teams of Cochrane Review Groups. Despite numerous reminders, we could not find evaluators for five pairs of non-automated PLS - semi-automated PLS, which were therefore excluded from the reading comprehension study. Another non-automated PLS and its semi-automated version were also excluded from the reading comprehension study because evaluators did not widely agree that the semi-automated PLS was accurate enough to be disseminated among the general public, possibly as a result of

alterations of content that were an unintended consequence of the use of Acrolinx. An analysis of the impact of Acrolinx on accuracy is beyond the scope of this thesis. However, it should be mentioned that the content alterations identified in this PLS might be the result of the experimental environment in which the authoring study was conducted (Section 4.6) — authors were asked to edit the PLS in one session, and we did not include a follow-up session where authors could check their edited PLS for accuracy. In a real life situation, both Cochrane authors and editors would evaluate the quality of a PLS prior to its dissemination (*Editorial Team 2018*).

Regarding the rest of the texts, for four pairs of non-automated PLS - semi-automated PLS, evaluators identified some pieces of information missing in both corpora of PLS (compared with the abstracts). For instance, one evaluator remarked that not all outcomes reported in the abstract were also reported in the PLS of the same systematic review⁴⁴. However, for these four pairs of PLS, evaluators widely agreed that the texts were accurate enough to be disseminated. We therefore included them as experimental materials in our study. For another pair of non-automated PLS - semi-automated PLS, evaluators did not widely agree that the non-automated PLS was accurate enough to be disseminated. The reason was missing information in the PLS, compared with the abstract. More precisely, one evaluator wrote that the adjective *non-fatal* was missing before *serious adverse effects*. We judged this lack of an adjective as non-detrimental for the well-being of the lay readers in our experiment, and we included this pair of PLS in our study. For the last pair of non-automated PLS - semi-automated PLS, evaluators did not widely agree that the non-automated PLS and the semi-automated PLS were accurate enough for dissemination. The reason was, again, missing information in the PLS. For instance, one evaluator wrote: “Under the heading called Abstract 8 studies are mentioned with follow up explanations. Under the heading called Plain language 8 studies are mentioned and only 6 are mentioned with follow up explanations [sic]”. However, we observed that the pieces of missing information

⁴⁴ Interestingly, some evaluators also pointed out that information that was present in the PLS was sometimes missing in the abstracts. Completeness of content is beyond the scope of this investigation. However, there seems to be a need to check the consistency with which information in Cochrane Systematic Reviews is reported across summary formats (such as abstracts and PLS).

reported by the evaluators were actually present in the PLS. These texts were therefore also used as experimental materials in our study.

In summary, six non-automated PLS, six semi-automated PLS, and six abstracts were tested for comprehensibility.

Text length might have represented a confounding variable influencing the number of phrases/idea units that participants would be able to recall. To make text length as similar as possible across corpora, we used text excerpts. We mainly removed sections from the abstracts, since the non-automated PLS and their semi-automated versions already contained the same rough number of words (Table 5.1). We ensured that the abridged versions were coherent and self-contained, and that the introductory and the concluding sections were always present. After selecting the excerpts, the texts used for the comprehensibility study contained between 290 and 490 words, similarly to what was reported in McNamara, Ozuru and Floyd (2011), where around 150-word variation was allowed among texts.

Similar to the authoring experiment (Section 4.6) and the MT evaluation experiment (Section 7.6), for this reading comprehension study we adopted a within-subject design — each participant was asked to read and answer questions on three texts, namely an abstract, a non-automated PLS, and a semi-automated PLS. The outcome of a reading task is influenced by the different cognitive skills of the subjects involved. A within-subject design, which allows researchers to control for individual differences, was therefore regarded as more appropriate (Lazar, Feng and Hochheiser 2010).

Prior to the experiment, we divided the experimental materials into three categories, based on loosely shared topics: category 1 included texts dealing with non-life-threatening conditions or disorders; category 2 included texts revolving around chronic conditions or disorders; and category 3 included texts focusing on treatments after life-threatening events. Each topic-based category contained two sets of abstract - non-automated PLS - semi-automated PLS. Subsequently, all texts were distributed across six groups (from A to F), with each group containing three texts belonging to different categories/topics and to different corpora.

Table 6.1 presents a summary of this design. *NonAuPLS* indicates that the text to be read was a PLS written following Cochrane guidelines; *AuPLS* indicates that the text to be read was a PLS edited using Acrolinx; and *Ab* indicates that the text to be read was an abstract. To indicate the Cochrane Review Groups to which texts belonged, we used the following codes: Cochrane Eyes and Vision Group (EVG); Cochrane Injuries Group (ING); Cochrane Stroke Group (STG); Cochrane Common Mental Disorders Group (MDG); Cochrane Cystic Fibrosis and Genetic Disorders Group (CFG); and Cochrane Vascular Group (VAG). The number in bracket indicates the category (or topic) of the text.

Groups	Reading task 1	Reading task 2	Reading task 3
A	Ab EVG (1)	AuPLS ING (2)	NonAuPLS STG (3)
B	AuPLS EVG (1)	NonAuPLS ING (2)	Ab STG (3)
C	NonAuPLS EVG (1)	Ab ING (2)	AuPLS STG (3)
D	AuPLS CFG (2)	NonAuPLS VAG (3)	Ab MDG (1)
E	NonAuPLS CFG (2)	Ab VAG (3)	AuPLS MDG (1)
F	Ab CFG (2)	AuPLS VAG (3)	NonAuPLS MDG (1)

Table 6.1: Experimental design of reading comprehension experiment

Both native and non-native speakers of English were randomly assigned to Groups A-F by means of an online randomisation programme⁴⁵. Therefore, each participant read and answered comprehension questions on three texts, each belonging to a different corpus. Participants were blinded to the design (Buljan et al. 2018), i.e. they were not aware of which type of text (whether abstract or PLS) they were reading.

When adopting a within-subject design, it is necessary to take into account and compensate for order effects — such as learning and fatigue effects — which might lead to bias (MacKenzie 2013). Within our study, to compensate for the fatigue effect, we counterbalanced the order in which texts from different corpora were presented. To give an example, the participants who were randomly assigned to Group A read the abstract

⁴⁵ The online randomisation programme is available at: <https://bit.ly/1g696SE> [Accessed 12 December 2018].

first, followed by the semi-automated PLS, and the non-automated PLS. In contrast, the participants who were randomly assigned to Group C read the non-automated PLS first, followed by the abstract, and the semi-automated PLS (Table 6.1). To compensate for learning effect, the three texts assigned to each participant dealt with three different health topics and originated from three different Cochrane Review Groups (Table 6.1). Had participants been assigned with three texts dealing with the same topic, they would have been more likely to recall content after the third reading task, regardless of the corpus to which the third text belonged.

6.7 Methods for Data Collection and Analysis on Text Comprehensibility

This section will start by delving into text-retelling, namely the method adopted to gather evidence on participants' reading comprehension through the information that they recalled (recall protocols). We will first describe text-retelling, and explain the rationale behind its adoption (especially by comparing it to other methods traditionally used for the measurement of reading comprehension). Subsequently, we will report on how the collected data (namely, recall protocols) were analysed (i.e. segmented and scored), with a view to ensuring the reproducibility of this research (MacKenzie 2013).

6.7.1 Characteristics of Text Retelling and Reasons for its Adoption

In text-retelling tasks, the reader is asked to tell or write what they recall from what they have read — along with inferences and extra-textual elaborations — in their own words, and without the possibility of rereading the text (Reed and Vaughn 2012; Crossley and McNamara 2016). In the case of *free recall*, readers are asked to tell or write everything they can remember about an entire text that they have read. In the case of *cued recall*, readers are prompted to tell or write everything they can remember about a specific section in a text (e.g. *write everything you can remember about the goals of this study*) (McNamara, Ozuru and Floyd 2011). The recall protocols produced by the readers are widely used to measure their comprehension of a text (Nilsson 2008). According to the Construction-Integration model developed by Kintsch (1998), an idea in a text is more likely to be recalled when a reader can identify how the idea is linked to other ideas in the text, i.e. when either the text is cohesive, or the reader can bridge gaps in cohesion

and identify links between ideas by relying on their prior knowledge (McNamara et al. 2014) — the identification of these links gives the reader “a retrieval route available to use” (Britton and Gülgöz 1991, p. 334). In turn, identification of the links between ideas is the basis of comprehension since it allows the reader to form a mental representation of the text (McNamara et al. 2014).



Figure 6.3: Exemplification of links between ideas in incoherent (left) and coherent (right) mental representation

Figure 6.3 has been taken from McNamara et al. (ibid., p. 19) to clarify why recall is used as a proxy measure of reading comprehension. Compared with the figure on the right, the figure on the left shows a mental representation of a text that has been comprehended less because there are fewer links between ideas, which are then less likely to be recalled. Similar to the usage of Coh-Metrix (Section 5.5.1), the adoption of text-retelling tasks is therefore supported by the theoretical framework of the Construction-Integration model, in which cohesion plays an important role.

In addition to being a theoretically-grounded method to measure reading comprehension, text-retelling has several benefits. Alderson (2000, p. 230) argues that “[t]his technique is often held to provide a purer measure of comprehension, since test questions do not intervene between the reader and the text”. Secondly, recall allows for equivalency of question format across texts (Reed and Vaughn 2012). In other words, the structure of the prompt (e.g. *write everything you can remember about...*) remains the same even though the text changes (see our free and cued recall questions to participants in Appendix K). This characteristic of the text-retelling task reduces bias, and ensures comparability of results (Choi and Pak 2005). In contrast, other tests of

reading comprehension (such as multiple-choice testing) require tailoring the questions to the specific passage that was read.

Thirdly, in multiple-choice testing, sections of the text are often repeated in the question — especially when readers are non-native speakers of a language — since researchers try to avoid using synonyms which might confuse the participants (Bernhardt 1983). Accordingly, multiple-choice testing acquires the characteristics of a “word recognition and matching exercise” (ibid., p. 28). In relation to this point, Hansen (1978) argues that text-retelling allows for the avoidance of questions that often contain clues. Furthermore, differently from multiple-choice testing, text-retelling does not allow participants to guess at their answers (Crossley and McNamara 2016), and does not expose participants to statements containing wrong answers, which might lead to readers’ long-term beliefs in incorrect statements (Roediger and Marsh 2005). Cloze tests are also widely used instruments of reading comprehension assessment. However, previous research has shown that, unlike reading comprehension tests based on recall, cloze tests mainly tap into readers’ word recognition/decoding skills rather than broader text comprehension (Keenan, Betjemann and Olson 2008). In contrast, text-retelling has been shown to correlate highly with results from open-ended, factual comprehension questions (Hansen 1978), thus providing evidence of the validity of this method.

It should be mentioned that, when testing comprehension in a foreign language — as was the case for a sample of participants in our experiment (Section 6.8.1) — translation tests have also been proposed (Chang 2006). However, Jones (1977, quoted in Bernhardt 1983) sheds light on the difficulty of assessing a translation since there are numerous competing levels of evaluation; e.g. the evaluator might focus on the grammatical characteristics of the text, rather than on its meaning. Moreover, the sample of non-native speakers of English in our experiment differed in terms of their L1 (Section 6.8.1), which made the adoption of a translation test (from English into their different L1) unfeasible.

In summary, we selected a text-retelling task to collect data on readers’ recall (and, indirectly, comprehension) since this task: (i) is theoretically supported by the Construction-Integration model; and (ii) allows for the collection of more comparable,

more valid, and less biased data, compared with other measures of reading comprehension. However, text retelling also entails several limitations.

It has been pointed out, for instance, that the completeness and accuracy of readers' written recall protocols is influenced by their writing fluency, particularly when a time limit is set for the text-retelling task (Peverly et al. 2007), as in the case of our experiment (Section 6.5). Another aspect likely to have an impact on the recall protocols is readers' ability to produce information in an organised manner (Johnston 1981). As discussed in Section 6.6, we adopted a within-subjects design to ensure that individual differences would not bias the data towards one of the three corpora. In other words, potential difficulties or skills of our participants in production/writing did not represent a confounding variable, as they would manifest consistently for each of the three texts that participants were asked to read and recall.

It has also been argued that it is difficult for researchers to determine if incomplete or inaccurate information in the recall protocols is due to comprehension difficulties, production/writing difficulties, or a combination of both (Johnston 1981). This point is summarised in Carlisle (1999, p. 12):

[T]he reader must understand and store the information and be able to retrieve it on demand; the reader must decide on a starting point and a path through the information; the reader must decide on a perspective from which to present the recall.

This limitation of the text-retelling task is particularly visible when readers are asked to produce recall protocols in L2, and when they have a low or intermediate level of proficiency in L2 (Brantmeier 2006). In line with this consideration, Alderson (2000, p. 230) remarked that "the recall needs to be in the first language, otherwise it becomes a test of writing as well as reading". It is likely that, in our experiment with non-native speakers of English, readers' writing skills in L2 influenced the completeness and accuracy of their recalls. However, it is important to remember that the goal of our experiment was not to test if readers had achieved an adequate level of comprehension (Section 8.4), but rather to compare comprehension across corpora (Section 6.1). In other words, instead of measuring accuracy and completeness of recall protocols per se,

we analysed the relative increase or decrease in accuracy and completeness across the three corpora of Cochrane summaries (Section 6.8.2). Furthermore, it is worth noting that, in order to be admitted at ASU, international students whose native language is not English need to provide proof of having at least an intermediate level of English proficiency (*ASU Admission* 2018).

With regard to the implementation of the text-retelling task in our experiment, a final consideration is needed. Differently from Buljan et al. (2018), we used an immediate (rather than a delayed) testing method, whereby readers were assigned the text-retelling task immediately after reading each text. The delayed testing method, which involves inserting neutral content between the reading and the testing tasks (*ibid.*), is adopted to assess longer-term recall of knowledge (Bell et al. 2008). For the purposes of our experiment, an immediate testing method was deemed more appropriate because it was assumed that lay readers do not generally seek health information with a view to learning/acquiring long-term medical knowledge (although desirable), but rather to applying the health information that has been accessed to the specific situation at hand. Finally, by using immediate testing, text-retelling can be regarded more as a test of comprehension and less as a test of readers' memory ability (Alderson 2000).

6.7.2 Procedure for the Analysis of Recall Protocols

Recall protocols produced by readers during text-retelling tasks are often analysed in terms of number, completeness, and accuracy of idea units recalled against the idea units contained in the texts which were read (Reed and Vaughn 2012). Idea units are difficult to define and researchers in the field of reading comprehension have rarely addressed this issue (Alderson 2000; Shin et al. 2016). Accordingly, there is wide variability in the criteria adopted by scholars to segment texts and recall protocols into idea units.

Riley and Lee (1996) argue that an idea unit might correspond to elements as diverse as an idea, a proposition, or a constituent structure. According to Carrell (1983), an idea unit coincides with either a proposition or a phrase. Similarly, Lee and Ballman (1987) state that idea units correspond either to individual sentences, semantic propositions, or phrases. McNamara and Magliano (2009) and Schiefele and Krapp

(1996) link idea units with propositions. In contrast, Bovair and Kieras (1981) make a distinction between idea units and propositions. Brantmeier, Strube and Yu (2014) report that any meaningful piece of information is traditionally treated as an idea unit. Similarly, Horowitz and Newman (1964, p. 642) define an idea as “an utterance that expresses a thought in a meaningful, relevant and unique way”. The authors (*ibid.*) go on to specify that a thought can be expressed even without a subject and a predicate. In her study on the impact of teaching English text structure on L2 reading, Carrell (1985) identifies idea units with single clauses, regardless of whether main or subordinate. Ellis and Barkhuizen (2005, p. 154) define an idea unit as “a message segment consisting of a topic and comment that is separated from contiguous units syntactically and/or intonationally”. In Best, Floyd and McNamara (2008) and McNamara, Ozuru and Floyd (2011), an idea unit is any utterance separated by connectives (e.g. *so* or *because*) and containing a subject, verb and direct object. Richards et al. (2016) categorise as an idea unit each noun or noun phrase, verb or verb phrase, object, and their modifiers. Moreover, other elements such as dates and connectives are treated by the authors (*ibid.*) as separate idea units. For instance, they divided the following sentence into three idea units, as follows: “babies / are born with / a sense of number” (*ibid.*, p. 150).

This review of previous studies has shown that idea units have been identified with a variety of structures, from phrases to sentences. Several studies show a tendency to link idea units with any meaningful piece of information, thus giving prominence to the semantic characteristics of a text. However, criteria for text segmentation based on semantics might be vaguer than syntax-oriented criteria. For this reason, we decided to identify each idea unit with a specific syntactic structure within a text. More precisely, for the purpose of this study, we linked idea units with phrases. The main reason for linking idea units with phrases is the fact that, compared with narrative texts, expository/specialised texts like the ones we used (Section 6.6) are characterised by higher conceptual density (Taylor and Samuels 1983). Therefore, it might have proven difficult for participants — who were not health domain experts (Section 6.8.1) — to recall larger units (such as entire clauses, propositions or sentences) in their entirety. In

other words, phrases were deemed more appropriate for the analysis of the content recalled by our lay readers.

Burton-Roberts (1986, pp. 14-18) provides various descriptions of phrases that help locate them within a sentence:

If a sequence of words can be omitted from a sentence leaving another good sentence, this is a good indication that the sequence is a phrase [...]. However, not all phrases are omissible. [...] [I]f you can replace a SEQUENCE OF WORDS in a sentence with a SINGLE WORD without changing the overall structure of the sentence, then that sequence functions as a constituent of the sentence and is therefore a phrase. [...] [A]nswers to 'WH' questions (that is, questions that contain one of the question words *who*, *which*, *what*, *why*, *where*, *when*, *whose*, and *how*) are phrases. [...] [P]hrases form not only SYNTACTIC UNITS (constituents in the structural form of sentences) but also SEMANTIC UNITS. By this I mean that they form identifiable parts of the MEANING of sentences; they form coherent units of sense. (Emphasis in original)

Similarly, the online *Collins English Dictionary* (2018) defines a phrase as one or more words forming a syntactic unit within a clause. In the online *Cambridge Dictionary* (2018), it is also specified that different word classes can have the function of head of the phrase (e.g. it is possible to have noun phrases, verb phrases, prepositional phrases, etc.).

To measure reading comprehension through the recall protocols produced by lay readers, we followed the same procedure reported in other works. Firstly, both the texts that participants had been asked to read and the recall protocols that participants had produced were segmented into idea units (in this case, phrases) (Bransford and Johnson 1972; Anderson and Pichert 1977; Best, Floyd and McNamara 2008). Subsequently, the phrases in the recall protocols produced by the participants were checked against the phrases contained in the texts in terms of accuracy and completeness. More precisely, each phrase in the recall protocols was assigned a score of 1 if it contained complete and accurate information compared with the corresponding phrase in the text; a score of 0.5 if it contained either incomplete or partially inaccurate information; and a score of 0 if the phrase was fully inaccurate (Schiefele and Krapp 1996; Best, Floyd and McNamara

2008). Correct inferences (produced by two participants only) were assigned a score of 2. Function words (such as conjunctions and pronouns) were always assigned a score of 0 when appearing in a phrase alone since the use of these words was assumed to be necessary for the formation of the sentences, and could not be treated as recalled content as such.

Participants were not penalised for spelling mistakes (Diao and Sweller 2007), nor for using singular instead of plural (or vice versa). Paraphrases and the use of synonyms were allowed, i.e. participants did not have to recall phrases verbatim to obtain a 1 or a 0.5 score (Bransford and Johnson 1972; Bovair and Kieras 1981). In relation to the use of synonyms, we observed that in some cases participants replaced specialised medical terms with their PL correspondents when recalling. This tendency is not surprising when considering that participants in our reading comprehension study did not have a health background (Section 6.8.1). To give just one example, participant Re25 used *fixing* (rather than *treating*) in their recall protocols. Participants were penalised when using words with a different meaning from the words in the texts, e.g. several participants wrote about removing (rather than repairing) an aneurysm. Participants were not penalised for using abbreviations (e.g. for writing *SCI* instead of *spinal cord injuries*). However, when only the initials were used, the phrase was marked as incomplete (0.5 score). For instance, participant R15 reported about the effects “of using B to cure S”.

If participants could not recall the name of the specific treatment or disease/disorder/condition which were the object of the PLS or abstract, and they only referred to them with generic terms (such as *treatment*, *drug*, or *disease*), the phrases containing these terms were marked as *partially accurate* and were assigned a score of 0.5 (rather than 1). However, if a participant mentioned the specific name of a treatment or of a disease/disorder/condition in their recall protocol at least once, and then referred to them with generic terms in the subsequent phrases, their scores were not penalised. Phrases describing the text as a study (rather than as a systematic review) were also treated as partially accurate, and assigned a 0.5 score — this inaccuracy appeared frequently in the recall protocols, suggesting that participants ignored our explanation of

what a systematic review is, as was observed in Glenton et al. (2010) (Section 6.5). Furthermore, participants were not penalised for reporting information in an order different from the order in the text.

Some researchers (e.g. Diao and Sweller 2007) assigned a different weight to different idea units, i.e. they applied a hierarchical structure to idea units in the text by differentiating between main ideas and supporting ideas. However, similarly to Shin et al. (2016), the texts adopted in our study were too conceptually dense for their content to be classified into major and minor chunks of information.

Free recall and cued recall scores were calculated separately (Section 6.8.2). Therefore, it was possible for participants to report the same content when answering free and cued recall questions. However, if a participant reported the same content in the same recall protocol (e.g. at the beginning and at the end), the second occurrence was treated as repetition and excluded from the analysis (Best, Floyd and McNamara 2008). Other types of content in the recall protocols which were excluded from the analysis were: (i) irrelevant content (e.g. when a participant was asked a cued recall question about a section in the text, but reported information from another section instead); (ii) unintelligible content; (iii) metatextual remarks (e.g. “the texts also defined” or “I remember it was discussing”), which could not be treated as recalled content; (iv) elaborations, namely participants’ incorrect inferences or speculations (e.g. “perhaps the effectiveness [...] will be better shown”); (v) extratextual comments (e.g. “do not remember”); (vi) incomplete sentences when their meaning was not clear — these incomplete sentences were usually observed at the end of the recall protocols, and were probably due to the time limit that had been set (Section 6.5); and (vii) content that had been copied and pasted from the text, as in the case of participant Re13.

For each recall protocol, we calculated both a raw score — obtained by summing the completeness/accuracy scores of all the phrases — and its percentage, out of the maximum score that participants might have obtained if they had reported: (i) all the phrases in the text accurately and completely (for free recall questions); and (ii) all the phrases in a specific section accurately and completely (for cued recall questions). For the analysis (Section 6.8.2), percentages were preferred to raw scores because, despite

being similar in length, each text and section contained a different number of phrases/idea units (Carrell 1985; Best, Floyd and McNamara 2008).

Both at the segmentation and at the scoring stage of recall protocols, it would have been desirable to have additional raters, so as to calculate inter-rater agreement. However, due to financial and time constraints⁴⁶, this was not possible. Several measures were therefore adopted to ensure that the impact of the researcher's subjectivity was reduced to a minimum both at the segmentation and at the scoring stage.

To reduce the impact of subjectivity when segmenting texts and recall protocols into phrases/idea units, we followed the criteria for segmentation reported in Burton-Roberts (1986) and Baker (1995). Specifically, we segmented at the level of the minimal phrase, which is defined as the head and the elements that are necessary to complete the meaning of the head, i.e. its complements (Baker 1995). For example, the verb *put* requires a noun phrase as direct object and a prepositional phrase as location complement — as can be seen in *John put the book on the table* — where *put the book on the table* is the minimal verb phrase (ibid.). Unlike complements, modifiers provide additional information which is not required by the head — see e.g. *on Sunday* in *They called their friend on Sunday* (ibid.). In our segmentation, we separated modifiers from the minimal phrase (i.e. from the head and its complements).

In addition to these general segmentation criteria, further and more specific criteria guided the segmentation of texts and recall protocols into phrases/idea units. These criteria are listed below, with the backslash indicating where a phrase finishes and the following starts:

- (i) Subjects/noun phrases were separated from verbs/verb phrases even when the verb was omitted, as in *data \ were sparse \ and \ no overall conclusions \ (were) possible*;
- (ii) Adverbs that only modified verb phrases were treated as part of the verb phrases (see e.g. *is frequently advocated*). However, adverbs that modified entire clauses or sentences were treated as separate phrases (e.g. *Accordingly, \ we \ observed that...*).

⁴⁶ Segmentation and scoring of recall protocols took about four months.

- (iii) In passive voice, the verb and the prepositional phrase indicating the agent (such as *were tested by the doctors*) were treated as components of the same high-level verb phrase (Burton-Roberts 1986, p. 127);
- (iv) Dependent clauses that complemented the verb (e.g. by having the function of direct object) were treated as part of the verb phrase (e.g. *researchers \ demonstrated that additional studies were needed*). However, when more than one clause had the function of direct object, the clauses were separated (e.g. *researchers \ demonstrated that additional studies were needed \ and \ that more participants had to be recruited*);
- (v) In existential constructions (e.g. *there is/are*): *there* was treated as the subject/noun phrase, and therefore separated from the verb (e.g. *there \ was evidence suggesting*) (Baker 1995, p. 427);
- (vi) Restrictive relative clauses were included into the high-level noun phrases in which they appeared (e.g. we treated *the treatments which were discussed in this review* as one noun phrase). In contrast, non-restrictive relative clauses were treated as separate from the noun (e.g. *the participants, \ who were all hospitalised*). Sometimes restrictive relative clauses were non-finite, as exemplified by the noun phrase *low-quality research involving few participants*;
- (vii) When nouns were followed by prepositional phrases that, like restrictive relative clauses, delimited their meanings, nouns and prepositional phrases were treated as one high-level noun phrase (e.g. *patients with sixth nerve palsy*);
- (viii) When WH-words (*when, where, who, what, why, how*) acted as adverbs marking the beginning of questions, or as conjunctions introducing subordinate clauses, they were separated from the rest of the sentence (Burton-Roberts 1986, pp. 185-189) (e.g. *What \ was \ the goal of this study?*). However, when they appeared in verb complements (e.g. *researchers \ do not know how this goal will be achieved*), they were included in the high-level verb phrase;
- (ix) Both coordinating and subordinating conjunctions (e.g. *although* or *because*) were separated from the clauses that they introduced (Burton-Roberts 1986, p. 180) (e.g. *although \ findings \ were not conclusive*). However, when coordinating conjunctions

introduced noun phrases (rather than clauses), they were not separated from them (e.g. *the primary and secondary objectives* was treated as one phrase);

(x) Nouns that were co-ordinate (e.g. by means of *and*, *or*) and had the same role in the sentence (e.g. direct objects) were treated as part of the same noun phrase, which, in those cases, had as many heads as nouns (Burton-Roberts 1986, p. 63).

As an additional check, prior to implementing these segmentation rules to texts and recall protocols, the author of this thesis selected a sample passage, segmented it into phrases, and compared her segmentation with that of another rater (the researcher's supervisor). Disagreements in segmentation were resolved through discussion.

With regard to the scoring of recall protocols, each phrase of the recall protocols was followed by the justification for its score (when 0 or 0.5). For instance, if a phrase was assigned a 0 score because it was inaccurate, this phrase was accompanied by the corresponding phrase in the text — in quotation marks — so that the accurate information was visible. An example is reported here: “[...] is known to be an effective treatment (0 – inaccurate information, ‘it is unclear how effective it is’)”. When the justification for a score required further explanation, this was signalled with either *RN** (researcher's note) or *SR** (extract of systematic review).

It should be mentioned that, since recall protocols were the result of participants' reformulation/paraphrase of texts, there was not an exact correspondence between phrases in the texts and phrases in the recall protocols. For instance, noun phrases in a text could appear as verb phrases in the recall protocols (and vice versa), depending on sentence structure. Moreover, in a few cases, and possibly due to the time limits set for the questions, participants' recall protocols showed changes in sentence planning (see e.g. “to provide the evidence that which treatment works and which doesn't”). In these cases, the segmentation and scoring were conducted based on the most likely interpretation of sentences.

6.8 Data Analysis and Results

This section describes the results obtained by collecting data on: (i) lay readers' eligibility and background characteristics, such as reading skills and topic knowledge;

(ii) lay readers' answers to free and cued recall questions; and (iii) lay readers' ratings of text ease. The main goal of this experiment was to determine if introducing Acrolinx into Cochrane's non-automated simplification approach would increase text comprehensibility (DV2.2). A secondary goal was to provide empirical evidence on the impact of simplification on comprehensibility, as opposed to lack of simplification efforts (as observed in the abstracts).

This section will be further divided into three sections. Section 6.8.1 will present the findings on the screening and background characteristics of lay readers recruited from the pool of ASU students; Section 6.8.2 will report results on participants' text-retelling tasks; and Section 6.8.3 will describe the ratings of text ease that participants assigned to the texts.

6.8.1 Lay Readers' Screening and Background Characteristics

Seventy-seven native speakers of English and 38 non-native speakers of English took part in our experiment. Before analysing the data, a screening was conducted in order to exclude those participants who spent either a limited amount of time reading the texts (possibly indicating that they did not engage in the reading tasks) or spuriously longer times (possibly as a result of lack of attention during the reading tasks) (Miller 1991). This screening was possible because Qualtrics, the software used for this study (Section 6.5), reported the seconds that each participant spent on the pages that contained the three texts to be read.

The screening was conducted per participant, i.e. for each participant, the mean and the SD of the duration of the three reading tasks was calculated. Subsequently, a grand mean and a grand SD were calculated for the entire sample, and the participants whose means were 2 grand SD above or below the grand mean were excluded from the analysis. The same method of participant screening is reported for example in Miller (1991). The author (*ibid.*) reports that researchers can choose to exclude extreme values that fall 2, 2.5 or 3 SD above or below the mean. In our study, selecting a threshold higher than 2 SD would have led to the inclusion of participants who spent as little as 2

seconds reading a text. Therefore, 2 SD was selected as a compromise between the need to exclude outliers while at the same time retaining values which were not extreme.

The screening process was conducted separately for the native and non-native sample because several studies have shown that reading time tends to differ depending on whether a text is written in the readers' L1 or L2. In particular, it has been observed that reading texts in L2 requires more time (Rai, Loschky and Harris 2014; Favreau and Segalowitz 1982). Accordingly, calculating a grand mean and a grand SD that included the duration of the reading tasks of both native and non-native speakers of English might have led to a bias in the screening, i.e. a higher number of non-native participants would have been excluded due to their longer mean reading times.

As a result of the screening, 17 participants were excluded from the sample of native speakers of English — nine of these had a lower mean reading time than the threshold of 2 SD (e.g. 35.34 seconds), while eight participants showed a higher mean reading time than the threshold of 2 SD (e.g. 464.26 seconds). In addition, within the native sample, one participant could not be included in the analysis because the software stopped working during the completion of the study. 15 participants were excluded from the sample of non-native speakers of English — eight of them had a lower mean reading time than the threshold of 2 SD (e.g. 5.77 seconds), while seven participants showed a higher mean reading time than the threshold of 2 SD (e.g. 441.86 seconds). To sum up, 59 native speakers of English and 23 non-native speakers of English were included in the analysis.

Among the 59 native participants, 31 were female, 27 were male and one participant selected the option *I prefer not to disclose* when asked about their gender. Participants reported ages between 17 and 45 years old, with the vast majority being between 18 and 19 years old (n=42). Of the 23 non-native participants, 9 were female and 14 were male. The age range was between 17 and 26 years old, and most participants were between 19 and 18 years old (n=12). Regarding the types of texts generally read in English, both native and non-native speakers of English mainly reported reading emails, followed by essays and notes.

Within the non-native sample, the participants' native languages were Chinese (n=7), Spanish (n=4), Arabic (n=3), Hindi (n=3), French (n=1), Gujarati (n=1), Korean (n=1), Mam (n=1), Telugu (n=1), and Urdu (n=1). Of these participants, only two reported also speaking English at home. When asked to self-report their level of English proficiency by answering the question *How well do you speak English?*⁴⁷ (Central Statistics Office 2016), the majority of participants answered either *very well* (n=10) or *well* (n=10). The remaining three participants replied *Not well*. According to Pandya, McHugh and Batalova (2011, p.12), “any person age 5 and older who reported speaking English less than very well” should be classified as a LEP respondent. Therefore, about half of the non-native speakers of English taking part in our experiment could be considered LEP respondents. It should however be noted that the majority of non-native participants (n=17) also reported having spent seven years or more speaking English, while the rest reported six years (n=3), five years (n=1), four years (n=1), and three years (n=1). This discrepancy with the self-reported level of English might be due to the subjectivity that characterises the self-reporting question (Section 7.8.1).

On the SONA system, we specified that students who were attending health-related courses were not eligible to participate in this study (Section 6.4). This decision was taken to avoid domain knowledge as a confounding variable — compared with low-knowledge readers, high-knowledge readers have been shown to have a richer and more complete recall of textual information (Spilich et al. 1979). Despite our requirement, when analysing the answers to the question on their School or College, it emerged that five participants were attending health-related courses, e.g. in the College of Nursing and Health Innovation at ASU. Even though they did not meet our requirement, these five participants were not excluded from the analysis because, from their answers to the following question on year of study, it emerged that they were all between their first and third year of study in college. In other words, even though five students were attending health-related courses, they were regarded as eligible to participate because they could not be considered health domain experts. In relation to this decision, Boshuizen and

⁴⁷ The same question on self-reported level of English proficiency was asked to MT evaluators (Section 7.8.1).

Schmidt (1992) treat medical students up to the third year as novices in their study on the impact of biomedical knowledge on clinical reasoning. Shapiro (2004) reports that recruiting novices is common practice when researchers aim to remove domain knowledge as a confounding variable.

Regarding topic knowledge, as assessed by means of nine multiple-choice questions⁴⁸ (Section 6.5), descriptive statistics for the number of correct answers were calculated for the entire sample of participants (n=82). We reported these results in Table 6.2. It emerges that, overall, our participants were not very familiar with the topics of the three texts assigned to them.

Descriptive statistics on topic knowledge of lay readers	Scores (out of 9)
<i>Mean (SD)</i>	4.91 (1.55)
<i>Min</i>	1
<i>Max</i>	7

Table 6.2: Descriptive statistics for lay readers' topic knowledge

Results on participants' reading skills, measured through the 48 multiple-choice questions of the GMRT (Section 6.5), are reported in Table 6.3, for the entire sample of lay readers, and then separately for native and non-native speakers of English. It can be observed that, on average, native English speakers had higher reading skills in English compared with non-native speakers of English, as could be expected.

⁴⁸ The limitations of multiple-choice questions that we discussed in Section 6.7.1 in relation to recall (e.g. the possibility that participants guess at their answers) can also be applied to our assessment of topic knowledge. Therefore, the scores in Table 6.2 should be interpreted with caution. Despite its limitations, we used multiple-choice questions for topic knowledge because this testing would give us a quicker overview of this background characteristic.

Participants/lay readers	Scores (out of 48)
	<i>Mean (SD)</i>
<i>Entire sample (n=82)</i>	29.84 (10.94)
<i>Native English speakers (n=59)</i>	33.72 (9.04)
<i>Non-native English speakers (n=23)</i>	19.86 (8.97)

Table 6.3: Descriptive statistics for participants' reading skills

6.8.2 Free and Cued Recall

For each participant, we gathered three free recall protocols (one per each of the texts that were read), and six cued recall protocols (two per each of the texts that were read) (Section 6.5). At the end of each recall protocol, both the raw scores and their conversion into percentages were reported. However, as specified in Section 6.7.2, percentage scores were used for the analysis. It should be noted that, since each participant was assigned two cued recall scores per text, we calculated the mean of their percentage scores before the analysis, and then used the means. For instance, participant Re53 obtained a percentage score of 37.5 on their first cued recall question on the abstract, and a percentage score of 27.77 on their second cued recall question on the abstract. Their average cued recall score for the abstract was therefore 32.63.

Regarding free recall, descriptive statistics for the entire sample of participants/lay readers, and then for native and non-native speakers of English separately, are reported in Table 6.4. It emerges that: (i) free recall of abstracts was consistently lower than free recall of PLS; (ii) differences in free recall between non-automated and semi-automated PLS were consistently slight; (iii) native and non-native speakers of English showed a similar trend of score variation across corpora; and (iv) free recall scores of non-native speakers of English were consistently lower than the free recall scores of native speakers — as specified in Section 6.7.1, this result might be due to reduced comprehension and/or to the difficulty of writing in L2.

Participants/lay readers	Free recall scores		
		<i>Mean (SD)</i>	
	<i>Abstracts</i>	<i>Non-automated PLS</i>	<i>Semi-automated PLS</i>
<i>Entire sample (n=82)</i>	8.01 (5.18)	11.98 (7.31)	11.58 (7.08)
<i>Native English speakers (n=59)</i>	9.08 (5.21)	13.39 (7.59)	12.87 (7.28)
<i>Non-native English speakers (n=23)</i>	5.26 (4.05)	8.36 (5.09)	8.26 (5.38)

Table 6.4: Descriptive statistics for lay readers' free recall scores

In order to determine if differences in free recall scores were statistically significant across the three corpora, we conducted a series of repeated measures (or within-subjects) ANOVA and post hoc tests with free recall as the DV, and the three corpora of texts as related groups of the same IV (i.e. non-automated simplification approach, semi-automated simplification approach, or lack of simplification efforts) (for a detailed description of the repeated measures ANOVA and its assumptions, see Section 5.6). Two different repeated measures ANOVA were run, one with data collected from native speakers of English, and the other with data gathered from the non-native sample. Concretely, the following procedure was adopted:

Checking assumptions of repeated measures ANOVA (free recall scores, native speakers of English)

In the case of the native sample of lay readers (n=59), we identified three outliers, which were excluded from the analysis, thus reducing our sample of native readers to 56. Regarding normality distribution, a Shapiro-Wilk test showed that this assumption was met for the free recall scores of abstracts ($z=1.611$, $p=0.05$) and for the free recall of non-automated PLS ($z=1.215$, $p=0.11$), but not for the free recall scores of semi-automated PLS ($z=1.841$, $p=0.03$). However, since our sample was quite large, this minor violation of the normality assumption did not require a non-parametric test. As reported in Ghasemi and Zahediasl (2012, p. 486), “[w]ith large enough sample sizes (> 30 or 40), the violation of the normality assumption should not cause major problems”.

Finally, we checked the G-G correction to account for potential violations of the sphericity assumption.

Running repeated measures ANOVA and post hoc tests (free recall scores, native speakers of English)

The repeated measures ANOVA on the free recall scores obtained by native speakers of English for the three corpora of texts showed that at least two means differed significantly, $F(2, 110)=10.3$, $p=0.0001$. This result was also confirmed when using the G-G correction, $\epsilon=0.8188$, $p=0.0003$. A Tukey post hoc test showed that the free recall of abstracts was significantly lower than the free recall of both non-automated PLS ($p=0.001$) and semi-automated PLS ($p=0.009$). Tukey post hoc results also showed no statistically significant difference between the free recall of non-automated PLS and semi-automated PLS ($p=0.752$).

Checking assumptions of repeated measures ANOVA (free recall scores, non-native speakers of English)

In the non-native sample of lay readers ($n=23$), no outliers were identified. Moreover, the distribution of the free recall scores for the three corpora met the assumption of normality, as indicated by the Shapiro-Wilk test ($p>0.05$). Mauchly's test did not show any violation of sphericity ($p=0.694$), so there was no need to use the G-G correction.

Running repeated measures ANOVA and post hoc tests (free recall scores, non-native speakers of English)

The repeated measures ANOVA on the free recall scores obtained by non-native speakers of English for the three corpora of texts indicated the presence of a statistically significant difference between at least two corpora, $F(2, 44)=6.576$, $p=0.003$. A Tukey post hoc test showed that the free recall of abstracts was significantly lower than the free recall of both non-automated PLS ($p=0.002$) and semi-automated PLS ($p=0.009$). Tukey post hoc results also showed no statistically significant difference between the free recall of non-automated PLS and semi-automated PLS ($p=0.913$).

Regarding cued recall, descriptive statistics for the entire sample of participants/lay readers, and then for native and non-native speakers of English separately, are reported in Table 6.5. It emerges that: (i) differently from free recall, cued recall of abstracts was consistently higher than cued recall of PLS; (ii) differences in cued recall between non-automated and semi-automated PLS were consistently slight; (iii) native and non-native speakers of English showed a similar trend of score variation across corpora; and (iv) cued recall scores of non-native speakers of English were consistently lower than the cued recall scores of native speakers — again, this result might be due to reduced comprehension and/or to the difficulty of writing in L2 (Section 6.7.1).

Participants/lay readers	Cued recall scores		
		<i>Mean (SD)</i>	
	<i>Abstracts</i>	<i>Non-automated PLS</i>	<i>Semi-automated PLS</i>
<i>Entire sample (n=82)</i>	28.84 (21.73)	12.49 (9.37)	13.46 (11.84)
<i>Native English speakers (n=59)</i>	29.77 (21.4)	14.9 (9.09)	15.72 (12.58)
<i>Non-native English speakers (n=23)</i>	26.45 (22.88)	6.3 (7.09)	7.68 (7.07)

Table 6.5: Descriptive statistics for lay readers' cued recall scores

To determine if differences in cued recall scores were statistically significant across the three corpora, we conducted a series of Friedman tests (i.e. the non-parametric version of the repeated measures ANOVA) and post hoc tests, with cued recall as the DV, and the three corpora of texts as related groups of the same IV. We selected the Friedman test after observing that the assumptions needed for the repeated measures ANOVA were not met (Section 5.6). The same procedure used for free recall scores was followed:

Checking assumptions of repeated measures ANOVA (cued recall scores, native speakers of English)

In the case of the native sample of lay readers (n=59), we identified two outliers, which were excluded from the analysis, thus reducing our sample of native readers to 57. By

conducting the Shapiro-Wilk test, we also observed that cued recall scores for none of the three corpora were normally distributed ($p < 0.05$). Since violations of normality affected all the three levels of the IV, we decided to run the non-parametric Friedman test and three separate Wilcoxon signed-rank tests with Bonferroni adjustment, rather than the repeated measures ANOVA.

Running Friedman test and post hoc tests (cued recall scores, native speakers of English)

The Friedman test showed that there was at least one statistically significant difference in cued recall scores of native speakers of English depending on the corpus of texts being read, $\chi^2(2) = 17.158$, $p = 0.000$. The Wilcoxon signed-rank tests with Bonferroni adjustment ($p < 0.017$ significance level) comparing the cued recall scores of non-automated PLS and semi-automated PLS showed that their difference was not statistically significant ($z = -0.719$, $p = 0.472$). In contrast, the increase in cued recall observed for the abstracts was statistically significant when compared with the cued recall of both non-automated PLS ($z = -4.31$, $p = 0.000$) and semi-automated PLS ($z = -4.604$, $p = 0.000$).

Checking assumptions of repeated measures ANOVA (cued recall scores, non-native speakers of English)

In the non-native sample of lay readers ($n = 23$), one outlier was identified. This participant was excluded from the analysis, and our final sample contained 22 non-native speakers of English. None of the three corpora had cued recall scores with a normal distribution, as shown by the Shapiro-Wilk test ($p < 0.05$). Accordingly, as in the case of native speakers of English, we ran the non-parametric version of the repeated measures ANOVA, namely the Friedman test. Subsequently, three separate Wilcoxon signed-rank tests with Bonferroni adjustment were conducted to identify where statistically significant differences lay.

Running Friedman test and post hoc tests (cued recall scores, non-native speakers of English)

The Friedman test identified at least one statistically significant difference in cued recall scores of non-native speakers of English depending on the corpus of texts being read, $\chi^2(2)=9.976$, $p=0.007$. The Wilcoxon signed-rank tests with Bonferroni adjustment ($p<0.017$ significance level) comparing the cued recall scores of non-automated PLS and semi-automated PLS showed that their difference was not statistically significant ($z=-1.288$, $p=0.198$). In contrast, the increase in cued recall observed for the abstracts was statistically significant when compared with the cued recall of both non-automated PLS ($z=-3.435$, $p=0.001$) and semi-automated PLS ($z=-3.077$, $p=0.002$).

To summarise, and in the interests of clarity, Table 6.6 presents free and cued recall scores obtained by native and non-native speakers of English for each of the three corpora of texts. To increase readability, we reported the means, but not the SD. Statistically significant differences are signalled with an asterisk. Different numbers of asterisks are used to indicate where the significant differences lie. For instance, in the case of free recall of native speakers, a significant difference was found between non-automated PLS and abstracts (signalled with one asterisk), and between semi-automated PLS and abstracts (signalled with double asterisk).

Type of recall	Participants	Abstracts	Non-automated PLS	Semi-automated PLS
Free recall	<i>Native English speakers</i>	9.08 ^{(*)(**)}	13.39 ^(*)	12.87 ^(**)
	<i>Non-native English speakers</i>	5.26 ^{(*)(**)}	8.36 ^(*)	8.26 ^(**)
Cued recall	<i>Native English speakers</i>	29.77 ^{(*)(**)}	14.9 ^(*)	15.72 ^(**)
	<i>Non-native English speakers</i>	26.45 ^{(*)(**)}	6.3 ^(*)	7.68 ^(**)

Table 6.6: Descriptive and inferential statistics for recall scores, per corpus and sample of participants

As emerged from the literature reviewed in Section 6.2, reading skills can interact with text characteristics and influence the comprehension of a text. Therefore, in our study, we treated reading skills as a covariate, namely as “a variable likely to be correlated with the dependent variable” (Huitema 2011, p. 123). To this end, we ran a series of within-subjects analyses of covariance (ANCOVA). This statistical test is used to measure the effect of the IV on the DV, adjusting for the covariate (Streiner 2016). A different within-subjects ANCOVA was run for each of the DV (free vs cued recall) and for each sample of participants (native vs non-native). For each ANCOVA, we also used the partial eta squared (η_p^2) as a measure of effect size, namely to calculate the proportion of variance in the DV that could be attributed to each effect (IV and covariate) (Becker 2000).

Testing assumptions and running within-subjects ANCOVA (free recall scores, native speakers of English)

Prior to running the within-subjects ANCOVA, we tested the assumption of homogeneity of regression slopes. In other words, we tested for homogeneity of the covariate in the prediction of the DV across the three conditions (Crossley, Yang and McNamara 2014). Figure 6.4 shows that this assumption was met, since the three lines representing free recall for the three conditions/corpora all increase as reading skills increase.

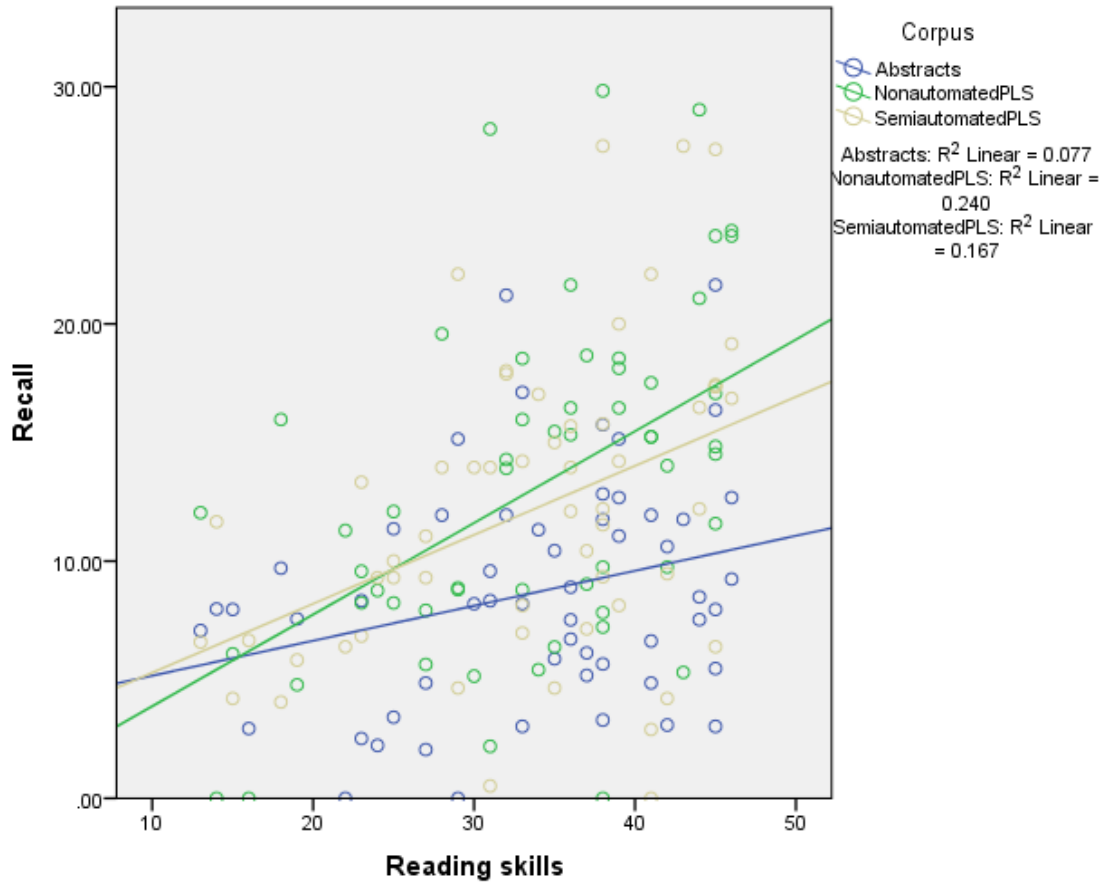


Figure 6.4: Regression slopes for free recall of native readers and reading skills

We report data from the G-G correction since Mauchly's test showed a violation of sphericity ($p < 0.05$). The within-subjects ANCOVA on free recall including reading skills as covariate was not significant, $F(1.584, 85.527) = 0.461$, $p = 0.587$, $\eta_p^2 = 0.008$. We observed a significant effect of reading skills on free recall of native readers, $F(1, 54) = 5.535$, $p = 0.022$, $\eta_p^2 = 0.093$. Therefore, by including reading skills as a covariate, the significant differences that had been observed between free recall of abstracts and free recall of PLS were no longer significant.

Testing assumptions and running within-subjects ANCOVA (free recall scores, non-native speakers of English)

The assumption of homogeneity of regression slopes was met for the free recall scores obtained by non-native speakers of English, as shown in Figure 6.5. The assumption of sphericity was also met, as shown by the result of the Mauchly's test ($p > 0.05$).

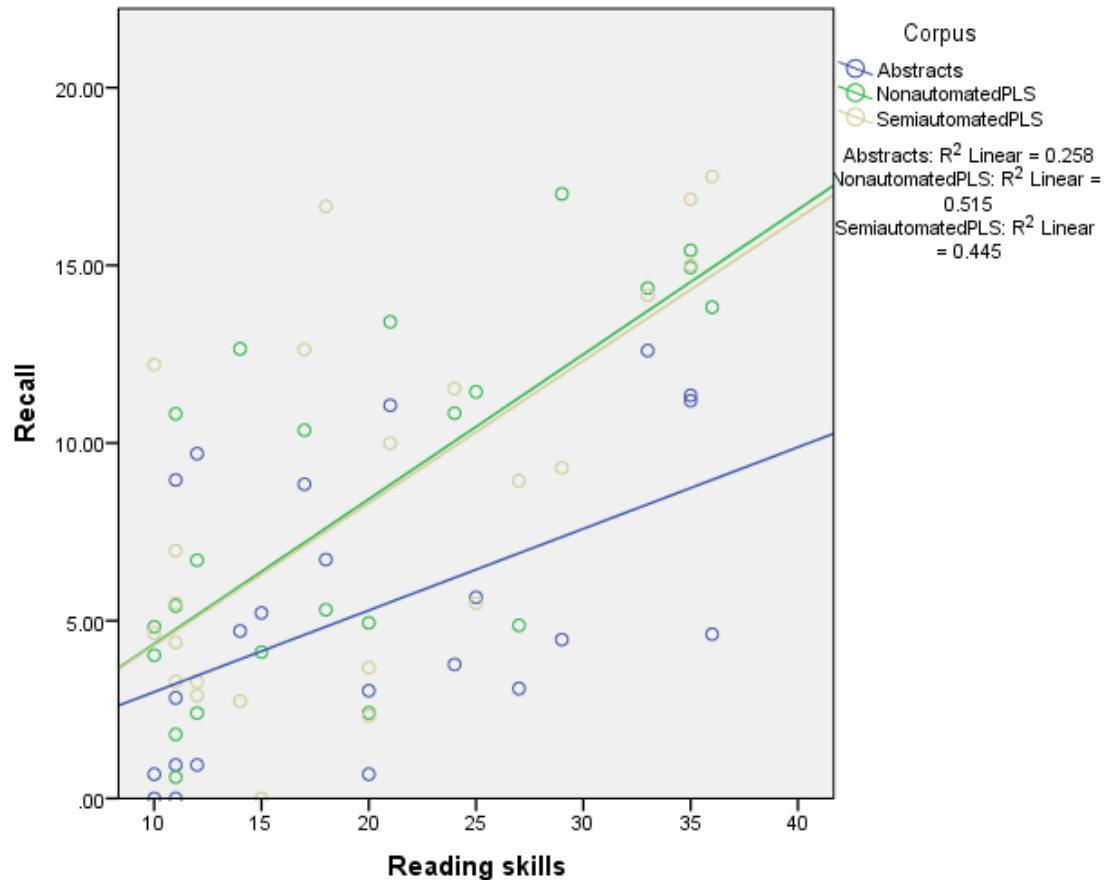


Figure 6.5: Regression slopes for free recall of non-native readers and reading skills

The within-subjects ANCOVA on free recall including reading skills as covariate was not significant, $F(2, 42) = 0.02$, $p = 0.980$, $\eta_p^2 = 0.001$. Moreover, we observed that reading skills did not have a significant effect on free recall of non-native readers, $F(1, 21) = 3.457$, $p = 0.077$. However, reading skills were shown to account for about 14% of the variance in free recall, $\eta_p^2 = 0.141$. As reported in Levine and Hullett (2002, p. 614), when the sample size is small — as in the case of our non-native speakers of English ($n = 23$) — “strong and important effects can be nonsignificant (i.e., a Type II error is made)”.

Testing assumptions and running within-subjects ANCOVA (cued recall scores, native speakers of English)

The distribution of the cued recall scores of native speakers of English was not normal. We tried to achieve normality of distribution by means of a logarithmic transformation of the data (Changyong et al. 2014). However, it was not possible to reduce their skewness. Since ANCOVA is quite robust to violations of normality (Barrett 2011), we decided to run this test despite the skewed distribution of our data. Figure 6.6 shows that the data met the assumption of homogeneity of regression slopes.

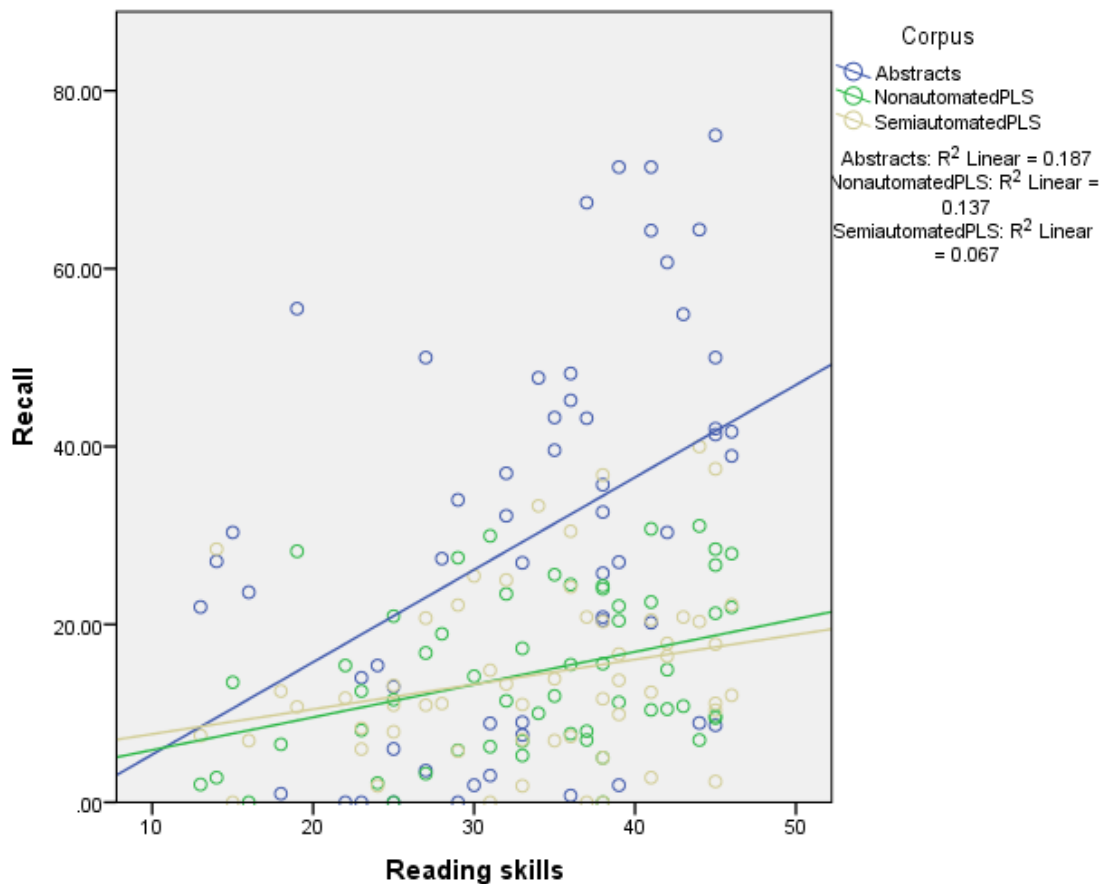


Figure 6.6: Regression slopes for cued recall of native readers and reading skills

Mauchly's test showed that the assumption of sphericity was not met ($p < 0.05$). Therefore, we used the G-G correction. The within-subjects ANCOVA on cued recall including reading skills as covariate was not significant, $F(1.569, 86.315) = 0.6$, $p = 0.512$,

$\eta_p^2=0.011$. We observed a significant effect of reading skills on cued recall of native readers, $F(1, 55)=4.628$, $p=0.036$, $\eta_p^2=0.078$. Therefore, by including reading skills as a covariate, the significant differences that had been observed between cued recall of abstracts and cued recall of PLS were no longer significant.

Testing assumptions and running within-subjects ANCOVA (cued recall scores, non-native speakers of English)

As in the case of the cued recall of native readers, the cued recall of non-native readers did not show a normal distribution. Logarithmic transformations were again unable to reduce the skewness of the data. Despite this violation, we ran the within-subjects ANCOVA (Barrett 2011). Figure 6.7 shows that the assumption of homogeneity of regression slopes was met.

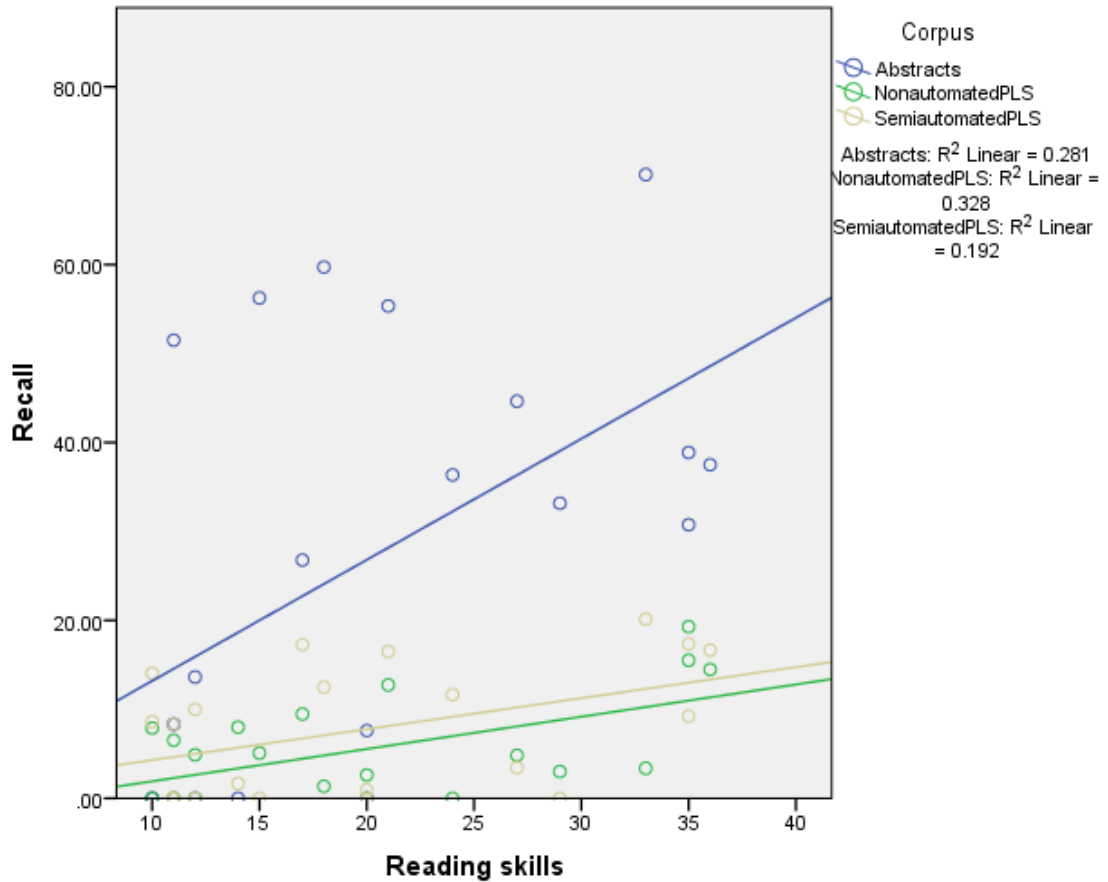


Figure 6.7: Regression slopes for cued recall of non-native readers and reading skills

Based on Mauchly's test, the assumption of sphericity had been violated ($p < 0.05$). Therefore, we used the G-G correction when reporting the results. The within-subjects ANCOVA on cued recall including reading skills as covariate was not significant, $F(1.143, 22.852) = 0.039$, $p = 0.874$, $\eta_p^2 = 0.002$. Moreover, we observed that reading skills did not have a significant effect on cued recall of non-native readers, $F(1, 20) = 3.697$, $p = 0.069$. However, reading skills were shown to account for about 15% of the variance in cued recall, $\eta_p^2 = 0.156$. The observed lack of a significant effect could be caused, again, by a Type II error due to the small sample size (Levine and Hullett 2002).

6.8.3 Ratings

While recall protocols were used to obtain an objective view of lay readers' comprehension of Cochrane PLS and abstracts, ratings allowed us to gather evidence on participants' perceptions of text ease, which might influence their willingness to engage

with a specific text type (Gambrell 2011). As explained in Section 6.5, after reading each of the three texts (namely, abstract, non-automated PLS, and semi-automated PLS) and completing the text-retelling task, participants were asked to indicate (on a 5-point scale) how strongly they agreed or disagreed with the fact that a text was easy to read (Appendix K). Rating results for native speakers of English are reported in Figure 6.8, while Figure 6.9 shows the ratings of non-native speakers of English.

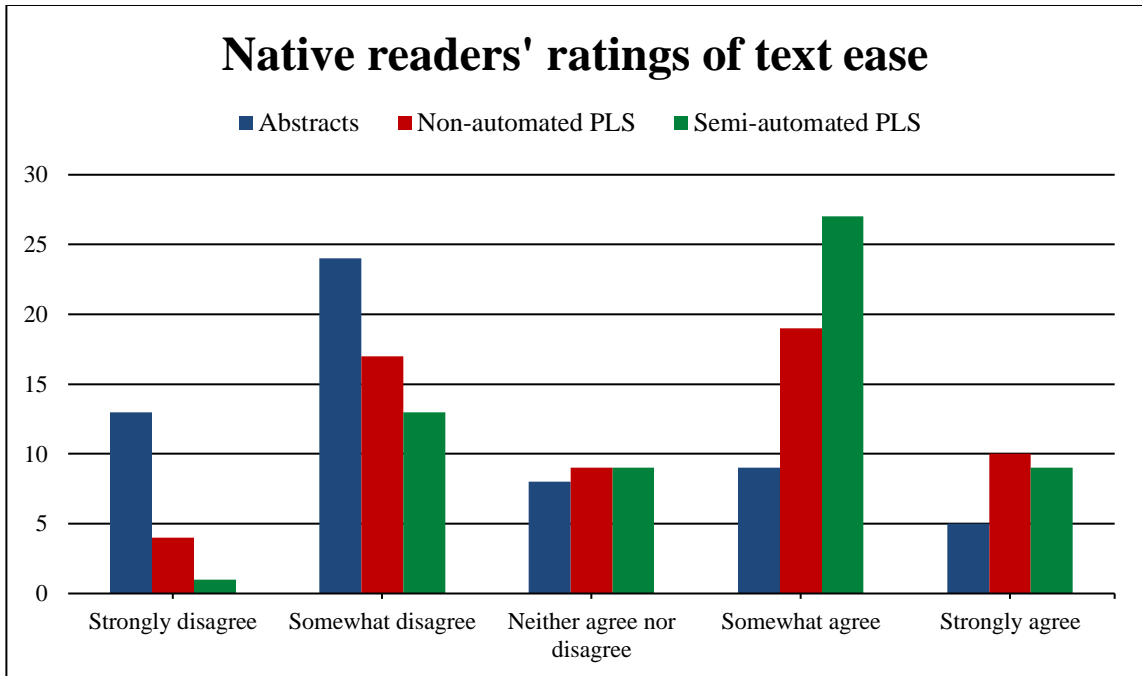


Figure 6.8: Ratings of text ease obtained from native readers

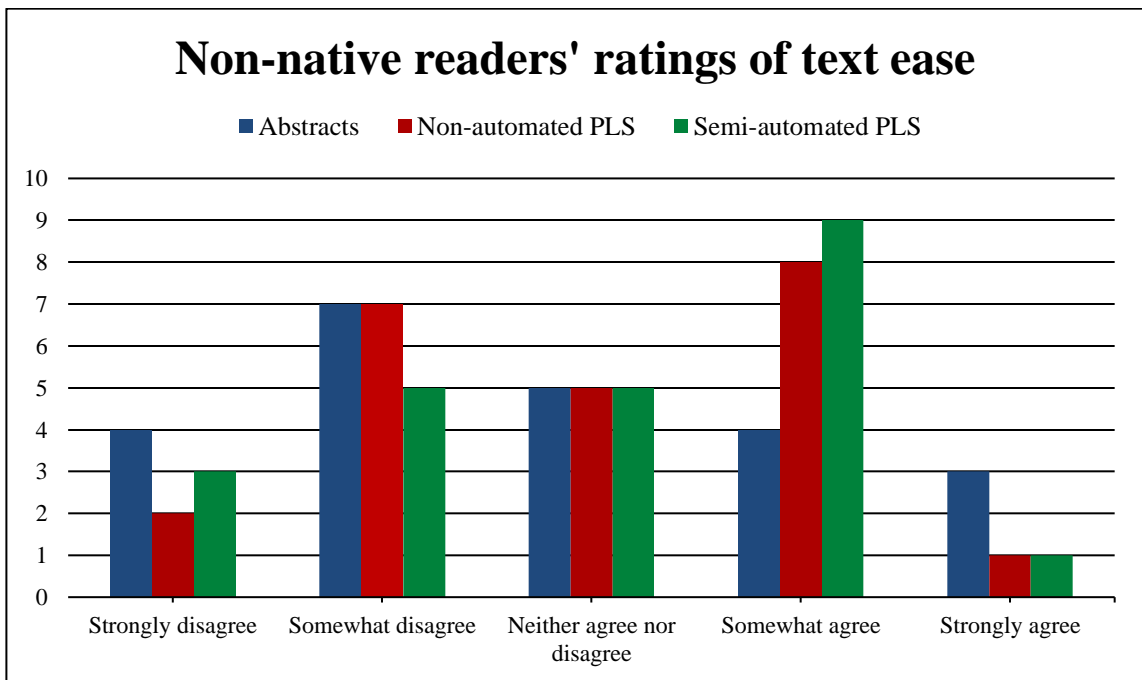


Figure 6.9: Ratings of text ease obtained from non-native readers

For both samples of participants, it can be observed that: (i) the majority (strongly or somewhat) disagreed with the fact that abstracts were easy to read; and (ii) compared with non-automated PLS, a slightly higher number of participants (strongly or

somewhat) agreed with the fact that semi-automated PLS were easy to read. In other words, most participants found abstracts difficult to read, and a slightly higher number of participants found semi-automated PLS easy to read, compared with non-automated PLS.

6.9 Discussion and Summary of the Study on the Comprehensibility of PLS and Abstracts

In this chapter we have described an experimental study aiming to answer RQ2.2 (Section 6.3), namely whether introducing Acrolinx as a CL checker into Cochrane's non-automated simplification approach (thus rendering it semi-automated) would increase the comprehensibility of Cochrane PLS (DV2.2). A secondary goal of our experiment was to gather empirical evidence on the impact of text simplification on comprehensibility. While several researchers have tested the comprehension of Cochrane PLS (Section 6.3), to the best of our knowledge, no previous studies have been conducted on the impact that editing a PLS with a CL checker would have on comprehension. Moreover, it seems that no previous studies with Cochrane PLS and abstracts have collected evidence on the reading skills of participants and on their impact on comprehension. In this final section, we will delve into the main findings, discuss their relevance for Cochrane, and briefly introduce the following chapter of this thesis.

We found that using the Acrolinx CL checker to edit non-automated PLS did not facilitate comprehension among readers who had different language backgrounds, were not health domain experts, and had relatively low knowledge of the health topics discussed. This type of reader often represents the target audience of Cochrane PLS (Langendam et al. 2013). Furthermore, we observed that there was only a slight increase in the number of participants who regarded semi-automated PLS as easy to read, compared with non-automated PLS (i.e. PLS that had not been revised with Acrolinx). Overall, the findings point to the acceptance of our null hypothesis (Section 6.3). In other words, semi-automating Cochrane's non-automated simplification approach by introducing the Acrolinx CL checker did not increase comprehension of PLS, which is

an important part of Cochrane's accessibility mission (Section 1.1). By reporting these results within the framework of usability (and, more precisely, effectiveness or goal completion), we could conclude that the effectiveness of Cochrane authors in terms of the level of comprehensibility achieved in the PLS was not enhanced by the adoption of Acrolinx. Our findings differ from Doherty's (2012), who found that applying CL rules to texts increased recall. However, differently from our investigation, Doherty (*ibid.*) used technical (rather than health-related) texts as experimental materials, and he assessed recall of the MT outputs (rather than the English source texts).

Readability findings reported in Section 5.7 had shown that, compared with non-automated PLS, semi-automated PLS were characterised by a statistically significant increase in syntactic simplicity, and by a statistically significant decrease in word length and sentence length. These characteristics of the semi-automated PLS did not seem to be beneficial in terms of comprehension. One reason could be the fact that, for our reading comprehension experiment, we only used six out of the 12 texts produced by Cochrane authors per corpus (Section 6.6) However, it is worth noting that similar studies also used a limited number of texts (e.g. nine texts in Crossley, Yang and McNamara [2014]).

Another reason for the lack of beneficial impact of Acrolinx on comprehension could be the fact that, unlike edits aimed to increase text cohesion, edits made at the word- and syntax-level do not facilitate the development of connections between the ideas in the text, and are therefore unlikely to result in an increase in recall (Figure 6.3). Tables 5.5 and 5.6 showed that the usage of the Acrolinx CL checker on PLS even resulted in a slight (not significant) decrease in the referential cohesion and deep cohesion. Our results are in line with those in Smith et al. (2011), who observed that recall of clinical documents increased when cohesion was enhanced, but not when assistance to understand the medical vocabulary was provided to readers. To increase the level of cohesion in simplified texts and, in turn, to facilitate their recall/comprehension, CL checkers might include a higher number of cohesion-oriented rules. For instance, to increase noun overlap (one of the components of cohesion), rules aiming at the avoidance of synonymy could be included (O'Brien 2003; Graesser et al.

2004). Another example of a CL rule aimed at increasing cohesion (for Spanish) is reported in Cascales (2002, p. 55): “Avoid the use of referring expressions such as pronouns and deictic determiners, instead repeat the concept”. For those aspects of cohesion that could not be formalised into a CL rule, authors could be trained to rely on their intuition of what makes a text cohesive, particularly at the macro/global level of a text. For example, authors would need to make sure that the presentation of arguments follows a logical order, and that each paragraph begins with a topic sentence that is consistent with the overall topic of a text (Kools et al. 2004; Smith et al. 2011).

Comprehension of abstracts (i.e. non-simplified summaries), as assessed through *free* recall, was significantly lower than comprehension of both corpora of non-automated and semi-automated PLS, among both native and non-native speakers of English. This result seems to confirm the beneficial effect of simplification on reading comprehension that was observed in previous works (Kurtzman and Greene 2016). The readability findings reported in Section 5.9 had shown that, compared with both automated and semi-automated PLS, abstracts were characterised by significantly lower levels of narrativity, referential cohesion, lexical coreferentiality, syntactic similarity, and word frequency, among others — these text characteristics are likely to have hindered the comprehension of abstracts. Moreover, the majority of our native and non-native lay readers disagreed that abstracts were easy to read.

However, we also found that comprehension of abstracts (measured through *cued* recall) was significantly higher than comprehension of both corpora of PLS, among both native and non-native speakers of English. This finding is surprising, especially considering that other studies analysing free and cued recall found that these reading comprehension measures were highly correlated (McNamara, Ozuru and Floyd 2011). To provide an explanation for cued recall results, we examined abstracts and PLS for characteristics other than language. We observed that: (i) sections in the abstracts usually contained less content than sections in the PLS; and (ii) while abstracts always had headings in bold formatting to separate the text into coherent sections, some PLS either were not divided into sections, or had section headings that did not coherently match the content of the sections. For example, it was not unusual for PLS to contain a

section titled *Background* where unrelated types of information were reported (e.g. the description of the medical condition under study and the objectives of the systematic review). Furthermore, two of the PLS that we used for our reading comprehension experiment did not have any headings to separate the sections. It is important to remember that cued recall questions asked participants to write everything they could remember about a specific section of a text (Section 6.5). Therefore, it can be assumed that these characteristics observed in the abstracts (namely, presence of bold headings coherently segmenting content into sections, and shorter sections) facilitated the recall of information when participants were prompted to answer specific questions on sections.

This claim seems supported by previous studies. Regarding the impact of using headings to distinguish different sections in a text, Kools et al. (2004, p. 723) argue that

[h]eadings and subheadings clarify overall text structure and can serve as “anchors” for the reader [...]. In general, it may be assumed that headings influence cognitive processing by (a) acting as cues for activating related prior knowledge, (b) accentuating the relationship among important concepts in a text, and (c) providing retrieval cues *for subsequent recall* of a text. (Emphasis added)

Moreover, Lonsdale (2014) reviewed studies showing that formatting features like the use of bold (as observed in the headings) can emphasise specific pieces of information. A similar remark is made in Rusko, Van der Waarde and Heiniö (2012), where it is stated that comprehensibility of information on pharmaceutical packages is influenced by typographic variables, including colours, layout, and use of headings. Regarding the reduced amount of content in each section, in their study on the effective presentation of information on the quality of healthcare providers, Kurtzman and Greene (2016) found that reducing the amount of information can avoid readers’ overload and facilitate comprehension. Interestingly, in the HSE (2017) guidance on how to communicate in PL, all these aspects are mentioned: from keeping paragraphs short to using headings and bold to highlight content.

In terms of implications for Cochrane and other (non-profit) organisations which produce simplified health content for lay readers, these results showed that, while

beneficial, text simplification might not be sufficient, especially if specific pieces of information within a text (e.g. prevention of a disease, or effectiveness of a treatment) need to be comprehended by the target audience. Formatting and text segmentation might all contribute to enhance the effectiveness of health communication.

Another qualification to the findings on the impact of simplification arose from the ANCOVA. In Section 6.2, we reviewed studies on the impact of reading skills on comprehension. In the case of our experiment, for both native and non-native readers, it emerged that reading skills tend to drive comprehension, often more than the type of text being read. In other words, the variance associated with comprehension could be better explained by individual differences in reading skills. Therefore, regardless of text characteristics (e.g. in terms of simple language, formatting, content segmentation), individuals with low reading skills are likely to show poorer comprehension than individuals with high reading skills. Although these results would need to be confirmed with a larger sample of participants — especially non-native speakers of English — our data indicates that it might be beneficial for lay users of the Cochrane website (as well as lay users of other websites disseminating health content) to be presented with information in multiple formats. This option would allow organisations to meet the different needs of users with various reading skills. In relation to this point, it is noteworthy that Cochrane is currently producing podcasts (Maguire and Clarke 2014) and comic strips based on some of their systematic reviews.

Finally, as emerged from Table 6.6, comprehension scores of non-native speakers of English were consistently lower than the comprehension scores of native English speakers — even though this result might be partly due to the difficulty of writing recall protocols in L2, it also underlines the need to use translation to present lay readers with health content in their L1, as will be discussed in Section 7.2. As argued in Huijsen (1998), CLs are used to improve text comprehension both by humans and by computer applications such as MT systems. Therefore, in the next chapter, we will present an experiment aimed to test the impact of the Acrolinx CL checker on the machine translatability of Cochrane PLS.

CHAPTER 7

ASSESSING THE MACHINE TRANSLATABILITY OF COCHRANE PLS

7.1 Aim of the Study on the Quality of Machine Translated Cochrane PLS and Overview of the Chapter

This chapter describes an experimental study which was carried out with the aim of determining whether, by introducing the Acrolinx CL checker into Cochrane’s non-automated simplification approach, there would be an increase in the machine translatability of Cochrane PLS. Alongside readability and comprehension, machine translatability (DV2.3) is treated as one of the three components of effectiveness — the DV2 of the empirical investigation described in this thesis. In Figure 7.1 we highlighted machine translatability to show how the experiment reported in this chapter relates to our broader investigation.

This chapter will begin with a summary of relevant research, followed by a description of the rationale behind this experiment, the RQ, and the research hypotheses that were tested. Subsequently, we will discuss: the recruitment of participants; the experimental environment and procedure; the tasks assigned to the participants; the experimental design; the texts under analysis; and the method adopted for the MT quality evaluation. Finally, we will present the analysis of the collected evidence and discuss the findings.

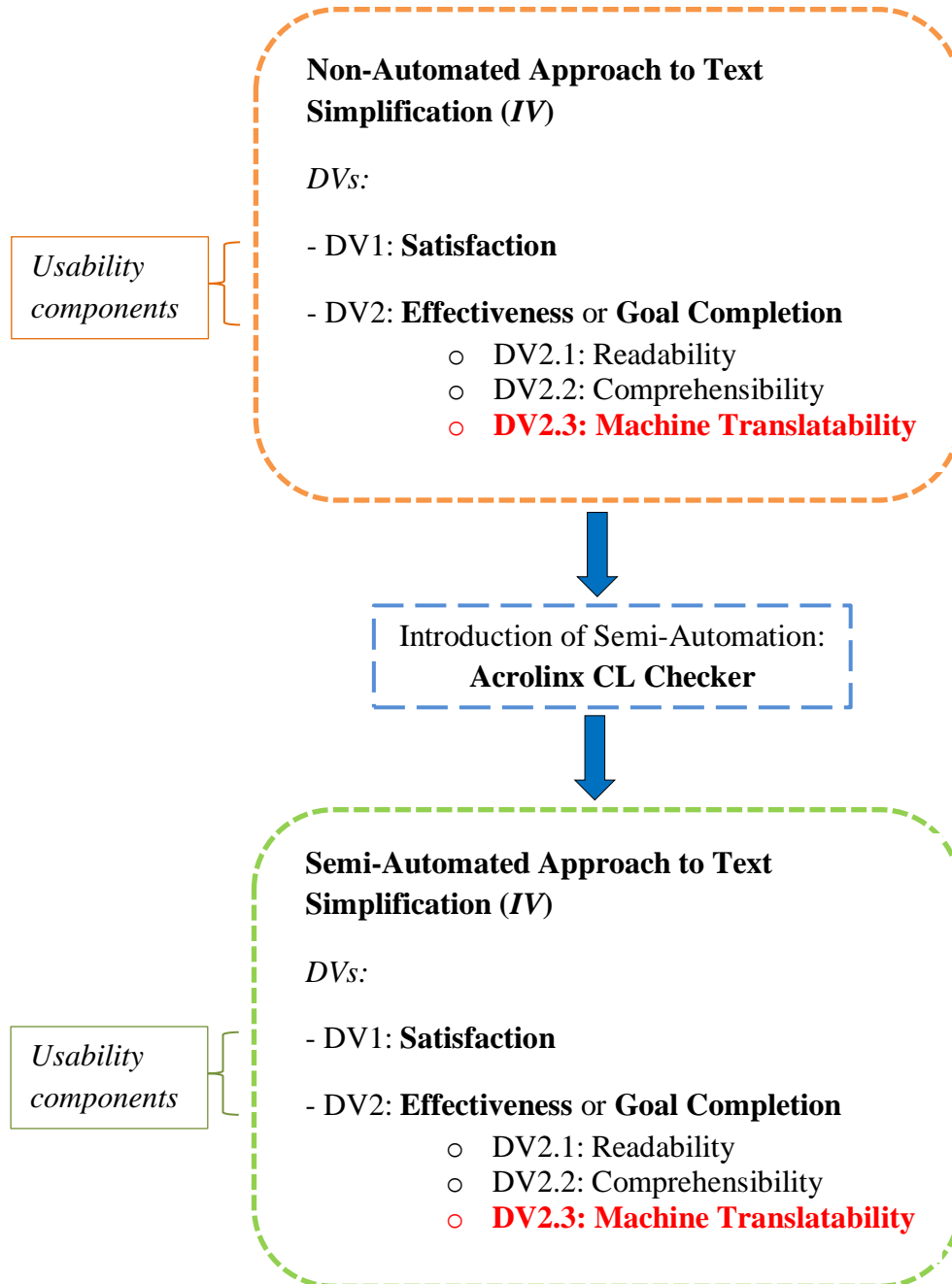


Figure 7.1: Machine translatability as DV2.3

7.2 Related Work on MT in the Health Domain

The majority of online health information is available in English only (Adams and Fleck 2015). To give a few examples, as of 2013, around 19% of the 155,000 Wikipedia medical articles were in English, while the rest was distributed across 255 natural

languages (Heilman and West 2015). At the European level, Internet users from the Czech Republic, Latvia, Bulgaria, Poland and Finland reported a dearth of online health information in languages that they could speak (European Commission 2014). Cochrane began to publish translations (from English) on its website only in 2012, and currently translations are available in a limited number of languages (Birch et al. 2017; *Translated Cochrane Evidence* 2018).

MT technology (in particular freely available MT systems such as Google Translate or Microsoft Translator) plays an important role in providing access to health information in multiple languages, thus addressing this language imbalance (Turner et al. 2015a). In particular, MT can prove beneficial for limited English proficiency (LEP) individuals (Dew et al. 2015). After conducting a survey on the understandability and usefulness of machine translated PLS among Cochrane users, Birch et al. (2018) found that about 75% of German respondents and around 50% of Czech and Romanian respondents preferred having machine translated texts (as opposed to English texts only). The authors (ibid.) argue that LEP users are more likely to benefit from the availability of an MT output. Interestingly, Birch et al. (2018) also point out that simplification of the English source texts (e.g. in terms of sentence length reduction) might improve the usefulness of MT.

There are numerous examples of development and use of MT in the health domain. MT is integrated into Canada's Global Public Health Intelligence Network (GPHIN), a multilingual early-warning system that collects and disseminates information on disease outbreaks and other public health threats (Blench 2008). GPHIN adopts a best-of-breed approach, whereby, for each language pair, the MT systems producing the highest quality levels are employed (ibid.). Epistemonikos⁴⁹, a freely available database of healthcare evidence, employs Google Translate when no official/human translation is available (Rada, Pérez and Capurro 2013). As of 2013, 99.6% of Epistemonikos articles had been translated with MT in languages other than English (ibid.). The French Cochrane Centre is using an English-French MT system

⁴⁹ The Epistemonikos database is available at: <https://bit.ly/2LORVul> [Accessed 12 December 2018].

developed by the multidisciplinary research consortium QUARTET M (Qualité de l'Aide à la Rédaction et de la Traduction; Evaluation du Transfert d'Information en Médecine) (The Translation Strategy Working Group 2014). Interestingly, the QUARTET M research consortium has also focused on the impact of simplified medical English on the machine translatability of Cochrane content (Von Elm et al. 2013). The Health in my Language (HimL) project aims to facilitate the multilingual dissemination of public health information by developing domain-adapted phrase-based and neural MT systems for translation from English into Czech, German, Polish and Romanian (Birch et al. 2017). Similar to the QUARTET M research consortium, the HimL project is particularly relevant to our investigation since the customised MT systems are meant to be integrated into the websites of both the National Health Service (NHS)⁵⁰ and Cochrane for the translation of PLS, among others (HimL 2016). Wołk and Marasek (2015) describe the effects of different training methods of neural and statistical MT systems for medical data and in the language pair Polish-English. In the Khresmoi project (whose goal is to build a multi-lingual and multi-modal system to access biomedical content), MT is used to translate user search queries and summaries of texts retrieved online (Dušek et al. 2014).

Despite their widespread use, there is wide agreement that MT systems tend to produce flawed outputs, particularly when translating specialised texts (Turner et al. 2014) from English into non-Western languages (Nguyen-Lu, Reide and Yentis 2010; Turner et al. 2015a). Evaluating the output of Google Translate on medical texts, Costajussà, Farrús and Pons (2012) found errors such as incorrect word order and word disagreement. Kirchhoff, Capurro and Turner (2012) showed that, among the different types of errors found in health-related texts machine translated into Spanish with Google Translate, word order errors were the most disliked type of error. Zeng-Treitler et al. (2010) employed the general-purpose MT system Babel Fish to translate medical records from English into Spanish, Chinese, Russian and Korean. Their results suggested that the MT output was often regarded as incomprehensible and inaccurate by the evaluators, especially in the case of Chinese, Russian and Korean translations.

⁵⁰ The NHS website is available at: <https://bit.ly/2MHwFas> [Accessed 12 December 2018].

However, in line with Birch et al. (2018) and the QUARTET M research consortium, Zeng-Treitler et al. (2010, p. 76) also consider the difficulty of the English source texts in their analysis:

The main cause of incomprehensible and incorrect translations appears to be the technical domain-related medical vocabulary on one hand, and irregular or complex syntax used by the original English sentences on the other. Longer sentences tend to have more complex syntax and a higher chance of containing difficult words.

Chen, Acosta and Barry (2016) also found that Google Translate — a statistical MT system at the time of their study — produced better quality when translating simple health-related sentences (as indicated by their Flesch-Kincaid Grade Level) from English into Spanish and Chinese.

In sensitive domains like health, a high level of accuracy is required since errors in the MT output may have detrimental effects on people's health (Costa-jussà, Farrús and Pons 2012). Therefore, when health-related texts need to be translated and distributed to the public, MT is traditionally used in combination with human validation and/or post-editing (PE), namely the editing and/or correction of MT output (Allen 2003). For instance, the Pan American Health Organisation (PAHO) has developed its own rule-based MT system (PAHOMTS), which has been operational since 1980 (Aymerich and Camelo 2009). PAHOMTS is adopted for more than 90% of the PAHO's translation jobs, in combination with PE carried out by in-house or external translators (*ibid.*). Similarly, Dew et al. (2015) describe the Public Health Automatic System for Translation (PHAST), a tool that allows for the integration of MT (the Microsoft Translator system) and PE into the translation workflow of public health departments in the United States, as well as for the remote collaboration of different health workers. The authors (*ibid.*) also suggest that the growing collection of texts translated through PHAST might lead to the development of a customised MT system for the public health domain.

Several studies have shown that, compared with human translation of healthcare information, MT followed by PE of health-related materials tends to result in substantial time and cost savings, while achieving the same level of quality as human translation

(Aymerich and Camelo 2009; Kirchoff et al. 2011; Gan 2012; Turner, Mandel and Capurro 2013; Turner et al. 2014). Nonetheless, this finding varies depending on the language pair analysed (Turner et al. 2015b).

To summarise, there is wide agreement that MT systems facilitate the assimilation and dissemination of online health content into multiple languages, and are particularly beneficial for LEP Internet users when human translation is not a feasible or affordable option. Numerous organisations are adopting or developing MT systems. However, due to the importance of accuracy of healthcare information, MT outputs require the validation and editing of language and/or domain experts prior to being circulated among the lay public.

This review of previous studies has also shown that simplification of the English source texts has been recognised as a factor that might increase MT quality (Chen, Acosta and Barry 2016). However, there is a dearth of empirical evidence on the impact of using plain/controlled English on the machine translatability of health-related texts when a neural MT system (like Google Translate at the time of our investigation) is used. As will be shown in Section 7.3, this research gap also extends to Cochrane content.

7.3 Motivation for Assessing the Quality of Cochrane PLS Machine Translated into Spanish, Research Question, and Research Hypotheses

Since 2014, Cochrane has been enacting a multi-language strategy whose goal is to increase the impact and accessibility of Cochrane evidence in non-English speaking countries (*Knowledge Translation in Multi-Languages* 2018). This strategy involves the translation of abstracts, PLS, podcasts, and entire cochrane.org websites into a variety of languages, including Croatian, French, Japanese, Polish, Tamil, Russian, and Spanish (ibid.). As of December 2017, 23,006 translations of abstracts and PLS had been published on Cochrane websites (Cochrane Translations 2018). Moreover, the Spanish version of the Cochrane Library, called *La Biblioteca Cochrane Plus*⁵¹, has been providing translations of Cochrane Systematic Reviews into Spanish since 2003 (*La*

⁵¹ *La Biblioteca Cochrane Plus* can be accessed at: <https://bit.ly/2OBasQP> [Accessed 12 December 2018].

Biblioteca Cochrane Plus 2018). On average, the Spanish version of the Cochrane Library was searched around 4,000,000 times per year between 2012 and 2014 (ibid.)

Translations are conducted through the Smartling Translation Management System⁵², and are produced by the Cochrane community, which is mostly characterised by volunteers with a health background. Moreover, even when translations are produced by professional translators, health experts revise the content for accuracy (*Translation at Cochrane: An Introduction* 2016). Translating Cochrane texts is a time-consuming and onerous task (Martikainen 2018), especially for volunteers. In order to make the translation effort more sustainable and to increase the number of available translations, the Cochrane translation community has been increasingly relying on MT (Von Elm et al. 2013; *Knowledge Translation in Multi-Languages* 2018). For instance, in the Cochrane translation strategy and business proposal (The Translation Strategy Working Group 2014), the development of MT software for Spanish is discussed, similar to the MT system developed for the English-French language pair by the QUARTET M research group (Section 7.2). An experiment conducted as part of the HimL project showed that, compared with translation from scratch, PE of the MT output produced by the HimL MT systems for Czech, German and Romanian resulted in time savings and was preferred by the translators (Birch et al. 2018).

In order to ensure accuracy, MT outputs are validated and post-edited by Cochrane health experts prior to being made available on the cochrane.org websites (Von Elm et al. 2013). Cochrane PE guidelines are oriented to light PE, “the objective of which is an accurate translation of essential content, in which stylistic considerations are secondary” (Martikainen 2018, p. 154). However, describing the adoption of MT for the translation of Cochrane abstracts, Ive (2017) reports that, when post-editing, Cochrane volunteer/health domain experts tend to either overlook or introduce errors (mainly inconsistent terminology). The author (ibid.) argues that the pre-editing of the source texts (namely, the identification and translation of difficult-to-translate segments carried out by professional translators prior to MT use) might improve the quality of the MT

⁵² Information on Smartling can be found at: <https://bit.ly/2CsSAQT> [Accessed 12 December 2018].

output, particularly in terms of terminological consistency. In addition to pre-editing, in Section 7.2, we presented other approaches adopted or considered at Cochrane for the improvement of MT quality and the reduction of the resulting PE effort — namely, the development of domain-adapted MT systems (e.g. as part of the HimL project), and the usage of a CL that would standardise and simplify the English source texts (Cochrane Collaboration Steering Group 2013).

To summarise, MT plays an increasingly important role in the Cochrane’s translation workflow since it facilitates the multilingual dissemination of evidence-based healthcare information by representing a preferred, faster and cheaper alternative to human translation. Health domain experts (with no linguistics background) need to validate and/or carry out PE in order to ensure content accuracy in the MT output. While there are studies on the benefits of using domain-adapted MT, there seems to be a lack of empirical evidence on whether and to what extent the adoption of a PL or a CL improves machine translatability into Spanish when a neural MT system like Google Translate is used (Section 7.2). The experiment described in this chapter aims to fill this research gap.

The impact of the Acrolinx CL checker on the machine translatability of Cochrane PLS was regarded as a worthy area of investigation because there is barely any mention or acknowledgement of (machine) translation in the PL guidelines that form Cochrane’s non-automated simplification approach (Section 4.3), and in PL guidelines in general (Section 3.3). In contrast, the Acrolinx CL checker automatically and consistently flags both readability and translatability issues in a text, while at the same time providing suggestions and examples on how to solve them (Section 4.3). Moreover, previous studies have mainly adopted domain-adapted MT systems, whose development is likely to be time-consuming and require funding. Here, we focused instead on the freely available MT system Google Translate, which is already being adopted by Epistemonikos to translate Cochrane abstracts into Spanish (Von Elm et al. 2013).

Concretely, we examined whether, compared with the implementation of Cochrane PL guidelines only, the introduction of the Acrolinx CL checker would

increase authors' effectiveness (DV2) and, more specifically, the level of machine translatability that they achieved in the PLS (DV2.3) as a result of simplification. In line with Izumi, Uchimoto and Isahara (2006, p. 484), we describe *machine translatability* as “a measure that indicates how well a given sentence can be translated by a particular MT system”. Machine translatability is therefore determined by the quality of the MT output.

Effectiveness is defined as “the accuracy and completeness with which users achieve certain goals” (ISO 9241-11:2018, 3.1.12). In Chapter 1 (Section 1.2), we specified that the RQ associated with effectiveness is the following:

RQ2: Does semi-automating a non-automated simplification approach by introducing a CL checker increase authors' effectiveness?

RQ2 was further segmented into three other questions (one per each of the goals of readability, comprehension, and machine translatability) (Figure 7.1). The RQ associated with machine translatability (RQ2.3) is the following:

RQ2.3: Does semi-automating a non-automated simplification approach by introducing a CL checker increase machine translatability?

The corresponding research hypotheses are:

H0: Semi-automating a non-automated simplification approach by introducing a CL checker does not increase machine translatability.

H1: Semi-automating a non-automated simplification approach by introducing a CL checker increases machine translatability.

In this chapter, we will describe the experiment conducted to address RQ2.3. It is important to remember that the 12 Cochrane authors that produced the PLS used as source texts had been asked to check their PLS for readability (rather than both readability and translatability) since they were likely to lack translation experience, and an instruction on translatability might have confused them (Section 4.5). However, we

did not deactivate Acrolinx translatability-oriented rules as they did not contravene Cochrane PL guidelines, and authors were not prevented from using them (Section 4.3).

7.4 Recruitment of Cochrane MT Evaluators

For the purpose of this experiment on MT quality, we aimed to recruit native speakers of Spanish among Cochrane health professionals/domain experts. The recruitment of these participants took place between October and November 2017, after receiving ethical approval from the Research Ethics Committee at DCU (DCUREC2017_149) (Letter of Approval in Appendix M), and after circulating the CFP (Appendix N). This CFP specified: (i) the names and affiliations of the researchers involved in the study; (ii) the requirements for participation (i.e. having a health background and being a native speaker of Spanish); (iii) the description of the tasks and the expected time commitment (i.e. around one hour); (iv) the indication that participation was on a voluntary basis, that participants could withdraw from the study at any point without repercussion, and that data would be treated confidentially; and (v) an invitation for interested participants to contact the researchers via email.

Similar to the studies described in Chapters 4, 5 and 6, this experiment was conducted within the context of the INTERACT project, in which Cochrane was one of the partners involved (Section 1.3). We could therefore avail of the assistance from Cochrane Iberoamérica, who advertised our CFP in their October newsletter and on their social media. Moreover, as was done for the recruitment of Cochrane authors, we published our CFP on TaskExchange (the online platform adopted by Cochrane contributors to advertise or volunteer for tasks). In summary, a random sampling recruitment technique was adopted, since every eligible member of the Cochrane community had an equal chance of being recruited as a participant (Saldanha and O'Brien 2013). In total, we recruited 41 eligible participants (Section 7.8.1). Since the CFP was distributed through different channels, it was not possible to calculate the response rate.

7.5 Experimental Environment, Procedure and Tasks

The eligible Cochrane contributors that contacted us via email or via TaskExchange to volunteer their participation were sent an email with the participant ID assigned to them (e.g. E05), and with the link to access the study, which was conducted online and remotely. By clicking on the link, participants could first read the PL statement describing the study, and the informed consent form. Subsequently, they could access a background questionnaire which contained five open-ended questions and six multiple-choice questions (Appendix O). The first question asked participants to insert the ID that had been assigned to them. The main aim of the background questionnaire was to collect data on the eligibility of the participants (i.e. being native speakers of Spanish and having a health background). Even though the eligibility criteria had been specified in the CFP (Appendix N), we asked these questions as an additional check. Furthermore, the background questionnaire was used to gather evidence on participants' characteristics that might have had an impact on their evaluation of the MT output, namely: (i) their level of familiarity with medical texts in English; (ii) their level of English proficiency; and (iii) their (frequency of) use of MT systems (Section 7.8.1).

After completing the background questionnaire, participants could access the first MT evaluation task. They were first presented with a detailed explanation of the task, along with instructions on how to conduct it (Appendix P). More precisely, participants/evaluators were told that the English PLS (source text) and its Spanish MT output (target text) would appear segmented at the sentence level, with each source sentence being followed by its corresponding target sentence. Evaluators had to answer two questions for each pair of sentences, one question dealing with the content/adequacy and one with the fluency of the Spanish translation (Appendix P), by scoring these aspects on a 4-point Likert scale (Section 7.7.2). After completing the first MT evaluation task, each participant could access the second (and last) MT evaluation task, which was characterised by the same instructions and adequacy/fluency questions, but on a different text (Section 7.6). A more detailed description of the adequacy and fluency measures will be provided in Section 7.7.2. However, in the interests of clarity, here we specify that adequacy (also called *accuracy* or *fidelity*) is determined by the

extent to which the MT output conveys the information of the source text (Linguistic Data Consortium 2002). On the other hand, fluency is described as the degree to which the MT output follows the grammar and the style of the target language (ibid.). Participants were not given any time limit to conduct the two MT evaluation tasks (Appendix P). Each participant conducted the evaluation tasks independently and anonymously, in order to avoid the influence of one evaluator over another (Doherty 2017). The presentation order of the sentences followed the structure of the text, rather than being randomised (ibid.).

Finally, after the last evaluation task was completed, each participant was sent a follow-up email in which they were asked if they had any comments on the quality of the machine translated texts, on the variety of Spanish (Section 7.6), or on any other aspect of the task (Appendix Q). Answers to these questions complemented the quantitative data collected during the adequacy and fluency scoring (Section 7.8.2).

Both the background questionnaire and the two MT evaluation tasks were presented to the participants on Google Forms. It should be noted that, prior to the main experiment, we conducted a small pilot study (with one participant having a background in linguistics) to ensure that there would be no technical issues when answering the questions on Google Forms.

7.6 Experimental Design and Experimental Materials

Similar to the authoring experiment described in Chapter 4, and the reading comprehension experiment described in Chapter 6, for this MT evaluation study we adopted a within-subject design — each participant was asked to evaluate the Spanish MT output of two Cochrane PLS, one non-automated and one semi-automated. We selected a within-subject design for the same reasons outlined in Section 4.6 — compared with between-subject designs, within-subject designs require a smaller sample of participants, and they allow researchers to control for individual differences (Lazar, Feng and Hochheiser 2010, pp. 48-49). Within-subject designs are therefore more appropriate when: (i) the sample is expected to be small because participants are highly educated professionals with busy schedules (as in the case of Cochrane contributors);

and (ii) the tasks involve activities whose success is heavily dependent on the participants' cognitive skills (such as the reading and evaluating of MT output in our experiment) (ibid.).

Despite these advantages, when adopting within-subject designs, it is important to take into account what MacKenzie (2013, p. 177) defines as *order effects*. In the case of our experiment, for instance, Cochrane participants might have been more lenient with the MT output in the second task after adjusting their quality expectations during the first task. They might also have felt more tired and less motivated while conducting the second MT evaluation task, thus showing less tolerance for inaccurate information or grammar errors. In order to compensate for order effects and avoid bias, we counterbalanced the order in which the MT outputs of non-automated and semi-automated PLS were presented to the evaluators. More precisely, the 41 participants recruited were divided into 12 groups. The first six groups were first presented with the Spanish MT output of a semi-automated PLS, followed by the Spanish MT output of a non-automated PLS. For the other six groups, the order was reversed. Evaluators were blinded to the design. In line with the need to avoid bias, we also asked each participant to evaluate MT outputs dealing with different health-related topics — if participants had evaluated the MT output of the same PLS (before and after the introduction of the Acrolinx CL checker), the evaluation of the second MT output might have been influenced by the comparison with the first one.

To obtain more accurate results, it would have been preferable to assign each participant with texts from their medical area of specialisation. However, this was not possible, mainly because each participant had to evaluate two texts on two different topics/areas (of which only one could be the area of specialisation). Therefore, we randomly assigned topics to groups by using an online random group creator⁵³. We assumed that, even though participants had never treated a specific illness or never prescribed a specific treatment/intervention, they would be familiar with the Spanish terms used to describe them considering their health background.

⁵³ The online random group creator is available at: <https://bit.ly/1g696SE> [Accessed 12 December 2018].

In the interest of clarity, this design is illustrated in Table 7.1, where we adopted the same method of text classification used in Chapter 6 (Table 6.1) — *NonAuPLS* indicates that the source text of the MT output was a non-automated PLS, while *AuPLS* indicates that the source text of the MT output was a semi-automated PLS. To indicate the Cochrane Review Groups to which texts belonged (and their topics), the following codes were used: Cochrane Eyes and Vision Group (EVG); Cochrane Injuries Group (ING); Cochrane Stroke Group (STG); Cochrane Common Mental Disorders Group (MDG); Cochrane Cystic Fibrosis and Genetic Disorders Group (CFG); Cochrane Vascular Group (VAG); Cochrane Dementia and Cognitive Improvement Group (DCG); Cochrane Heart Group (CHG); Cochrane Gynaecological, Neuro-oncology and Orphan Cancer Group (GNG); Cochrane Breast Cancer Group (BCG); and Cochrane Work Group (CWG). Table 7.1 also shows that between three and four participants evaluated each pair of MT outputs. Dyson and Hannah (1987) recommend at least three evaluators.

Groups	MT evaluation task 1	MT evaluation task 2	Number of evaluators
1	AuPLS (STG)	NonAuPLS (MDG)	4
2	AuPLS (MDG)	NonAuPLS (STG)	3
3	AuPLS (DCG)	NonAuPLS (CHG)	4
4	AuPLS (CHG)	NonAuPLS (DCG)	4
5	AuPLS (GNG)	NonAuPLS (BCG)	3
6	AuPLS (BCG)	NonAuPLS (GNG)	3
7	NonAuPLS (CWG)	AuPLS (VAG)	3
8	NonAuPLS (VAG)	AuPLS (CWG)	4
9	NonAuPLS (EVG)	AuPLS (VAG)	4
10	NonAuPLS (VAG)	AuPLS (EVG)	3
11	NonAuPLS (ING)	AuPLS (CFG)	3
12	NonAuPLS (CFG)	AuPLS (ING)	3

Table 7.1: Experimental design of MT evaluation experiment

A total of 24 PLS/source texts (between non-automated and semi-automated) were available to us (Section 5.4). Similar to the readability analysis described in Chapter 5, for this MT evaluation experiment we also adopted texts in their entirety. Non-automated PLS contained between 18 and 32 sentence pairs (source sentence and target sentence), while semi-automated PLS contained between 19 and 46 sentence pairs. When considering that participants would be asked to evaluate two MT outputs and answer two questions (on adequacy and fluency) on each of their sentences, we assumed that these tasks would already involve a considerable time commitment. For this reason, we decided to exclude abstracts from the MT evaluation. By including abstracts, our results would have been more comparable to the readability and comprehensibility scores.

The Spanish MT outputs of the PLS were obtained from Google Translate between October 6 and October 7, 2017. We selected Google Translate because this neural MT system is already being used by the Cochrane community (e.g. at Cochrane France) (Martikainen 2018). Therefore, its adoption enhanced the ecological validity of

our experiment. Moreover, Google Translate has been shown to produce high-quality MT output when translating titles of biomedical texts from English into Spanish, among other languages (Wu et al. 2011).

Prior to conducting the experiment, we searched for online information regarding the variety of Spanish that Google Translate would produce. However, we were unable to find an answer to this question. Therefore, we ran a small test with English words whose translations would vary between Castilian Spanish and Latin American Spanish, such as *car*, *computer*, *bus* or *apartment*. We observed that Google Translate would not consistently produce translations of a specific variety of Spanish. For instance, *car* was translated as *coche* (Castilian Spain), while *computer* was translated as *computadora* (Latin American Spanish). Even though this was just a small-scale test, it shed light on the need to take the variety of Spanish into consideration, as this might have influenced the scores assigned by participants. For this reason, in our follow-up email (Appendix Q), we also asked participants if they had any comments on the variety of Spanish characterising the MT outputs.

7.7 Method Adopted for Evaluating MT Quality

This section will deal with the method adopted to evaluate MT quality in our experiment, namely to determine whether there was an increase in the machine translatability of Cochrane PLS after the introduction of the Acrolinx CL checker. We will start by providing a brief overview of MT quality evaluation, with a special focus on non-profit settings. We will then present the most common automatic and human approaches to the evaluation of MT quality. Subsequently, we will delve into the measures of adequacy and fluency, the rationale behind their selection, and the way in which they were used for the purposes of our study. Finally, we will discuss the reasons for selecting domain experts as evaluators, as well the implications of our choice.

7.7.1 Overview of MT Quality Evaluation

The evaluation of the quality of MT output has been the object of numerous studies since the early development of MT itself (Castilho et al. 2018a). Similar to the assessment of text readability (Chapter 5), the evaluation of MT quality can be assigned

to the category of product-oriented research (Saldanha and O'Brien 2013). MT evaluation is also a component of the broader area of translation quality assessment (TQA), whose aim is “to ensure a specified level of quality is reached, maintained, and delivered to the client, buyer, user, reader, etc., of translated texts” (Doherty 2017, p. 131). As discussed in Castilho et al. (2018a), TQA represents a key area of investigation and debate in both academia and industry. However, the very notion of (machine) translation quality and the approaches used to measure it are characterised by substantial variability (ibid.; Drugan 2013). While the field of Translation Studies has traditionally adopted a theoretical and principled approach to the broader TQA, academic and industry research on MT quality has adopted a more pragmatic approach, in which evaluation of the MT output is more “a means to an end” (Doherty 2017, p. 133). For instance, in academia, MT evaluation determines future development needs; in industry, it allows for the testing and comparison of different commercial MT engines (Castilho 2016).

Interestingly, there seems to be a dearth of research on the quality of translations (either produced by humans or MT engines) at non-profit organisations like Cochrane. With regard to human translation, an exception is represented by Gigliotti (2017), who evaluated the quality of a corpus of texts translated by volunteers for four non-profit organisations (namely, Translators without Borders, The Rosetta Foundation⁵⁴, PerMondo⁵⁵, and Translations for Progress⁵⁶) by using an evaluative framework which included, among others, target language competence, textual competence, and cultural competence. Her findings showed that, in general, the quality of translations was poor. However, as the author herself (ibid.) points out, the small scale of the corpus used (i.e. 12 texts) does not allow for generalisability of results. As far as the quality of MT outputs for non-profit organisations is concerned, to the best of our knowledge, the only example of evaluation was conducted as part of the HimL project (Section 7.3), in which

⁵⁴ Translators without Borders and The Rosetta Foundation have now merged. The Translators without Borders website is available at: <https://bit.ly/2fnRGvx> [Accessed 12 December 2018].

⁵⁵ The website of PerMondo is available at: <https://bit.ly/2nasjhi> [Accessed 12 December 2018].

⁵⁶ The website of Translations for Progress is available at: <https://bit.ly/2RYFDnf> [Accessed 12 December 2018].

Cochrane was involved. More precisely, Birch et al. (2016) adopted a human semantic evaluation measure (HUME), whose goal was to determine the amount of meaning of the source sentence that is preserved in the MT output, similarly to adequacy (Section 7.7.2).

The scarcity of TQA, and in particular of MT evaluation, in non-profit settings is surprising when considering that organisations like Cochrane and Translators without Borders are increasingly relying on MT technology to speed up the translation process (*Introducing Kiswahili for Microsoft Translator* 2015). Castilho et al. (2018a, p. 31) argue that “[a]ny TQA method aims to minimise risk, whether this is a risk to communication, to reputation, or a risk of injury or death”. A wider adoption of TQA would therefore be particularly recommendable in non-profit settings, where translated texts often contain critical information, and where inaccuracies are likely to have detrimental effects on the health and well-being of end users (*Our Work* 2018).

7.7.2 Selection and Adoption of Adequacy and Fluency Measures

TQA, including evaluation of MT quality, can be either automatic or carried out by human evaluators. Numerous automatic evaluation metrics (AEMs) have been developed, which are software programmes providing a numeric score on the MT output based on its similarity with the human translation of the same sentence/text (Doherty 2017; Saldanha and O’Brien 2013). In other words, human translation represents the reference against which MT is automatically evaluated. Examples of widely adopted AEMs are: BLEU (Bi-Lingual Evaluation Understudy) (Papineni et al. 2002); TER (Translation Error Rate) (Snover et al. 2006); and GTM (General Text Matcher) (Turian, Shen and Melamed 2003). Although AEMs provide quick and consistent/objective quality scores, they have numerous limitations. For instance, it is not clear whether the human translation used as reference is assessed for quality (Saldanha and O’Brien 2013). More importantly, by focusing on the similarities between the target sentences/texts (i.e. MT output and corresponding human translation), AEMs do not consider the source text, thus not accounting for potential adequacy issues, such as content omissions and/or inaccuracies (Way 2018). As discussed in Section 7.2,

accuracy plays a paramount role in the health domain. Therefore, with their focus on fluency only, AEMs were not deemed appropriate for the purpose of this study.

We adopted human evaluation instead — more specifically, adequacy and fluency measures. As reported in Section 7.5, adequacy is determined by the amount of source text content that is kept in the target text/MT output; on the other hand, fluency refers to the extent to which the MT output is grammatically correct and reads naturally (Saldanha and O’Brien 2013, p. 104). Evaluation of MT quality based on adequacy and fluency can be assigned to the broader category of declarative evaluation, “which addresses how an MT system performs relative to various dimensions of translation quality” (Way 2018, p. 164). One of the first attempts at human MT evaluation was a four-year initiative at the Defense Advanced Research Projects Agency, where comprehension, adequacy and fluency measures were adopted, among others (White, O’Connell and O’Mara 1994). It is worth mentioning that adequacy and fluency measures were also adopted by Miyata et al. (2017) in a similar experiment, where the usability of a CL authoring assistant was evaluated in terms of the machine translatability of the source (Japanese) texts produced (Section 2.3).

As reported in Section 7.5, non-automated and semi-automated PLS were presented to participants segmented at the sentence level, with each source sentence immediately followed by the corresponding target sentence produced by Google Translate. While this set-up is very common for adequacy and fluency measures (Saldanha and O’Brien 2013), it entails the limitation of not considering characteristics of the MT output at the document level (Section 8.4). For instance, it is not possible for evaluators to determine whether text cohesion is maintained in the MT output (Doherty 2017).

Each pair of source sentence - target sentence was followed by two questions, one on adequacy and one on fluency. Since the MT evaluators in our experiment did not have a linguistics/translation background (Section 7.7.3), we made the concepts of adequacy and fluency explicit in the questions (Appendix P). This decision is in line with one of the recommendations in Doherty (2017, p. 141), who argues that “[i]nstructions and guidelines should be written clearly [...] and contain explicitly

operationalized definitions appropriate to the evaluator group”. More precisely, instead of asking participants to rate *adequacy*, we asked “[h]ow much of the information contained in the English source sentence (SS) appears in the Spanish target sentence (TS)?”, from 1 (none of it) to 4 (all of it). Similarly, the question on *fluency* was the following: “[i]ndicate the extent to which the Spanish target sentence (TS) is in grammatically well-formed and fluent Spanish”, from 1 (incorrect and disfluent) to 4 (correct and fluent). A similar wording of the questions is reported in Castilho et al. (2017). A 4-point Likert scale was adopted in order to avoid mid-point bias (ibid.).

In addition to adequacy and fluency, other popular types of human evaluation of MT output are error typology and ranking. Error typology involves the adoption of a list of error types, which are assigned penalties and classified as major or minor (Saldanha and O’Brien 2013). While allowing for a somewhat objective scoring of translations, applying error typology is a time-consuming task, which also requires training the evaluators (ibid.). Since our MT evaluators were health professionals with busy schedules who were volunteering their time, error typology was not deemed appropriate for this experiment. Ranking is another type of human evaluation which is typically employed to compare the output of two or more different MT systems from the same source text (Castilho et al. 2018a). More precisely, a source sentence is usually displayed along with the different target sentences produced by the MT systems under evaluation (ibid.). Contrastive ranking can lead evaluators to make more informed judgments (Koehn and Monz 2006). Nonetheless, in this study, it was not possible to conduct a ranking task because we adopted only one MT system and two different source texts (namely non-automated and semi-automated PLS) (Section 7.6).

7.7.3 Rationale behind the Recruitment of Domain Experts and Implications

We recruited MT evaluators among health professionals/domain experts to enhance the ecological validity of our experiment — in Section 7.3 we specified that, at Cochrane, volunteer domain experts carry out most of the translations, and are also involved in the process of validation and/or PE of MT outputs prior to their dissemination to the public (*About Translation at Cochrane* 2018). Moreover, due to resource constraints, it was not

possible to access evaluators with a linguistics/translation background, or to train them. As pointed out in Castilho et al. (2018a), in MT research scenarios, the recruitment of trained evaluators represents the exception. As an alternative, it would have been interesting to recruit evaluators among lay people (not familiar with Cochrane/medical content), and assess MT quality by means of reading comprehension tests of the MT outputs, similarly to what was done with the English source texts in this thesis (Chapter 6). However, this group would not have represented a realistic target audience, since Cochrane does not disseminate raw/unedited MT output among the lay public (*About Translation at Cochrane* 2018). Furthermore, considering the importance of content accuracy in the health domain, presenting lay people with the raw output might have led them to read and remember potentially incorrect and harmful information.

One of the benefits of recruiting evaluators with a health background was that, differently from professional translators, their scores were unlikely to be biased by the perception of MT as a threat (Roturier 2006; Cadwell, O'Brien and Teixeira 2018). However, preferring this sample of participants over professional translators/linguists entailed the limitation of not being able to ensure that their level of English proficiency was sufficient to conduct the evaluation task — especially in the case of adequacy, whose assessment also requires source language competence (Castilho et al. 2018a). To identify and account for potential differences in English proficiency among our MT evaluators/domain experts (all native speakers of Spanish), and as part of our background questionnaire (Appendix O), we asked participants: (i) to self-report their level of English proficiency by indicating how well they spoke English, i.e. whether very well, well, not well, or not at all (Vickstrom et al. 2015); and (ii) to take a short online English test available on the Cambridge English website⁵⁷ and containing 25 multiple-choice questions (Parra Escartín et al. 2017). The Cambridge English test provides a score and the corresponding Common European Framework of Reference for Languages (CEFR) level, from A1 to C2. During data analysis, we calculated mean adequacy scores from all evaluators first, and then only from evaluators who received a

⁵⁷ The Cambridge English test is available at: <https://bit.ly/2MeunUi> [Accessed 12 December 2018].

score corresponding to a B1 CEFR level or higher, in order to identify and remove potential bias caused by low levels of English proficiency (Section 7.8.2). B1 was selected as a threshold because, differently from A1-A2 levels, users at B1 level “[c]an understand the main points of clear standard input on familiar matters regularly encountered in work, school, leisure, etc.” (Council of Europe 2011, p. 5).

7.8 Data Analysis and Results

This section presents the analysis of the data collected from Cochrane MT evaluators through: (i) a background questionnaire on their eligibility and characteristics; (ii) two evaluation/scoring tasks of Spanish MT outputs of Cochrane PLS; and (iii) follow-up questions. The main goal of this TQA was to determine whether the integration of Acrolinx into Cochrane’s standard non-automated simplification approach would increase the effectiveness of Cochrane authors in terms of the level of machine translatability of the PLS that they produced (DV2.3).

This section will be further divided into two sections. Section 7.8.1 will present the findings on the eligibility and background characteristics of the Cochrane MT evaluators who took part in our study. Section 7.8.2 will report quantitative and qualitative evidence on the MT evaluation tasks.

7.8.1 MT Evaluators’ Background Characteristics

Fifty-two Cochrane contributors contacted us (either via TaskExchange or via email) to show interest in participating in our study. They were therefore sent the link to access our study. Of these 52, ten did not complete the study, even though we sent them reminders. One participant was excluded because they did not work in the health field. Therefore, the data presented here have been collected from a sample of 41 participants, all native speakers of Spanish and all with a background in health.

Most evaluators worked as health practitioners (n=22); eight of them were academics, and two of them reported working both as academics and health practitioners. Of those remaining, eight were students/trainees and one was a pharmaceutical consultant. Similarly to what was done for Cochrane authors (Section 4.8.1), when answers regarding jobs were vague, we tried to collect more data from the

participants' email signatures or from their online profiles. In the emails that they sent to volunteer for our study, eight of the 41 evaluators also reported having some experience as linguists or translators. For instance, E02 wrote:

I have a background in the areas of dentistry, public health, epidemiology and biostatistics, and pedagogy. I currently work as a teacher in the areas of health, languages and social sciences. I am also training to become a certified translator. I believe I am suitable to take part in this study, and I hope you take me into account.

All but one evaluator reported reading health-related texts in English. However, there was variability in the number of years participants had spent reading English medical texts — results for this question are reported in Figure 7.2. The vertical axis represents the number of respondents who selected each option. One participant whose job was internal medicine physician answered “84”. We assumed that that number referred to months (rather than years). It can be observed that the vast majority of evaluators (n=36) had been reading English health-related texts for at least five years.

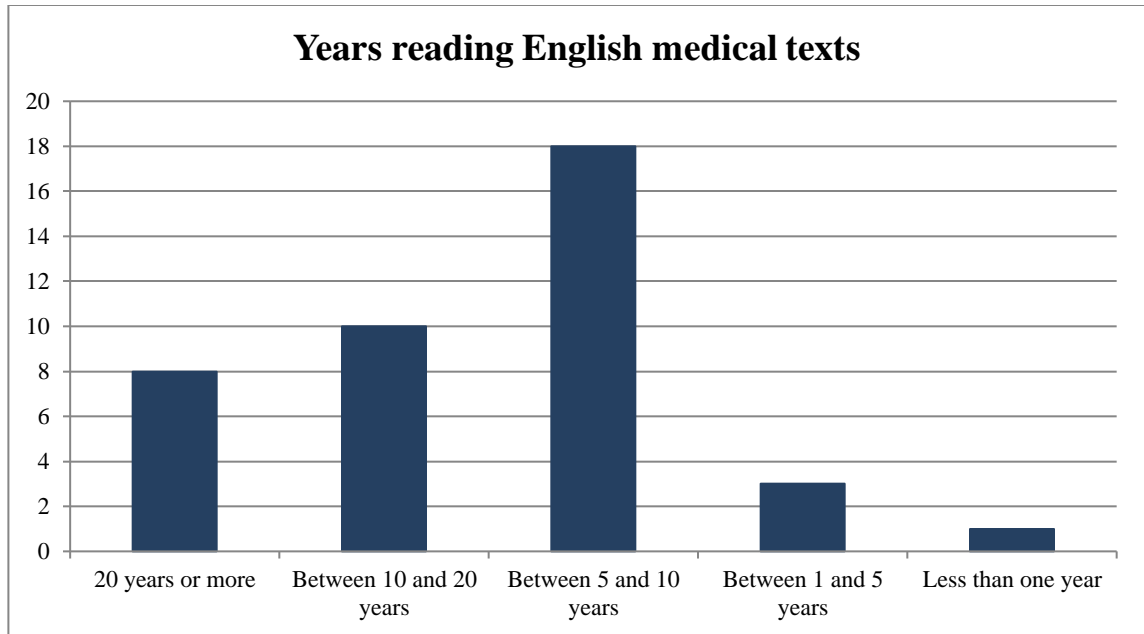


Figure 7.2: Years MT evaluators had been reading English health-related texts

To collect further data on evaluators' familiarity with medical texts in English, we also asked them to provide an estimate (as accurate as possible) of the hours spent reading these texts per month, on average. Results are reported in Figure 7.3, where the vertical axis represents the number of participants who selected each option. One answer was excluded because it was unintelligible. It emerged that most participants (n=25) read health-related texts in English at least 10 hours per month, on average. Overall, data in Figure 7.2 and 7.3 seem to indicate that our MT evaluators were familiar with accessing health content in English, despite being native speakers of Spanish. This finding is not surprising when considering that most online healthcare information is available in English only (Adams and Fleck 2015).

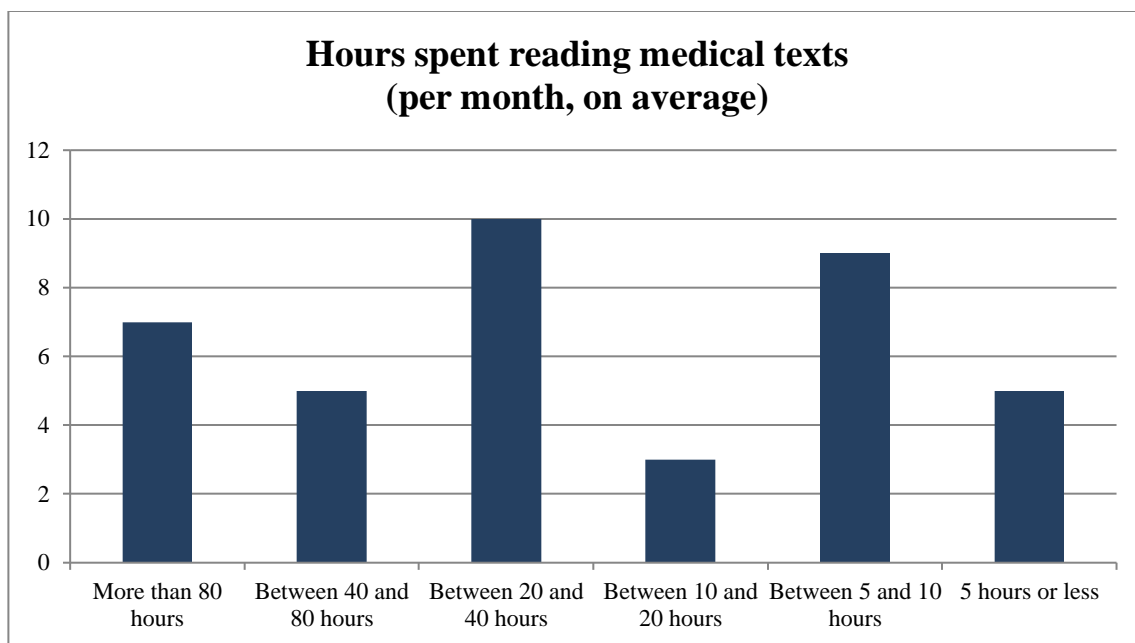


Figure 7.3: Hours (per month) spent by MT evaluators reading medical texts, on average

With regard to level of English proficiency, participants were asked to: (i) self-report it; and (ii) take a short English test that would assign them to a specific CEFR level. In Table 7.2, we show the number of participants automatically assigned to each CEFR level, alongside the self-reported level that participants selected for themselves.

For some participants, there was not an exact correspondence between self-reported level and CEFR level automatically assigned by the Cambridge English test. For instance, one participant answered “Not at all” to the self-rated English-ability question, but was then classified as C1 (i.e. proficient) English user. On the other hand, five participants who were assigned an A2 level (indicating that they were basic users) answered “Well” to the self-rated English-ability question. Although several studies have demonstrated the validity of the self-reporting question (Vickstrom et al. 2015), this lack of correspondence might be due to: (i) participants not engaging with the short Cambridge English test, thus receiving a lower score than the one matching their actual proficiency level; or (ii) participants having different standards against which self-assess their English proficiency (e.g. against native speakers of English vs. other native speakers of Spanish) (Siegel, Martin and Bruno 2001). When observing the more objective results of the Cambridge English test (Table 7.2), it emerges that the majority

of participants (n=32) had a B1 level of English or higher, indicating that they could understand the main content of familiar texts, such as medical texts (Section 7.7.3).

<u>CEFR level</u> <u>(and n. of participants assigned to it)</u>	<u>Answers to</u> <u>How well do you speak English?</u> <u>(and n. of participants</u> <u>which selected each answer)</u>
A1 (1)	Not well (1)
A2 (8)	Not well (3) / Well (5)
B1 (13)	Not well (1) / Well (10) / Very well (2)
B2 (5)	Well (5)
B2 - C1 (4)	Well (3) / Very well (1)
C1 (2)	Well (1) / Not at all (1)
C1 - C2 (5)	Well (3) / Very well (2)
C2 (3)	Very well (3)

Table 7.2: MT evaluators' (self-reported) level of English proficiency

Regarding familiarity with MT systems, the vast majority of participants (n=36) reported using MT systems such as Google Translate when encountering information in an unknown language. Of these 36, the majority (n=28) also stated that they used MT systems either always or frequently (Figure 7.4).

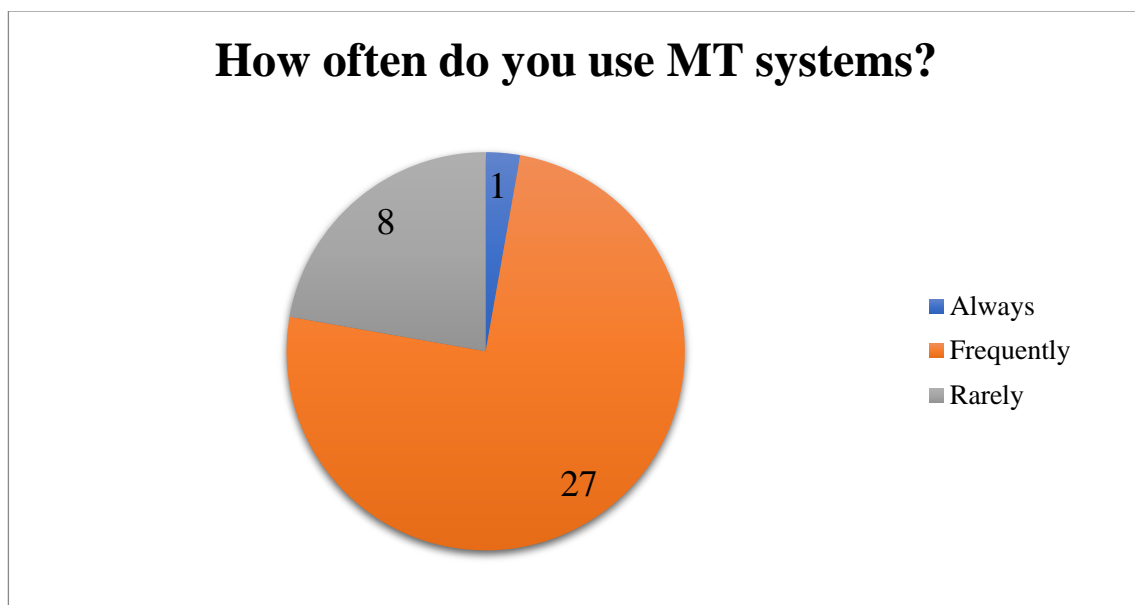


Figure 7.4: MT evaluators' frequency of use of MT systems

7.8.2 Quality of Cochrane PLS Machine Translated into Spanish

As discussed in Section 7.6, each participant conducted two scoring/evaluation tasks on two different Spanish MT outputs, one obtained from a non-automated Cochrane PLS while the other from a semi-automated Cochrane PLS. To analyse the quality of the Spanish MT outputs (and, in turn, the machine translatability of the PLS), we first calculated the average adequacy and fluency score (across all text sentences) per evaluator (Koehn and Monz 2006). Moreover, we calculated a grand mean of fluency and adequacy scores for both corpora to get an overall picture of the evaluations. Subsequently, for each evaluator, and for both the adequacy and the fluency measure, we conducted an independent-samples t-test to determine if the average score that each evaluator assigned to the output of the non-automated PLS was significantly different from the average score that they assigned to the output of the semi-automated PLS.

The independent-samples t-test assesses if the means of two independent groups are significantly different (Meyers, Gamst and Guarino 2013, p. 463). In our independent-samples t-test: the IV (categorical) was the simplification approach adopted, and its levels were represented by the absence or integration of Acrolinx; the DV (continuous) were the adequacy and fluency scores; the independent samples were the non-automated and the semi-automated PLS in each group, which belonged to a

different corpus and dealt with a different health-related topic (Table 7.1). When the assumption of normality was not met (as indicated by the p-value of the Shapiro-Wilk test lower than 0.05), we ran the nonparametric version of the independent-samples t-test, namely the Mann-Whitney U test (*SPSS Tutorials: Independent Samples t Test* 2018).

Table 7.3 shows: (i) the mean adequacy score (and its SD), across all sentences, assigned by each MT evaluator to both the non-automated and the semi-automated PLS; and (ii) the grand means of the adequacy scores, across all MT evaluators. Table 7.4 reports the same results, but for the fluency measure. In both tables, statistically significant differences are signalled with an asterisk.

Group	Evaluator	Adequacy scores of non-automated PLS	Adequacy scores of semi-automated PLS
		Mean (SD)	Mean (SD)
1	E13	3.73 (0.44)	3.89 (0.31)
	E25	4 (0)	4 (0)
	E01	3.43 (0.66) (*)	3.94 (0.22) (*)
2	E37	4 (0)	4 (0)
	E14	3.88 (0.47)	4 (0)
	E02	4 (0)	3.96 (0.2)
3	E26	3.88 (0.47)	3.68 (0.69)
	E15	3.76 (0.53)	3.72 (0.45)
	E03	4 (0)	3.97 (0.16)
4	E27	4 (0)	3.89 (0.31)
	E39	3.52 (0.81)	3.13 (0.78)
	E16	4 (0)	3.86 (0.44)
	E04	3.89 (0.4)	3.82 (0.46)
5	E40	4 (0)	3.89 (0.3)
	E28	3.93 (0.25)	3.82 (0.6)
	E17	4 (0)	4 (0)
	E05	2.7 (0.6)	2.54 (0.8)
6	E41	3.48 (0.75)	3.68 (0.47)
	E18	3.55 (0.82)	3.5 (0.84)
	E42	3.9 (0.44)	3.75 (0.61)
7	E30	3.95 (0.22)	3.93 (0.25)
	E31	3.85 (0.6)	4 (0)
	E49	3.25 (0.98) (*)	3.78 (0.61) (*)
8	E19	3.29 (0.77) (*)	3.95 (0.21) (*)
	E08	3.86 (0.35)	3.89 (0.3)
	E20	3.93 (0.25)	4 (0)
	E32	3.55 (0.73)	3.34 (0.76)
9	E44	3.82 (0.46)	3.93 (0.37)
	E21	3.93 (0.25)	3.96 (0.19)
	E09	4 (0)	4 (0)
	E33	3.83 (0.46)	3.85 (0.36)
10	E45	3.86 (0.34)	3.77 (0.42)
	E10	3.63 (0.72)	3.77 (0.61)
	E34	3.77 (0.52)	3.83 (0.37)
11	E46	3.86 (0.35)	3.96 (0.17)
	E23	3.82 (0.38)	3.76 (0.6)
	E50	3.73 (0.54)	3.8 (0.54)
12	E52	2.47 (0.59) (*)	3.5 (0.69) (*)
	E12	3.34 (0.7)	3.5 (0.65)
	E24	3.84 (0.36)	3.75 (0.44)
	E48	3.46 (0.62)	3.7 (0.46)
GRAND Means (SD)		3.72 (0.33)	3.78 (0.27)

Table 7.3: Descriptive and inferential statistics for adequacy scores

Group	Evaluator	Fluency scores of non-automated PLS	Fluency scores of semi-automated PLS
		Mean (SD)	Mean (SD)
1	E13	3.56 (0.58)	3.47 (0.77)
	E25	3.17 (1.07)	3.21 (1.13)
	E01	2.26 (1.17)	2.73 (1.04)
	E37	3.65 (0.57)	3.63 (0.68)
2	E14	3.11 (0.9) (*)	3.64 (0.56) (*)
	E02	3.38 (0.69)	3.36 (0.63)
	E26	3.33 (0.76)	3.12 (1.01)
3	E15	3.71 (0.56)	3.62 (0.59)
	E03	3.71 (0.46)	3.59 (0.64)
	E27	3.04 (1.07)	3.13 (0.91)
	E39	2.8 (0.81)	2.67 (0.7)
4	E16	3.79 (0.49)	3.65 (0.61)
	E04	3.79 (0.55)	3.65 (0.72)
	E40	3.24 (0.57)	3.44 (0.63)
	E28	3.68 (0.66)	3.75 (0.73)
5	E17	2.85 (0.76)	3.13 (0.63)
	E05	2.07 (0.72)	1.95 (1.04)
	E41	2.55 (1.01)	2.77 (0.86)
6	E18	3.9 (0.3)	3.72 (0.54)
	E42	3.45 (0.75)	3.54 (0.72)
	E30	2.55 (1.19)	2.7 (1)
7	E31	3.59 (0.79) (*)	3.92 (0.34) (*)
	E49	3.22 (0.93)	3.53 (0.67)
	E19	3 (0.96)	3.36 (0.58)
8	E08	3.75 (0.51)	3.86 (0.35)
	E20	3.89 (0.3)	3.96 (0.18)
	E32	3.13 (1.02)	2.62 (1.04)
	E44	3.68 (0.54)	3.62 (0.67)
9	E21	3.8 (0.48)	3.51 (0.7)
	E09	3.86 (0.34)	3.85 (0.45)
	E33	3.46 (0.62)	3.59 (0.57)
	E45	3.36 (0.88)	3.37 (0.92)
10	E10	3.09 (1.1)	3.61 (0.84)
	E34	2.22 (0.97)	2.09 (0.9)
	E46	3.04 (0.84)	3.35 (0.7)
11	E23	3.69 (0.47)	3.84 (0.36)
	E50	3.78 (0.51)	3.63 (0.71)
	E52	1.04 (0.2) (*)	1.6 (1.1) (*)
12	E12	3.78 (0.42) (*)	4 (0) (*)
	E24	3.9 (0.29)	3.87 (0.33)
	E48	3.31 (0.78)	2.91 (0.77)
GRAND Means (SD)		3.27 (0.6)	3.33 (0.55)

Table 7.4: Descriptive and inferential statistics for fluency scores

As can be observed in Tables 7.3 and 7.4, the average adequacy score per evaluator ranged from 2.47 to 4, while the average fluency score per evaluator ranged from 1.04 to 4. In terms of adequacy, the number of evaluators who rated the output of semi-automated PLS higher (n=19) was just one greater than the number of evaluators who rated the output of non-automated PLS higher (n=18) — the remaining four evaluators assigned the same average score to both MT outputs. With regard to fluency, the number of evaluators who assigned a higher average fluency score to the output of semi-automated PLS (n=22) was slightly higher than the number of evaluators who assigned a higher average fluency score to the output of non-automated PLS (n=19).

Unsurprisingly then, for both the adequacy and the fluency measure, differences in the grand means of the two corpora were slight. Moreover, most differences in average scores were not statistically significant. The only statistically significant increases in adequacy and fluency scores (i.e. for eight participants in total) were observed for the corpus of MT outputs of semi-automated PLS (Tables 7.3 and 7.4). After excluding evaluators with A1-A2 level of English proficiency (Section 7.7.3) and recalculating the grand mean of adequacy scores, we observed that the difference between the grand mean score assigned to the outputs of non-automated PLS (M=3.76, SD=0.28) and the grand mean score assigned to the outputs of semi-automated PLS (M=3.82, SD=0.25) remained slight.

Overall, it was not possible to identify a majority of evaluators assigning higher scores to the MT outputs of a specific corpus. Moreover, there was little difference in evaluators' mean ratings of fluency and adequacy between the two corpora of PLS, thus indicating that the MT system used (i.e. Google Translate) did not consistently produce better raw MT quality when Acrolinx was employed.

Even though the differences in machine translatability between the two corpora of PLS were slight, two other findings emerged from the data in Tables 7.3 and 7.4. First of all, the grand means of the adequacy scores (Table 7.3) were higher than the grand means of fluency scores (Table 7.4). Secondly, it emerged that the mean fluency and adequacy scores were relatively high, which suggests that Google Translate produced reasonably good raw MT quality. We further delved into these findings by combining

the quantitative data in Tables 7.3 and 7.4 with the comments provided by the evaluators as a reply to our follow-up email (Appendix Q), bearing in mind that “[w]hen assessing a complex ‘product’ like MT output, users are notoriously poor at analyzing their own judgments and stating them in explicit terms, especially when they lack linguistic training” (Kirchhoff, Capurro and Turner 2014, p. 5). Moreover, despite several reminders, 15 of the 41 evaluators did not reply to our follow-up email. Therefore, qualitative data could only be collected from 26 participants.

The comments provided by the evaluators fully or partially supported the scores that they had assigned. For example, E25 rated the adequacy of both PLS higher than their fluency. They explained the reason for these scores in the following comment, from which it emerges that all the information in the English source PLS appeared in the Spanish MT output (adequacy). However, vocabulary and sentence structure in the MT output were not always perceived as natural (which decreased fluency scores):

In all the sentences in Spanish I could find all the components that the sentence in English has. But sometimes the words that the machine used wasn't the right word or not the most used word in Spanish, then the sentence in Spanish sound [sic] weird.

Similarly, E42, E26, E32, E21, E46 and E14 assigned higher scores to the adequacy of both PLS, compared with fluency. This is reflected in their remarks. For instance, E14 discussed the segments that, although grammatically correct, would be perceived as unusual by a native speaker, such as the translation of *patient-reported outcomes* with *resultados informados por el paciente*, rather than the more common *desenlaces reportados por los pacientes*. E14 also added: “[t]he majority of the translations were great”. E46 wrote: “When you read it, you know that is a machine translator and not a native speaker”. Similarly, E26 commented:

The quality of the machine translated texts is very good, although [sic] as it is natural, we might change some words sometimes which makes it sound better. But still the message is clear and complete in all the translations.

In relation to fluency issues, participant E21's remark was:

My impression of the translators is that they make the translation very literally, and that makes lose [sic] the fluency in reading the translated, however the general context of what the translated document is about is well understood, which is very favourable.

E37 assigned high mean scores to both the semi-automated and the non-automated PLS, and to both the fluency and adequacy measures, but their mean fluency scores were lower than their mean adequacy scores. Their comment only partially reflects these scores, since, unlike E25, E14 or E26, this evaluator did not comment on fluency as being lower than adequacy: “[i]n general, I think the two group [sic] of sentences had a great level, not only in the content of he [sic] information but also in the grammatical structure”. Participant E34 also rated the fluency of the PLS lower than their adequacy. However, in their remark, E34 only discussed the overall quality of the MT output as good. Partial confirmation of the scores has also been provided by E27, who assigned lower scores to fluency than adequacy, but then in their comment reported the presence of both adequacy and fluency issues in the MT output:

Some subtitles in the document as “Background” are really wrong since this word in a systematic review means “Antecedentes” and in all examples reviewed use [sic] “Fondo”. [...] Overall, machine translations are well defined (about 75-80%); however, there are large sentences which miss a correct idea.

In line with the high mean scores provided, participants E31, E33, E20, E09 and E19 commented on the overall good quality, accuracy and spelling of the Spanish translations. E20, for example, wrote: “I think the overall machine translated texts were pretty great, I did found [sic] some minor synthaxis [sic] mistakes in very few long phrases, but besides that, everything was pretty much awesome”. Similarly, E09 commented: “the translations from English to Spanish accomplished the required level of understanding for clinical purposes”.

E15, who assigned high mean scores to both PLS, and on both measures of adequacy and fluency, made a comment on the overall quality that confirmed their scores: “[i]n general the translation was very good, only in [sic] some expressions don’t have the same meaning when they are traslated [sic] literally but these were very few translations”. Similarly, E04, E10 and E30 commented both on the overall quality of the

MT output and on some fluency issues — like the participants above, E04, E10 and E30 gave higher adequacy (than fluency) scores. E04’s remark was: “I consider that the translation has a good quality, and it is completely understandable. The only flaw I found is the translation of the word ‘background’ -> ‘fondo’. I think that it would be more appropriate [sic] the word ‘antecedents’ [sic]”. The same issue with the translation of *background* was identified by participant E10. E30 commented that the quality of the translation was quite acceptable, but it was often literal, thus shedding light, again, on potential issues of fluency.

Despite the high mean scores, fluency issues were also identified by evaluator E49, who commented:

I think most of the translated text goes from acceptable to high quality (I have to admit I've been very demanding while scoring the results). However there are some sentences where the translation loses [sic] -at least partially- its original meaning, e.g., although the Spanish verb “intimidar” can be considered a synonym of the English word “bullying”, I think the Spanish term “acosar” would have been more appropriate in this context. On the other hand, it may be that machine translatability gets into trouble when facing some verbal tenses, e.g., the Spanish subjunctive (simple past in the English grammar) was not well captured by the machine, which makes the sentence not to sound [sic] fluent.

In the case of E01, who rated fluency lower than adequacy, fluency issues had a negative impact on the overall perception of the MT output:

I was appaled [sic] by the really poor translations. [sic] [...] The biggest issue was that the machine translated literally and word by word, which occasionally [sic] even disrupted what the original sentence meant to either change the meaning or make it incomprehensible [sic]. [...] I also observed that when I was trying to think of how you would say things in Spanish, because I already had this Spanish sounding translation it was actually really hard to think differently and more naturally.

Unsurprisingly, E01 was one of the few participants who reported not using MT systems for assimilation purposes (Section 7.8.1). This lack of familiarity with MT might have influenced their expectations regarding the quality of the output.

Further examples of full correspondence between scores and follow-up comments could be observed for participants E41 and E05. E41 rated fluency lower than adequacy for both texts, and also assigned lower mean scores to the non-automated PLS (presented to them as the second text), compared with the semi-automated PLS. These differences in scores were reflected in their comment:

The first text was better translated than the second one. It was quite good transferring [sic] the information. It was quite easy and nice to read. This second text had significant gramatical [sic] mistakes, most of them into the parentheses. The information was not properly translated, the meaning of the sentences sometimes was not the same. [...] It was very strange to find the word “SENO” in a scientific text. The correct word should be “MAMA”. [...] “BACKGROUND” word’s meaning, in the contest [sic], never would be “FONDO”. It could be “ANTECEDENTES”.

E41 also added that the variety of Spanish was appropriate. In the case of E05, the mean adequacy and fluency scores assigned were relatively low, compared to the rest of the evaluators. Unsurprisingly, this participant commented:

The quality of the machine translated texts in some sentences is acceptable, but other sentences are hard to understand, lead to confusion and misunderstanding and lack coherence and cohesion.

Similar to E41, participant E18 reported that they did not find any issues with the variety of Spanish adopted. However, differently from the evaluators listed above, E18 identified more adequacy than fluency issues and, as a result, provided mean adequacy scores that were lower than mean fluency score. Their remark was:

[T]he items I rated as 1 have crucial errors which in some cases changed the meaning of the sentence and would mislead the reader of the translated version. These instances, to me, present a crucial hurdle to overcome before widespread use of machine translation.

Finally, rather than discussing their scores, a few evaluators commented on the characteristics of source texts that, in their opinion, might have a negative impact on machine translatability. For instance, E48 mentioned the different word order for the combination of noun and adjective between English and Spanish.

In summary, only one evaluator mentioned potential differences between semi-automated and non-automated PLS, which confirms the very slight differences in the mean scores assigned to the two corpora (Tables 7.3 and 7.4). In contrast, numerous evaluators paid attention to the difference between the two measures of adequacy and fluency, and to the overall quality of the translations. In particular, most evaluators identified more stylistic and grammatical issues than content problems, a result which is in line with the overall lower fluency scores. Furthermore, as emerges from several comments reported above, generally evaluators were positively impressed with the quality of the MT output, which, again, confirms the high grand means reported in Tables 7.3 and 7.4. The variety of Spanish did not seem to influence the evaluation of the MT output.

Finally, we calculated inter-rater agreement (namely, the variation between multiple evaluators assessing the same sentences) using the intraclass correlation coefficient (ICC) (Koo and Mae 2016). We used a two-way mixed-effects model, whereby we assessed the reliability of the specific evaluators/raters recruited. In terms of measurement protocol, we adopted the mean of multiple evaluators, rather than the single measures option, which uses the scores of a single evaluator as the basis of the reliability measurement (*ibid.*). Furthermore, we focused on absolute agreement, which is determined by the extent to which the same evaluators assigned the same scores to the same sentences (*ibid.*). Absolute agreement was preferred to consistency, which is instead determined by the degree to which scores are linearly related (McGraw and Wong 1996).

Since the texts assigned varied across the 12 groups of evaluators (Table 7.1), a separate ICC was calculated for each of the 12 groups, and for both adequacy and fluency scores. It is important to remember that each group had between three and four evaluators (Section 7.6). ICC values are reported in Table 7.5, separately for adequacy and fluency. Particularly low ICC values are in bold. Koo and Mae (2016) provide guidelines on the interpretation of ICC — values lower than 0.5 indicate poor reliability; values between 0.5 and 0.75 suggest moderate reliability; values between 0.75 and 0.9

are indicative of good reliability; and ICC values higher than 0.9 suggest excellent inter-rater reliability.

Based on these guidelines, the ICC values reported in Table 7.5 show that: (i) when evaluating both adequacy and fluency, inter-rater agreement in most groups was either poor or moderate — a good level of agreement was reached only in group 4, for adequacy; (ii) in five of the 12 groups, the same level of inter-rater agreement was reached in fluency and adequacy evaluations — for instance, in group 1, inter-rater agreement was moderate for both fluency and adequacy; and (iii) in four groups, a higher level of inter-rater agreement was reached for adequacy scores; whereas the remaining three groups showed higher inter-rater agreement when evaluating fluency.

<i>Groups</i>	<i>ICC (adequacy)</i>	<i>ICC (fluency)</i>
1	0.68	0.73
2	0.34	0.51
3	0.29	0.49
4	0.8	0.72
5	0.14	0.55
6	0.54	0.36
7	0.62	0.49
8	-0.05	0.31
9	0.36	0.62
10	0.67	0.4
11	0.43	0.12
12	0.48	0.04

Table 7.5: Inter-rater agreement of MT evaluators (per group) on adequacy and fluency scores

Overall, there was variability in the adequacy and fluency scores assigned by evaluators in each group. This result is not surprising when considering the subjectivity that characterises the human evaluation of MT output quality (Turchi, Negri and Federico 2014). The poor to moderate inter-rater agreement might also be due to the fact that, even though we collected data on participants' familiarity with online MT systems (Section 7.8.1), we did not train them in MT evaluation prior to the beginning of this experiment. This limitation will be further discussed in Sections 7.9 and 8.4. We also observed that evaluators did not seem to agree more on fluency than adequacy, or vice versa. This result is different from the findings in Mitchell (2015), where higher agreement was achieved on fluency scores among domain experts. However, it should be noted that, in Mitchell's study (ibid.), fluency was rated against a gold-standard translation, which is likely to have acted as a common benchmark for the evaluations.

7.9 Discussion and Summary of the Study on the Machine Translatability of Cochrane PLS

In this chapter, we have presented an experimental study whose goal was to answer RQ2.3 (Section 7.3), namely if introducing the Acrolinx CL checker into Cochrane’s non-automated simplification approach (thus rendering it semi-automated) would increase the machine translatability (into Spanish) of Cochrane PLS (DV2.3). To the best of our knowledge, no previous investigation has been conducted on the impact of Acrolinx on the MT quality of Cochrane PLS. In this section, we will summarise and discuss the main findings. We will also discuss the implications of this study for Cochrane and briefly introduce the last chapter of this thesis.

The evidence collected showed that using the Acrolinx CL checker to check Cochrane PLS for translatability (and readability) issues did not result in an increase in their machine translatability into Spanish, as assessed by human evaluators who were health domain experts and native speakers of Spanish. In other words, by reporting these results within the framework of usability (and, more precisely, effectiveness or goal completion), we could conclude that the effectiveness of Cochrane authors in terms of the level of machine translatability achieved in the PLS was not enhanced by the usage of Acrolinx. Therefore, the data point to the acceptance of our null hypothesis (Section 7.3).

Results reported in Section 5.7 had shown that, compared with non-automated PLS, semi-automated PLS were characterised by a significant decrease in word length and sentence length, and by a significant increase in syntactic simplicity. Considering that sentence length has been shown to influence the output of neural MT systems — with quality dropping for long sentences containing more than 20 tokens (Castilho et al. 2018b) — our results are surprising. In other words, since the introduction of the Acrolinx CL checker resulted in significantly lower sentence length, we would expect the machine translatability of semi-automated PLS to be significantly higher, based on findings from previous works. To explain this finding, we examined the mean sentence length of non-automated PLS (using Coh-Metrix 3.0) and found that, on average, sentences in non-automated PLS already contained less than 20 words ($M=17.07$,

SD=2.88). Therefore, despite being significant, the decrease in sentence length observed after using Acrolinx might have had only an incremental impact on the machine translatability of the sentences.

The analysis of the quantitative and qualitative data collected from the 41 MT evaluators also showed that the neural system Google Translate produced Spanish MT outputs of relatively high quality in terms of adequacy and fluency. This finding, which is in line with the results reported in Wu et al. (2011)⁵⁸, seems promising for Cochrane and other non-profit organisations that are increasingly relying on this technology to streamline their translation workflow and to encourage contributions from volunteers (Section 7.3). In turn, an increase in the number of contributions/translated texts is likely to make online health content more accessible for non-native speakers of English (especially LEP individuals and individuals with no knowledge of English).

We also observed that, while the style of the MT output was often described as unnatural by the evaluators, the content of the source English PLS was often translated fully and accurately into the Spanish output. This result is also encouraging since fluency issues are both easier to correct and less likely to have a detrimental effect on the well-being of readers, compared with adequacy/content errors (Koponen 2010; Stymne 2013). Moreover, stylistic issues seem to be of secondary importance for the Cochrane PE community (Section 7.3).

All our MT evaluators had a health background, and the vast majority of them were familiar with MT systems (Section 7.8.1). However, inter-rater agreement was between poor and moderate. This result might have been caused by the lack of shared training or experience in MT evaluation, which requires specific analytical skills, differently from using online MT systems for assimilation purposes. Recruiting participants who have (similar) experience in evaluating MT output, or providing them with training on MT evaluation might increase inter-rater agreement. However, despite experience or training, some degree of subjectivity is likely to always be present in human judgement (Castilho et al. 2018a).

⁵⁸ It should be noted that, when Wu et al. (2011) conducted their study, Google Translate was a statistical (rather than a neural) MT system.

With this chapter, we have concluded the experiments on the usability components under investigation (namely, the satisfaction of authors of PLS and their effectiveness in achieving the goals of readability, comprehension, and machine translatability before and after the introduction of Acrolinx) (Section 1.2). In the following (and final) chapter, we will outline the findings, implications, and contributions of the entire thesis. We will also present the limitations of this investigation and suggest areas for future research.

CHAPTER 8

CONCLUSIONS

8.1 Content and Organisation of the Chapter

This final chapter begins with an overview that summarises the goals of this thesis, its findings, and its practical implications, in line with the applied nature of this research (Section 1.3). Subsequently, we will highlight our main contributions to empirical knowledge, methodological knowledge, and knowledge of practice. Finally, we will discuss the limitations of this work and suggest how future research might address them.

8.2 An Overview of the Thesis: Goals, Findings, and Implications

Our intent with this thesis was to fill several research gaps in relation to the authoring, reading comprehension, and machine translation of online simplified health content. More precisely, we set out to empirically investigate the *usability* of the *text simplification* approach currently adopted at Cochrane for the production of PLS to be published online, and the impact of introducing the Acrolinx CL checker on the usability of the aforementioned approach. As specified in Section 1.2, our overarching RQ was:

RQ: Does semi-automating a non-automated simplification approach by introducing a CL checker increase usability?

To achieve this overarching goal, and within the framework of the ISO 9241-11:2018 definition of usability, we conducted experimental research on the *satisfaction* and *effectiveness* of Cochrane authors of PLS before and after the introduction of Acrolinx. In particular, effectiveness was measured in terms of the level of *readability*, *comprehensibility*, and *machine translatability* achieved by Cochrane authors in the PLS (Section 1.2).

Our departure points, grounded in the review of related literature, have been that: (i) manually simplifying medical content can represent an unsatisfactory experience — particularly for volunteer authors with no linguistics background — because checking and remembering different sets of PL guidelines are difficult and time-consuming tasks, and because authors are not provided with automatic feedback on the level of readability

or translatability achieved in their texts; (ii) manually simplifying medical content can lead to texts with low readability, comprehensibility, and translatability, as a result of contradictory and vague PL guidelines, or of their inconsistent application; and (iii) providing authors of simplified medical texts with technological assistance might be beneficial in terms of their satisfaction and effectiveness as PL writers. Here we report the results for each of the components of usability that represented our DVs (Section 1.2).

Our DV1 was the usability component of satisfaction, and the RQ associated with it was:

RQ1: Does semi-automating a non-automated simplification approach by introducing a CL checker increase authors' satisfaction?

The quantitative and qualitative data collected from a sample of Cochrane volunteer authors of PLS who had a medical background pointed to the acceptance of our alternative hypothesis, according to which introducing semi-automation (i.e. the Acrolinx CL checker) into Cochrane's current simplification approach could be beneficial in terms of authors' satisfaction. In particular, **we found that, on average, our participants were more satisfied with Acrolinx than they were with Cochrane PLS guidelines alone, and that they would welcome the introduction of the CL checker to complement the current non-automated simplification approach.** We described a possible scenario for the integration of these two tools — Cochrane PLS guidelines might be revised to only focus on content, and to be used exclusively when authors need to summarise an entire systematic review from scratch; on the other hand, with its focus on simple language and style, the Acrolinx CL checker, or similar, might be used for the subsequent simplification phase (i.e. to ensure that the summary is written in PL), either in Microsoft Word or in RevMan. With this integration, authors would only have to check guidelines dealing with content, while compliance with PL guidelines would be automatically and consistently checked. In particular, as shown by the responses of our participants regarding their typical workflow of PLS production, it is not uncommon for authors to check the PLS guidelines at the end of the writing task

— without disrupting this workflow, authors could therefore run an Acrolinx check at the end of the summarisation task. This scenario might reduce the overall effort and time-commitment required by PLS writing. It is worth mentioning that the introduction of (semi-)automation is already being considered for the different stages that lead to the production of a systematic review. In particular, Tsafnat et al. (2014) describe a workflow in which the manual/non-automated tasks requiring authors’ intuition and common sense (e.g. interpretation of the collected studies) are complemented by more automated tasks, such as machine learning-based screening of the abstracts of the studies to be included in the systematic review. The authors (*ibid.*, p. 3) define automation as “software that streamline processes by automating even only the trivial parts of a task”. This definition could also be applied to CL checkers.

At the very beginning of this thesis (Section 1.1), we discussed how lay users with different language background are increasingly turning to the Internet as a source of health content. It is therefore pivotal to develop a simplification/writing environment that maximises the satisfaction of authors and reduces their effort and commitment, so as to motivate them to contribute simplified/accessible online medical content. The increasing availability of accessible health-related texts might be beneficial not only for the (lay) users of the Cochrane Library website⁵⁹, but also for other websites that already disseminate or are considering disseminating Cochrane PLS, such as Epistemonikos and PubMed Health (Rada, Pérez and Capurro 2013; McIlwain and Tovey 2018).

There are, however, a few caveats to consider. For our authoring experiment, we used a version of Acrolinx that had not been tailored to Cochrane medical content. This occasionally resulted in the CL checker giving suggestions that were rejected by the authors. Should Acrolinx, or another CL checker, be integrated into a text simplification approach, it is likely that its tailoring might make the interaction with the tool smoother, and further increase satisfaction among authors. Tailoring CL checkers to authors’ preferred spelling might also have a beneficial effect on their interaction. Furthermore, as we observed, the introduction of any form of semi-automated solution should be

⁵⁹ Visits to cochrane.org increased from 5.7 million in 2015 to over 15 million in 2017 (*Cochrane Organisational Dashboard* 2017).

preceded by a training stage, when authors can practise and experiment with a new tool in their own time. As argued in Patel and Kaufman (2006, p. 134) in relation to the introduction of new software into health-related settings, “mastery of the system necessitates an individual and collective *learning curve* yielding incremental improvements in *performance* and *satisfaction*” (emphasis added). The cost of introducing technological assistance in a non-profit organisation like Cochrane should also be considered. Although limited in their functionalities, some websites provide freely available tools for PL writing (Section 4.9).

While the use of the Acrolinx CL checker, or similar tool, might facilitate the simplification task, as far as summarisation-related guidelines are concerned, the comments and rankings provided by our participants underlined the need to revise the set of Cochrane PLS guidelines in terms of characteristics considered, level of detail, and soundness. Furthermore, it might be beneficial to combine all the guidelines on summarisation/content into one document, which could then be made available to all authors regardless of their Cochrane Review Group.

Our investigation also underlined the importance of providing Cochrane volunteer authors with some form of guidance or support (whether non-automated or semi-automated) when writing PLS. More precisely, we found that most authors checked the guidelines for each PLS they wrote, and none of the authors reported that they would write a PLS without any form of support. These results are not surprising considering that Cochrane authors tend to have a health background, and are likely to be more familiar with using specialised medical language rather than with writing in PL. These characteristics of Cochrane authors might also explain why one participant welcomed the automatic feedback received by Acrolinx on readability issues as a way to develop their PL writing skills.

Our DV2 (i.e. the second usability component under investigation) was effectiveness or goal completion. This DV2 was further segmented into DV2.1 (readability), DV2.2 (comprehensibility), and DV2.3 (machine translatability).

The RQ associated with DV2.1 (readability) was the following:

RQ2.1: Does semi-automating a non-automated simplification approach by introducing a CL checker increase readability?

Our investigation showed that **the introduction of semi-automation** (i.e. the Acrolinx CL checker) into Cochrane's non-automated simplification approach **resulted in a statistically significant increase in the syntactic simplicity of Cochrane PLS, as well as in a statistically significant decrease in their word length and sentence length.** This result might be due to: (i) the higher specificity of Acrolinx rules dealing with, for example, the maximum number of words allowed in a sentence (as opposed to the vagueness which sometimes characterises Cochrane guidelines); (ii) the way in which suggestions are presented by Acrolinx (i.e. automatically and consistently flagged in the text, with no risk that authors would forget them); or (iii) a combination of both these aspects.

Regarding the other readability measures under analysis (namely, narrativity, word concreteness, referential/deep cohesion, and L2 readability), we found no significant impact of editing the PLS with Acrolinx. Therefore, the results pointed to the partial acceptance of our alternative hypothesis. In other words, introducing the Acrolinx CL checker into Cochrane's non-automated simplification approach made authors of PLS more effective in achieving readability, but only for a limited set of text characteristics. Regarding narrativity, word concreteness, and referential/deep cohesion, we even observed a slight, not significant decrease when the PLS were edited with Acrolinx. Despite being slight, this decrease sheds light on the fact that introducing some form of (semi-)automation might have unintended consequences on the texts, and that authors should be trained to take into account these unintended changes, e.g. by running multiple checks with a CL checker.

We also found that, regardless of being non-automated or semi-automated, simplification results in texts with higher levels of narrativity, referential cohesion, and L2 readability, as well as lower word and sentence length, compared with non-simplified texts (represented by Cochrane abstracts, in the case of our experiment). Collectively, these findings provide empirical evidence of the beneficial effects of both

manual and semi-automated simplification on text readability. As a further remark, we observed that, even though Cochrane authors had a medical background, they seemed able to intuitively increase some readability measures (e.g. deep/referential cohesion) in their PLS, compared with their abstracts. This finding underlines the need to train authors to also rely on their intuition (and common sense) of what makes a text readable, regardless of the simplification approach adopted (Section 5.7).

The RQ associated with our DV2.2 (comprehensibility) was:

RQ2.2: Does semi-automating a non-automated simplification approach by introducing a CL checker increase comprehensibility?

As explained in Section 5.2, an increase in readability does not automatically result in enhanced comprehensibility or accessibility, as the latter is mainly determined by the reader's characteristics. The quantitative data collected through our reading comprehension experiment showed that, despite having higher syntactic simplicity and lower word/sentence length than non-automated PLS, **the PLS edited with Acrolinx were not comprehended more** by lay readers who had different language backgrounds and had relatively low knowledge of the medical topics discussed in the texts. In addition, **compared with non-automated PLS, only a slightly higher number of native and non-native participants regarded semi-automated PLS as easy to read.** Therefore, the results indicated that introducing the Acrolinx CL checker into Cochrane's non-automated simplification approach did not make authors of PLS more effective in achieving the goal of comprehensibility, in line with our null hypothesis. One explanation for this result might be the fact that the usage of Acrolinx did not lead to an increase in cohesion, one of the text characteristics shown to influence comprehension (Smith et al. 2011). Previous studies had underlined the reduced comprehensibility of Cochrane PLS, even though these texts play a key role in Cochrane's accessibility mission (Sections 1.1 and 6.3) — should this organisation, or a similar one, consider the introduction of some form of technological assistance to increase the accessibility of their online PLS for lay readers, the tool should address a higher number of text characteristics (including characteristics at the discourse level,

such as cohesion). Furthermore, as we specified above in relation to readability findings, Cochrane authors seem able to intuitively identify and fix some cohesion gaps in the texts. They might therefore be trained to further harness their intuition, particularly to address cohesion issues at the macro level of the text (e.g. the presence of a topic sentence at the beginning of each paragraph), which are unlikely to be signalled by a CL checker or an authoring support tool.

By introducing the abstracts (i.e. non-simplified texts) as a baseline in our reading comprehension experiment, we expanded our results on comprehensibility. More precisely, we found that, compared with lack thereof, simplification can increase the perceived and actual accessibility of online medical texts, and more precisely the amount of information recalled from them, among both native and non-native speakers of English. However, we also observed that recall of specific sections was significantly higher in the case of abstracts, compared with PLS. Drawing upon previous studies (e.g. Kools et al. 2004; Lonsdale 2014; Kurtzman and Greene 2016), we explained this finding with the two following characteristics of the abstracts, compared with PLS, namely: (i) a more marked separation of sections through bold headings; and (ii) shorter sections. In terms of implications for Cochrane, this result shows the need to consider text formatting and segmentation — in addition to simple language — with a view to producing more accessible PLS. Again, a CL checker or an authoring support tool might facilitate the task of authors, for example, by reminding them to shorten a section or to use a different formatting when a string of text is marked as heading. Acrolinx, for example, allows users to configure the maximum number of sentences allowed in a paragraph — this functionality might in turn reduce the length of the section (Carter 2018).

From our experiment on comprehensibility, it also emerged, somewhat unsurprisingly, that reading skills of both native and non-native speakers of English have an impact on their comprehension of both simplified (i.e. PLS) and non-simplified texts (i.e. abstracts). In other words, regardless of text characteristics (e.g. use of a simple language, formatting, text segmentation), individuals with low reading skills are likely to show poorer comprehension, compared to individuals with high reading skills.

These findings caution against a one-size-fits-all approach to achieving accessible online health content, and underline the importance of tailoring the way in which health information is presented. As discussed in Section 6.9, lay users of the Cochrane website (as well as lay users of other websites disseminating health content) might benefit from the availability of different formats of communication of medical content (e.g. image and audio). In relation to the audio mode of communication, a study conducted by Maguire and Clarke (2014) showed that, compared with written summaries, audio summaries (i.e. podcasts) of Cochrane Systematic Reviews can facilitate understanding and identification of key messages in the text. Moreover, Houts et al. (2006) and Meppelink et al. (2015) found that images that accompany simplified medical texts are particularly beneficial when the recipients have limited reading skills.

The final result emerging from our comprehensibility experiment was that the amount of English content recalled/comprehended by non-native speakers of English was consistently lower than amount of information recalled/comprehended by native English speakers. This finding indicates that, to avoid putting Internet users with a L1 different from English at a disadvantage when searching for accessible online health content, translation might be needed.

The RQ associated with our DV2.3 (machine translatability) was:

RQ2.3: Does semi-automating a non-automated simplification approach by introducing a CL checker increase machine translatability?

Our machine translatability experiment — conducted with human evaluators who were health domain experts and native speakers of Spanish — showed that **the introduction of semi-automation (i.e. the Acrolinx CL checker) into Cochrane’s non-automated simplification approach did not result in an increase in the machine translatability of PLS from English into Spanish**. In other words, the effectiveness of Cochrane authors in terms of the level of machine translatability achieved was not enhanced by the use of Acrolinx, in line with our null hypothesis. This result was unexpected considering that semi-automated PLS had significantly lower sentence length than non-automated PLS, and that sentence length has been shown to impact the quality of MT output

obtained from neural MT systems (Castilho et al. 2018b). A possible explanation might be the fact that, on average, sentences in non-automated PLS were already quite short (Section 7.9). Accordingly, despite being significant, the impact of Acrolinx on sentence length and, in turn, on machine translatability might have not been substantial.

By analysing the quantitative and qualitative data collected from our Spanish-speaking health domain experts, we also found that, regardless of the simplification approach adopted, the quality of the Spanish MT outputs produced by the neural system Google Translate was relatively high in terms of adequacy and fluency. Furthermore, even though the style/language of the MT output was often described as unnatural, the information in the source English PLS was often translated fully and accurately into Spanish. Collectively, these findings seem promising for Cochrane since this organisation is increasingly relying on MT to reduce the workload of their volunteer translators/health domain experts, and to encourage their involvement with the translation tasks — the benefits (in terms of costs and time) of post-editing health-related texts rather than translating them from scratch have already been shown (Kirchhoff et al. 2011; Turner et al. 2014). As specified in Section 7.3, PLS are often translated as part of Cochrane’s strategy to make online medical evidence accessible to lay readers with no or limited knowledge of English. The importance attributed to translation is also visible in the recently redesigned Cochrane Library website, where Internet users can conduct searches in different languages (Anthony 2018).

Similar to native speakers of English with limited reading skills and no medical background, lay readers whose L1 is not English might represent a vulnerable group (depending on their knowledge of English) when searching for health content online. The availability of translated content can partially address their needs as the advantages (in terms of comprehension) of receiving healthcare information in one’s L1 have already been observed (Todd and Hoffman-Goetz 2011; O’Brien and Cadwell 2017). However, as in the case of simplification, translation should not be used as a one-size-fits-all approach. In other words, Internet users with low reading skills are likely to encounter comprehension issues even when content is presented in their L1. Therefore, to maximise accessibility of medical content on their websites, Cochrane and other

organisations might present translated information in different formats (such as, again, with image or audio). In relation to this point, it is noteworthy that Cochrane is also translating its podcasts (*Translated Cochrane Evidence* 2018).

In summary, **we empirically showed that semi-automating a non-automated simplification approach by introducing a CL checker increased the satisfaction component of usability and, only partially, its effectiveness component — the use of Acrolinx was beneficial for some readability measures, but there was no significant impact of this tool on comprehensibility and machine translatability.**

8.2.1 Practical Implications

As the discussion in this section has shown, our investigation has several practical implications for Cochrane and other organisations sharing similar missions. In other words, our findings could be used to support and build on the work that these organisations are already carrying out to make online health content available and accessible to Internet users with no medical background and with different reading skills and native languages. In the interests of clarity, we summarise and list here the main recommendations emerging from our findings for a satisfactory production and effective (multilingual) dissemination of health content:

- *Provide volunteer authors of simplified medical content with some form of technological assistance that can reduce their workload and further develop their writing skills:*
 - *Any form of technological assistance would need to be tailored and introduced gradually into the existing workflow;*
- *For authoring tasks that cannot be (semi-)automated, ask volunteer authors for feedback on the resources provided to them;*
- *Provide volunteer authors of simplified medical content with some form of technological assistance that reduces vagueness in the simplification guidelines and ensures their automatic and consistent application:*

- *Any form of technological assistance would need to account for a broad range of text characteristics (including characteristics at the discourse level);*
- *Train volunteer authors to combine technological assistance with their intuition of what makes medical texts accessible;*
- *In addition to focusing on simple language, consider text segmentation and formatting:*
 - *Having technological assistance might help authors consider text segmentation and formatting;*
- *Do not focus the accessibility strategy on text format only — instead, make health content available online also in audio and visual formats;*
- *Translate health content into readers' L1;*
- *Introduce MT (followed by PE/validation conducted by volunteer health domain experts) into the translation workflow.*

8.3 Contributions of the Thesis

This thesis contributes to knowledge in three ways: empirically, methodologically, and in relation to practice. Along each of these three dimensions, this thesis supports and develops existing knowledge, while also providing new knowledge on the usability of text simplification. These contributions are summarised in Table 8.1 — which has been adapted from Farndale (2004) — and discussed in detail thereafter.

Domains of contribution	What has been supported?	What has been developed?	What is new?
<i>Empirical evidence</i>	<ul style="list-style-type: none"> * Different findings from subjective and objective components of usability * Evidence of the benefits of text simplification in terms of readability * Evidence of the benefits of text simplification in terms of comprehension of entire texts * Evidence that text formatting and segmentation can influence recall of specific sections * Evidence that reading skills can influence comprehension * Evidence of the benefits of receiving health content in one's L1 	<ul style="list-style-type: none"> * Testing the impact (in terms of usability) of introducing semi-automation in a non-automated simplification approach for health-related texts 	<ul style="list-style-type: none"> * Evidence that authors of health content would welcome the introduction of technological support as a CL checker * Evidence that authors of health content are effective at increasing text readability at the word and syntax/sentence level when using a CL checker * Evidence that authors of health content are not effective at increasing text comprehensibility and machine translatability when using a CL checker * Evidence that Google Translate (as a neural MT system) produces output of relatively high quality with health-related texts, regardless of the simplification approach, for the English-Spanish translation direction * Evidence that the output of Google Translate (as a neural MT system) was rated higher in terms of adequacy than in terms of fluency, regardless of the simplification approach, for the English-Spanish translation direction
<i>Method</i>	<ul style="list-style-type: none"> * Advantages and limitations of the SUS 	<ul style="list-style-type: none"> * Adaptation of the SUS to PL guidelines 	<ul style="list-style-type: none"> * Strategies for the adaptation of the SUS

	<ul style="list-style-type: none"> * Advantages of using Coh-Metrix measures * Advantages and limitations of text-retelling * Advantages and limitations of human evaluation of MT output * Benefits of complementing quantitative data with qualitative data 	<ul style="list-style-type: none"> * Application of Coh-Metrix to Cochrane PLS and abstracts * Adoption of immediate text-retelling for Cochrane PLS and abstracts * Application of fluency and adequacy measures to machine translated Cochrane PLS 	<ul style="list-style-type: none"> to PL guidelines * Procedure for segmentation and scoring of recall protocols * Learning point on the importance of assessing the fluency of machine translated Cochrane PLS
<i>Knowledge of practice</i>	<ul style="list-style-type: none"> * Use of RevMan to produce Cochrane PLS * Variability in the guidelines provided to Cochrane authors for the production of PLS * Infrequent production of PLS at Cochrane (per author) * Different opinions of authors depending on the set of PLS guidelines 	<ul style="list-style-type: none"> * Possible workflow for integrating a CL checker and Cochrane PLS guidelines 	<ul style="list-style-type: none"> * Vagueness and contradictions in Cochrane PLS guidelines on language/style * Consultation of PLS guidelines mostly before or after the authoring task * Widespread reliance on Cochrane PLS guidelines * Variability in Cochrane authors' opinions on the completeness of Cochrane PLS guidance

Table 8.1: Contributions of this thesis

8.3.1 Empirical Contribution

This thesis provides empirical evidence to support the view of usability as a broad and multifaceted concept, including subjective (namely, satisfaction) and objective components (i.e. effectiveness, and efficiency) that are relatively independent from one another, and either not correlated or weakly correlated (Frøkjær, Hertzum and Hornæk 2000; Brooke 2013). This investigation also supports the body of knowledge on: (i) the benefits of simplification, compared with lack thereof, in terms of readability and comprehension (Wilson and Wolf 2009; Eltorai et al. 2015; Meppelink et al. 2015); (ii) the importance of considering text formatting and segmentation to enhance

recall/comprehension (Kools et al. 2004; Lonsdale 2014; Kurtzman and Greene 2016); (iii) the impact of reading skills on comprehension (Ozuru, Dempsey and McNamara 2009); and (iv) the advantages in terms of comprehension of receiving health information in one's native language (Dew et al. 2015; O'Brien and Cadwell 2017). Collectively, this body of evidence can inform and support best practices for the effective tailoring and dissemination of health content. These best practices, which have been outlined in Section 8.2.1, can be relevant for organisations whose goal is to provide online health information that is accessible for lay readers with different skills and language backgrounds, such as the CDC, the Campbell Collaboration, the PAHO, and the HSE (Sections 3.3 and 7.2). For example, the HSE has developed the Under the Weather⁶⁰ website, whose goal is to educate the public on the appropriate use of antibiotics by means of readable and user-friendly content.

In addition to providing the aforementioned supporting empirical evidence, this thesis also develops and contributes new empirical evidence. To the best of our knowledge, this is the first investigation on the usability of a manual text simplification approach for health content, and on the impact of introducing a CL checker in the aforementioned approach. The new empirical evidence emerging from this thesis is summarised here:

1. While the need to provide health domain experts with support during PL writing has been acknowledged (Smith et al. 2011), this thesis represents the first empirical study on: (i) the satisfaction of health domain experts with sets of PL guidelines and a CL checker; and (ii) their authoring preferences and reasons behind their preferences. Concretely, our findings have shown that authors would welcome the introduction of semi-automation in the authoring workflow (Section 4.9). In addition to having implications for authors at Cochrane (Section 8.2), these findings could benefit other online environments relying on volunteer authors and editors, such as Wikipedia and Simple English Wikipedia. For example, similar to Cochrane authors, Simple English Wikipedia editors are asked to read and manually apply a series of guidelines, which correspond to a modified version of Ogden's Basic English (Schwitter 2015). These

⁶⁰ The Under the Weather website is available at: <https://bit.ly/1qRy5xz> [Accessed 12 December 2018].

guidelines are spread across different webpages, which is likely to make their checking and manual implementation difficult and time-consuming⁶¹. Moreover, authors/editors do not receive automatic feedback on the quality (e.g. readability and translatability) of their texts. Therefore, manual simplification is unlikely to represent a learning experience in PL/CL writing. All these factors might discourage potential volunteers from getting involved (Section 4.2);

2. This thesis also represents the first empirical study on the effectiveness of health domain experts as PL/CL writers before and after the introduction of semi-automation in the form of a CL checker. We observed that: (i) when presented with a CL checker, authors mainly implemented changes at the word and syntax/sentence level (Section 5.7); and (ii) by implementing these edits, authors were not effective in increasing the comprehensibility and machine translatability of texts (Sections 6.9 and 7.9). In the case of comprehensibility, the observed ineffectiveness of the edits made by the authors with the CL checker might be due to the fact that they did not consistently address cohesion, a text characteristic shown to have an impact on comprehension/recall (Sections 6.2 and 6.7.1). Therefore, our evidence points to the need to develop and test CL checkers that address a broader range of issues at the discourse level (including cohesion) (Section 6.9);

3. In relation to the effectiveness of health domain experts in terms of machine translatability achieved, to the best of our knowledge, this thesis represents the first empirical investigation on the impact of text simplification on the machine translatability of online health-related texts with a neural MT system (Sections 7.2 and 7.3). Even though machine translatability did not increase as a result of Acrolinx, our additional findings on the translation direction English-Spanish have shown that: (i) both before and after the introduction of the CL checker, the MT quality obtained with Google Translate was relatively high; and (ii) both before and after the introduction of the CL checker, the adequacy/content of the MT outputs was rated higher than their fluency/style (Section 7.9). Collectively, these findings show that, regardless of being

⁶¹ The guidelines on Basic English for Simple English Wikipedia editors are available at: <https://bit.ly/2t4ayS1> [Accessed 12 December 2018].

non-automated or semi-automated, the combination of text simplification and freely available neural MT systems (followed by PE/validation) can represent a viable alternative to human translation of online health content, thus reducing the time and effort of the volunteers involved in the translation tasks. In addition to Cochrane, these findings are encouraging for other organisations relying on volunteers and MT for the dissemination of multilingual content — such as Translators without Borders (Section 7.7.1) — and underline the benefits of editing the source texts with a view to simplifying them prior to MT adoption.

8.3.2 Methodological Contribution

Each of the four experimental chapters in this thesis (Chapters 4-7) contributes to methodological knowledge. Below we explain how they support existing knowledge, and how they develop new knowledge of methods:

Satisfaction study (Chapter 4). In line with the studies discussed in Sections 4.7.1 and 4.7.2, this thesis confirms the advantages and limitations of adopting the SUS to measure satisfaction. Regarding the advantages, we showed that the SUS: (i) is a technology-agnostic instrument that can be used for products/systems as diverse as a set of PL guidelines and a CL checker; (ii) requires a short time commitment to fill out; and (iii) provides an overall single score that can be easily interpreted, particularly if associated with an adjective descriptor (Bangor, Kortum and Miller 2008). With regard to the limitations of the SUS, we observed that this questionnaire is unable to provide diagnostic information on the issues that participants might encounter when using a product or system, and that follow-up questions are therefore required to complement the quantitative data (Sections 4.8.3 and 4.8.4). In addition to supporting methodological knowledge emerging from previous studies, this thesis represents the first attempt to use the SUS on a set of PL guidelines⁶². The novelty of this use of the SUS led us to consider strategies for its tailoring to a non-automated simplification approach. In particular, in statement 5 of the SUS, we replaced the term *functions* with *documents* (Section 4.7.2). A similar edit to the wording of the SUS might be implemented in future

⁶² On the other hand, as explained in Section 4.7.1, the SUS had already been used on a CL checker (Miyata et al. 2017).

studies aiming at testing the satisfaction of authors who are asked to follow written guidelines;

Readability study (Chapter 5). This thesis supports existing methodological knowledge on the advantages of using Coh-Metrix rather than traditional readability formulas, which only account for shallow text characteristics (Section 5.5.1). Concretely, we showed that simplified and non-simplified texts differ along a variety of dimensions — such as narrativity and cohesion — that traditional readability formulas are unable to capture. Furthermore, as far as we are aware, this thesis represents the first application of Coh-Metrix to Cochrane PLS and abstracts, whose readability has traditionally been measured through traditional formulas (Section 5.3). Considering the ever growing interest in the readability of health-related texts in general (Section 5.2), and of Cochrane texts in particular (Section 5.3), this thesis shows the viability of using Coh-Metrix to gain a broader picture of the text characteristics that might influence comprehension;

Comprehensibility study (Chapter 6). This thesis supports existing methodological knowledge on the advantages and limitations of using text-retelling to measure reading comprehension (Section 6.7.1). With regard to the advantages, we showed how text-retelling: (i) allows for question formats that are equivalent across texts, thus ensuring the comparability of recall scores (Appendix K); (ii) allows for the avoidance of clues in the questions; and (iii) does not allow participants to guess at their answers (Crossley and McNamara 2016). Regarding the limitations of text-retelling, we showed how a within-subjects design is needed to prevent readers' differences in writing skills (in L2) from biasing the results (Section 6.7.1). As far as new methodological knowledge is concerned, this thesis describes the first adoption of an *immediate* text-retelling task to assess comprehension of Cochrane PLS and abstracts — to the best of our knowledge, previous studies have either adopted delayed text-retelling or alternative methods such as multiple-choice testing (Section 6.3). Moreover, this thesis describes a detailed procedure for the segmentation and scoring of the recall protocols obtained from text-retelling — this procedure can be replicated in other reading comprehension studies (Section 6.7.2);

Machine translatability study (Chapter 7). This thesis supports the widely recognised advantages and limitations of human evaluation of MT output (Section 7.7.2), particularly with health domain experts (Section 7.7.3). For example, regarding the advantages, we discussed how, differently from AEMs, human evaluation allows researchers to identify potential content inaccuracies (Way 2018), which are particularly detrimental in the health domain. Regarding the limitations of human evaluation, we observed the impact of the evaluators' subjectivity, particularly when they do not have a linguistics background, nor shared training or experience in MT evaluation (Section 7.9). In addition to supporting already existing methodological knowledge, this thesis describes the first application of both adequacy and fluency measures to Cochrane content machine translated with a neural MT system. As part of the HimL project (Sections 7.2 and 7.3), Birch et al. (2016) adopted HUME (a measure that reflects adequacy) to evaluate the machine translatability of Cochrane PLS and abstracts. However, the authors (ibid.) did not focus on fluency. This thesis shows that, even though fluency/style plays a secondary role when it comes to health content, by including the fluency measure it is possible to gain a broader understanding of the types of issues that health domain experts (at Cochrane) might be asked to fix when post-editing an MT output. In turn, this broader understanding might help explain their preferences for human translation over PE, or vice versa.

As a final remark on the methodological contributions of this thesis, the analysis that followed the experiment on authors' satisfaction (Sections 4.8.3 and 4.8.4) and the experiment on machine translatability of PLS (Section 7.8.2) supports the idea that combining quantitative and qualitative data allows researchers to enhance the credibility of their conclusions (Frey et al. 1991). Creswell and Plano Clark (2018, p. 8) underline the benefits of combining quantitative and qualitative data by arguing that “[o]ne type of evidence may not tell the complete story”. For example, in Section 7.8.2, we showed how the comments provided by the participants after the MT evaluation tasks fully or partially supported the scores that they had assigned.

8.3.3 Contribution to Knowledge of Practice

Chapter 4 of this thesis both supports existing knowledge of the authoring/simplification workflow at Cochrane, and provides new knowledge of the aforementioned workflow. Specifically, through the questionnaires on Cochrane authors' background characteristics and typical interaction with PLS guidelines (Sections 4.8.1 and 4.8.2), we observed that, along with Microsoft Word, RevMan is an important part of the Cochrane Review ecosystem, as is also evidenced by the training provided to authors and by the consistent updates of the software (*Introduction to RevMan Web* 2018). In line with Glenton (2017), we also observed that: (i) not all authors of PLS are provided with the same guidelines; (ii) each author tends to produce a low number of Cochrane Systematic Reviews (and corresponding PLS); and (iii) not all sets of PLS guidelines might be regarded as equally useful by the authors.

In addition to supporting existing knowledge of the authoring/simplification practice at Cochrane, this thesis broadens its understanding. Concretely, to the best of our knowledge, this thesis provides the first analysis of the vagueness and contradictions that sometimes characterise Cochrane PLS guidelines dealing with simplified language and style (Section 4.3), and the first survey of the varied opinions of authors on the completeness of the guidelines dealing with both content and style (Section 4.8.2). As far as we are aware, this thesis also represents the first study of the stage at which Cochrane authors check the guidelines when authoring PLS, and of the frequency with which they consult them. Specifically, we observed that: (i) most authors check the guidelines either at the beginning or at the end of the authoring task; and (ii) most authors need to consult the guidance for each PLS they write (Section 4.8.2). Collectively, these results underline the need for assistance in PL writing, and can inform the design of alternative authoring workflows. In the case of our study, we explained how the Acrolinx CL checker and the sets of Cochrane PLS guidelines dealing with content/summarisation could be integrated (Section 8.2). With this alternative workflow, authors would not have to check language/style-related guidelines either before starting a PLS or after finishing it, as they could just run a check with Acrolinx once the PLS has been written. Other non-profit organisations which produce

PLS of systematic reviews might benefit from a study of how volunteer authors use the guidelines that they are given, and of potential difficulties that they might encounter in applying them. An example is the Campbell Collaboration, whose PLS guidelines deal with both simplification and summarisation, and are implemented with a non-automated approach similar to the one adopted at Cochrane (Campbell Collaboration 2016). Volunteers at the Campbell Collaboration might therefore also welcome the introduction of technological assistance while authoring.

8.4 Limitations of the Thesis and Future Research

The findings reported in Section 8.2 and the contributions outlined in Section 8.3 need to be understood within the limitations of this thesis. Below we delve into these limitations and explain how future research might address them.

As far as our satisfaction experiment is concerned, despite trying different recruitment techniques, it was not possible to recruit a large and homogenous sample of participants (Section 4.4). Accordingly, our sample of authors was reduced and varied in terms of native language, experience in PL writing, and time of production of latest PLS. Our findings would need to be confirmed with future research involving a larger and more homogeneous group of Cochrane authors. With a larger sample of participants, it might be also possible to compare groups based on their background characteristics. It might also be necessary to only recruit authors who have produced a PLS with the non-automated approach just a few days before the experiment, as they might provide more reliable answers on their level of satisfaction with Cochrane PLS guidelines. Furthermore, when asking Cochrane authors about their interaction with Cochrane PLS guidelines, we could not distinguish between summarisation- and simplification-oriented guidelines since these appear in the same documents and are not clearly distinguished (Section 4.5).

Another limitation of our satisfaction experiment is linked with the CL checker used. More precisely, even though Acrolinx rules contradicting Cochrane PLS guidelines had been deactivated, this CL checker had not been tailored to Cochrane content or to authors' spelling preferences (Section 4.3). Future research might examine

whether authors' satisfaction would be higher with a more tailored tool. In addition, considering that the Acrolinx CL checker is a commercial tool, it might be worth investigating authors' interaction and satisfaction with resources for text simplification that are freely available online (Section 4.9).

We collected data on authors' interaction with the simplification approaches exclusively by means of questionnaires. Future work might expand on our research by also adopting ethnographic methods of observation in real-world settings, or recordings of participants' interaction with a tool, e.g. by means of eye tracking, screen recording, or keystroke logging (Risku, Milošević and Pein-Weber 2016; Teixeira and O'Brien 2017; Olohan 2018). Methods that are traditionally adopted for the study of the interlingual translation process could therefore be applied to the study of text simplification as an intralingual translation process (Kajzer-Wietrzny, Whyatt and Stachowiak 2016).

As a final remark on our experiment with Cochrane authors, we assumed that developing an environment that reduces their effort and time-commitment, while also providing automatic feedback on their PL writing, could motivate them to continue contributing simplified online health content. While previous works have shown that this assumption might prove correct in some cases (Section 4.2), it should also be noted that motivation is a complex psychological construct — in Section 4.2, we presented the six motivational categories in Clary et al. (1998). Previous studies have also shown that volunteers involved in online communities can be self-motivated to write in the first place (Joyce and Kraut 2006), or that motivation to contribute is determined by the perceived importance and uniqueness of the contributions (Ling et al. 2005). Qualitative data collected from Cochrane authors (e.g. through structured or semi-structured interviews) might shed light on the factors that motivate them.

With regard to our readability experiment, the main limitation was the small number of texts per corpus (Section 5.4). For our findings to be generalisable, the statistical tests described in Section 5.6 should be repeated by using larger corpora of texts. Moreover, the limited impact of the Acrolinx CL checker on readability might be due to the fact that the texts we used had already been written in accordance with PL

guidelines. Future work would need to test the impact of this or other CL checkers on non-simplified texts (such as Cochrane abstracts). Even though Cochrane abstracts target health professionals (rather than lay readers), it would be beneficial to also check the readability of these texts since studies have shown that abstracts might be difficult to comprehend even for health domain experts (Zhelev, Garside and Hyde 2013).

As far as the experiment on comprehensibility is concerned, one limitation was the lack of additional raters for the segmentation and scoring of recall protocols. Despite adopting several measures to reduce the influence of the researcher's subjectivity (Section 6.7.2), calculating inter-rater agreement would have allowed us to test the reliability of our procedure (Hallgren 2012). Future work could therefore apply the same procedure for the segmentation and scoring of recall protocols, while also ensuring that agreement between different raters is reached. Additionally, regarding the comparison of recall scores between native and non-native speakers, the latter group should also be tested in terms of their L2 writing skills, as these could influence the accuracy and completeness of the recall protocols produced (Section 6.7.1). Another limitation of the comprehensibility experiment was the relatively small sample of non-native speakers of English involved, which might have had an impact on the results of the statistical tests (Section 6.8.2). Our findings should be confirmed with studies involving a larger sample of L2 readers.

Moreover, as part of the comprehensibility study, we tested increase or decrease in recall across the three corpora of abstracts, non-automated PLS and semi-automated PLS (Section 6.8.2), while not examining whether the amount of information recalled by lay readers would result in them being able/willing to apply the acquired knowledge in real life. This type of investigation could involve recruiting lay readers who are at risk of being affected by an illness, presenting them with a simplified text on the effectiveness of a recommended preventive action, and then running focus groups to discuss their compliance with the preventive action, or lack thereof (Lee 2000). It should also be pointed out that comprehension is only the first step towards patients' adherence to healthcare interventions (Smith et al. 2011; Grene, Cleary and Marcus-Quinn 2017) — characteristics of the target audience (such as levels of self-efficacy, risk perceptions,

and cultural backgrounds) all determine the practical implications that a health-related message can have, and should therefore be taken into account (Lee 2000; Weinstein et al. 2007; Ryan et al. 2008). Interestingly, these characteristics of the target audience have also been shown to influence the effectiveness of communication in crises and disasters (see e.g. Kammerbauer and Minnery 2018), which is the focus of the INTERACT project (Section 1.3).

Similar to the study reported in Meppelink et al. (2015), it might be worth testing whether accompanying PLS with images would positively influence recall. Furthermore, while we focused on PLS and abstracts, research on comprehension using recall could be expanded to also include podcasts or another summary format developed at Cochrane, namely blogshots (*Cochrane Blogshots Used Globally* 2018). As a final remark on the comprehensibility experiment, we presented our participants with a short background questionnaire so as not to overload them since all our tasks were being conducted in one session (Section 6.5). In other words, we had to prioritise the most relevant questions. In future experiments, it might be worth asking participants to conduct the tasks on different days, as this would give researchers the possibility to expand each task. In particular, when completing the background questionnaire, participants might also be asked to take one of the standardised tests on health literacy, such as the Test of Functional Health Literacy in Adults (Parker et al. 1995).

With regard to the machine translatability experiment, this was influenced by the same limitation identified for the readability experiment, namely a limited number of texts per corpus (Section 7.6). Our findings would need to be confirmed with studies using larger corpora of English PLS and corresponding MT outputs. It might also be necessary to conduct the same MT evaluation at the document level, rather than the sentence level (Section 7.5). In relation to this point, Läubli, Sennrich and Volk (2018) found that document-level MT evaluation allows for the identification of errors and inaccuracies (e.g. in terms of cohesion) that are difficult to recognise when evaluation is conducted at the sentence level. Furthermore, repeating the same experiment after training participants on MT evaluation might at least partially increase inter-rater agreement.

There are other areas for future research. To give a few examples, it might be interesting: to assess the quality of the post-edited MT output in terms of its comprehensibility and/or acceptability among end users/readers (Castilho et al. 2018a); to combine simplification of medical texts with domain-adapted MT or with human translation; to complement adequacy and fluency measures with error typology (Birch et al. 2016); and to examine the machine translatability of Cochrane PLS (or other English health-related texts) when translated into other languages, including low-resourced languages. In relation to this point, Dew et al. (2018) point out that neural MT systems seem to perform worse when training data is scarce. Another aspect to consider in relation to the adoption of MT and PE is that, even when the source text has been controlled/simplified, the task of PE might reintroduce complexity in the text as post-editors “embellish” the MT output in addition to correcting it (Nyberg, Mitamura and Huijsen 2003, p. 274). Interestingly, the same issue is discussed in Warde et al. (2018, p. e55) in relation to human translation: “inadvertent revision away from plain language can occur during the process of translation by professional medical translators”. Future research should therefore examine the impact of involving health domain experts in the PE/translation tasks on the readability and comprehensibility of the translated texts.

In this thesis, we examined the usability of a text simplification approach. Future work might focus on text summarisation instead. For example, Cochrane authors might be asked questions on their satisfaction with the guidelines dealing with the content of PLS (Section 1.1). Similar to simplification, summarisation of content plays an important role on the Internet, where users are presented with a massive amount of information that they need to filter (Bayomi et al. 2016). Moreover, the (source or translated) texts themselves might be the objects of a usability investigation (see e.g. Castilho 2016). While the need to address the usability of health-related texts has been recognised (Kools 2007), to the best of our knowledge, medical texts have not been analysed along the usability dimensions of satisfaction, effectiveness, and efficiency. Future research might therefore investigate the time and mental effort required to read medical texts (efficiency), their ability to lead to behavioural change (effectiveness), and the recipients’ physical, cognitive and emotional responses to the texts (satisfaction).

Further attention should also be devoted to a potential trade-off between the user experience⁶³ of the readers of the source texts and the user experience of the readers of the corresponding (machine) translated texts. For example, while potentially beneficial in terms of comprehension, edits at the discourse level might not have an impact on the characteristics of the MT outputs obtained with neural MT systems (Bawden et al. 2018). Bowker (2015) found that, when translatability-oriented rules were applied to texts extracted from a university website and then machine translated, the user experience of the target-language readers increased, while the user experience of the source-language readers decreased. As specified in Section 4.3, the Acrolinx CL checker in our experiment presented authors with both readability- and translatability-oriented rules. Future research should therefore isolate the impact of each of these rule categories on the usability/user experience of medical texts.

Finally, despite examining the accessibility of health content on the online Cochrane Library, this thesis merely focused on the comprehensibility of this content, and did not consider other factors that contribute to web accessibility, namely: (i) users' ability to perceive the information and the user interface through one of their sense (e.g. by means of text alternatives for blind users); (ii) users' ability to navigate and operate the interface (e.g. for people who do not use a mouse, a keyboard must provide access to all functionalities); and (iii) users' ability to access content robust enough to be compatible with various user agents, including different browsers and assistive technologies (Rodríguez Vázquez 2016; *Accessibility Principles* 2017). All these different aspects of web accessibility deserve further investigation. To quote Yesilada et al. (2012, npn): “[a]ccess is what the web is ‘about’, it is the motivation behind its creation”.

⁶³ For an explanation of the differences and similarities between *usability* and *user experience*, see Section 2.2.

REFERENCES

A

Abdi, H. and Williams, L. 2010. Tukey's Honestly Significant Difference (HSD) test. *IN: Salkind, N. (ed.) Encyclopedia of Research Design*. Thousand Oaks, California: Sage.

About Cochrane 2018. Available at: <https://bit.ly/2E4SW1r> [Accessed 12 December 2018].

About Translation at Cochrane 2018. Available at: <https://bit.ly/2BagXDF> [Accessed 12 December 2018].

About Us 2018. Available at: <https://bit.ly/2E6AnKD> [Accessed 12 December 2018].

Abrahamsson, E., Forni, T., Skeppstedt, M. and Kvist, M. 2014. Medical text simplification using synonym replacement: Adapting assessment of word difficulty to a compounding language. *IN: Williams, S., Siddharthan, A. and Nenkova, A. (eds.) Proceedings of the Third Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)*. Gothenburg, Sweden, 27 April. Stroudsburg, Pennsylvania: Association for Computational Linguistics, pp. 57-65.

Accessibility Principles 2017. Available at: <https://bit.ly/2Eo0ixn> [Accessed 12 December 2018].

Acrolinx 2012. *Acrolinx Donates Software to Translators without Borders*. Available at: <https://bit.ly/2Du0wTe> [Accessed 12 December 2018].

Adams, P. and Fleck, F. 2015. Bridging the language divide in health. *Bulletin of the World Health Organization*, 93(6), pp. 365-366.

Adnan, M., Warren, J. and Orr, M. 2010. Assessing text characteristics of electronic discharge summaries and their implications for patient readability. *IN: Maeder, A. and Hansen, D. (eds.) Proceedings of the Fourth Australasian Workshop on Health Informatics and Knowledge Management (HIKM 2010) - Volume 108*. Brisbane, Australia, 1 January. Darlinghurst, Australia: Australian Computer Society, Inc., pp. 77-84.

Aikawa, T., Schwartz, L., King, R., Corston-Oliver, M. and Lozano, M. 2007. Impact of controlled language on translation quality and post-editing in a statistical machine translation environment. *IN: Maegaard, B. (ed.) Proceedings of the Machine Translation Summit XI*. Copenhagen, Denmark, 10-14 September. European Association for Machine Translation, pp. 1-7.

Alderdice, F., McNeill, J., Lasserson, T., Beller, E., Carroll, M., Hundley, V., Sunderland, J., Devane, D., Noyes, J., Key, S. and Norris, S. 2016. Do Cochrane summaries help student midwives understand the findings of Cochrane systematic

reviews: The BRIEF randomised trial. *Systematic Reviews*, 5(40). doi:10.1186/s13643-016-0214-8.

Alderson, J. 2000. *Assessing Reading*. Cambridge: Cambridge University Press.

Alexander, P., Kulikowich, J. and Schulze, S. 1994. The influence of topic knowledge, domain knowledge, and interest on the comprehension of scientific exposition. *Learning and Individual Differences*, 4, pp. 379-397.

Allen, J. H. 2003. Post-editing. IN: Somers, H. (ed.) *Computers and Translation: A Translator's Guide*. Amsterdam: John Benjamins, pp. 297-317.

Aluísio, S. M., Specia, L., Pardo, T. A., Maziero, E. G. and Fortes, R. P. 2008. Towards Brazilian Portuguese automatic text simplification systems. IN: *Proceedings of the Eighth ACM Symposium on Document Engineering (DocEng 2008)*. São Paulo, Brazil, 16-19 September. New York: Association for Computing Machinery, pp. 240-248.

Anderson, R. C. and Pichert, J. W. 1977. Recall of previously unrecallable information following a shift in perspective. *Journal of Verbal Learning and Verbal Behavior*, 17(1), pp. 1-12.

Anthony, J. 2018. *Cochrane Library: An Improved Online Platform to Guide Health Decision-Making across the World*. Available at: <https://bit.ly/2M4n7KN> [Accessed 12 December 2018].

ASU Admission 2018. Available at: <https://bit.ly/2DwCvee> [Accessed 12 December 2018].

Aymerich, J. and Camelo, H. 2009. The machine translation maturity model at PAHO. IN: *Proceedings of the Machine Translation Summit XII*. Ottawa, Canada, 26-30 August. Available at: <https://bit.ly/2MRH6Im> [Accessed 12 December 2018].

Azzam, A., Bresler, D., Leon, A., Maggio, L., Whitaker, E., Heilman, J., Orlovitz, J., Swisher, V., Rasberry, L., Otoide, K., Trotter, F., Ross, W. and McCue, J. D. 2017. Why medical schools should embrace Wikipedia: Final-year medical student contributions to Wikipedia articles for academic credit at one school. *Academic Medicine*, 92(2), pp. 194-200.

B

Badarudeen, S. and Sabharwal, S., 2010. Assessing readability of patient education materials: Current role in orthopaedics. *Clinical Orthopaedics and Related Research*, 468(10), pp. 2572-2580.

Baker, C. L. 1985. *English Syntax* (2nd ed.). London and Cambridge (Massachusetts): The MIT Press.

- Bangor, A., Kortum, P. T. and Miller, J. 2008. An empirical evaluation of the System Usability Scale. *International Journal of Human-Computer Interaction*, 24(6), pp. 574-594.
- Barrett, T. J. 2011. Computations using analysis of covariance. *Wiley Interdisciplinary Reviews: Computational Statistics*, 3(3), pp. 260-268.
- Basch, C. H., MacLean, S. A., Romero, R. A. and Ethan, D. 2018. Health information seeking behavior among college students. *Journal of Community Health*, 43, pp. 1094-1099.
- Bawden, R., Sennrich, R., Birch, A. and Haddow, B. 2018. Evaluating discourse phenomena in neural machine translation. *IN: Proceedings of the 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT)*. New Orleans, Louisiana, 1-6 June. Stroudsburg, Pennsylvania: Association for Computational Linguistics, pp. 1304-1313.
- Bayomi, M., Levacher, K., Ghorab, M. R., Lavin, P., O'Connor, A. and Lawless, S. 2016. Towards evaluating the impact of anaphora resolution on text summarisation from a human perspective. *IN: Métais, E., Meziane, F., Saraee, M., Sugumaran, V. and Vadera, S. (eds.) Natural Language Processing and Information Systems*. Basel, Switzerland: Springer International Publishing, pp. 187-199.
- Baytiyeh, H. and Pfaffman, J. 2010. Volunteers in Wikipedia: Why the community matters. *Educational Technology & Society*, 13(2), pp. 128-140.
- Beck, I. L., McKeown, M. G., Sinatra, G. M. and Loxterman, J. A. 1991. Revising social studies text from a text-processing perspective: Evidence of improved comprehensibility. *Reading Research Quarterly*, 26(3), pp. 251-276.
- Becker, L. A. 2000. *Effect Size (ES)*. Available at: <https://bit.ly/2RQ4zx5> [Accessed 12 December 2018].
- Beinborn, L., Zesch, T. and Gurevych, I. 2012. Towards fine-grained readability measures for self-directed language learning. *IN: Borin, L. and Volodina, E. (eds.) Proceedings of the SLTC 2012 Workshop on NLP for CALL*. Lund, Sweden, 24-26 October. Linköping: Linköping University Electronic Press, pp. 11-19.
- Bell, D. S., Harless, C. E., Higa, J. K., Bjork, E. L., Bjork, R. A., Bazargan, M. and Mangione, C. M. 2008. Knowledge retention after an online tutorial: A randomized educational experiment among resident physicians. *Journal of General Internal Medicine*, 23(8), pp. 1164-1171.
- Benjamin, R. G. 2012. Reconstructing readability: Recent developments and recommendations in the analysis of text difficulty. *Educational Psychology Review*, 24(1), pp. 63-88.

- Berkman, M. I. and Karahoca, D. 2016. Re-assessing the usability metric for user experience (UMUX) scale. *Journal of Usability Studies*, 11(3), pp. 89-109.
- Berkman, N. D., Davis, T. C. and McCormack, L. 2010. Health literacy: What is it? *Journal of Health Communication*, 15(S2), pp. 9-19.
- Bernhardt, E. B. 1983. Testing foreign language reading comprehension: The immediate recall protocol. *Die Unterrichtspraxis/Teaching German*, 16(1), pp. 27-33.
- Best, R. M., Floyd, R. G. and McNamara, D. S. 2008. Differential competencies contributing to children's comprehension of narrative and expository texts. *Reading Psychology*, 29(2), pp. 137-164.
- Best, R. M., Rowe, M., Ozuru, Y. and McNamara, D. S. 2005. Deep level comprehension of science texts: The role of the reader and the text. *Topics in Language Disorders*, 25(1), pp. 65-83.
- Bevan, N., Carter, J. and Harker, S. 2015. ISO 9241-11 revised: What have we learnt about usability since 1998? IN: Kurosu, M. (ed.) *Human-Computer Interaction: Design and Evaluation* (Part 1). Berlin: Springer, pp. 143-151.
- Bevan, N., Kirakowski, J. and Maissel, J. 1991. What is usability? IN: Bullinger, H. J. (ed.) *Proceedings of the Fourth International Conference on Human-Computer Interaction*. Stuttgart, Germany, 1-6 September. Amsterdam: Elsevier, pp. 651-655.
- Bhagoliwal, B. 1961. Readability formulae: Their reliability, validity, and applicability in Hindi. *Journal of Education and Psychology*, 19, pp. 13-26.
- Bingel, J. 2018. *Personalized and Adaptive Text Simplification*. PhD thesis. University of Copenhagen.
- Birch, A., Abend, O., Bojar, O. and Haddow, B. 2016. HUME: Human UCCA-based evaluation of machine translation. IN: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2016)*. Austin, Texas, 1-5 November. Red Hook, New York: Curran Associates, pp. 1264-1274.
- Birch, A., Bojar, O., Rosa, R., Ried, J., Hassan, H. and Davenport, C. 2017. *D5.5: Report on User Surveys, Impact Assessment and Automatic Semantic Metrics*. Available at: <https://bit.ly/2E4Hjb6> [Accessed 12 December 2018].
- Birch, A., Ried, J., Davenport, C., Huck, M. and Marecek, D. 2018. *D5.6: Report on Third Year's MT Evaluation*. Available at: <https://bit.ly/2ud4yIo> [Accessed 12 December 2018].
- Blench, M. 2008, Global public health intelligence network (GPHIN). IN: *Proceedings of the Eight Conference of the American Machine Translation Association (AMTA)*. Waikiki, Hawai'i, 21-25 October. Available at: <https://bit.ly/2KCKWgT> [Accessed 12 December 2018].

- Boone, H. N. and Boone, D. A. 2012. Analyzing Likert data. *Journal of Extension*, 50(2), pp. 1-5.
- Booth, P. 1989. *An Introduction to Human-Computer Interaction*. Hove, United Kingdom: Lawrence Erlbaum Associates Publishers.
- Bormuth, J. R. 1969. *Development of Readability Analyses*. Final Report, Project No. 7-0052. Washington, D.C.: U.S. Office of Education.
- Borsci, S., Federici, S. and Lauriola, M. 2009. On the dimensionality of the System Usability Scale: A test of alternative measurement models. *Cognitive Processing*, 10, pp. 193-197.
- Boshuizen, H. and Schmidt, H. 1992. On the role of biomedical knowledge in clinical reasoning by experts. *Cognitive Science*, 16, pp. 153-184.
- Bovair, S. and Kieras, D. E. 1981. *A Guide to Propositional Analysis for Research on Technical Prose* (Technical Report No. 8). Tucson: Department of Psychology, University of Arizona.
- Bowker, L. 2015. Translatability and user experience: Compatible or in conflict? *Localisation Focus — The International Journal of Localisation*, 14(2), pp. 13-27.
- Brajnik, G. 2008. Beyond conformance: The role of accessibility evaluation methods. *IN: Hartmann, S., Zhou, X. and Kirchberg, M. (eds.) Web Information Systems Engineering – WISE 2008 Workshops. Lecture Notes in Computer Science 5176*. Berlin and Heidelberg: Springer, pp. 63-80.
- Bransford, J. D. and Johnson, M. K. 1972. Contextual prerequisites for understanding: Some investigations of comprehension and recall. *Journal of Verbal Learning and Verbal Behavior*, 11(6), pp. 717-726.
- Brantmeier, C. 2006. The effects of language of assessment and L2 reading performance on advanced readers' recall. *The Reading Matrix*, 6(1), pp. 1-17.
- Brantmeier, C., Strube, M. and Yu, X. 2014. Scoring recalls for L2 readers of English in China: Pausal or idea units. *Reading in a Foreign Language*, 26(1), pp. 114-130.
- Bredenkamp, A., Crysmann, B. and Petrea, M. 2000. Looking for errors: A declarative formalism for resource-adaptive language checking. *IN: Proceedings of the Second International Conference on Language Resources and Evaluation (LREC 2000)*. Athens, Greece, 31 May-2 June. Available at: <https://bit.ly/2BHAppF> [Accessed 12 December 2018].

Britton, B. K. and Gülgöz, S. 1991. Using Kintsch's computational model to improve instructional text: Effects of repairing inference calls on recall and cognitive structures. *Journal of Educational Psychology*, 83(3), pp. 329-345.

Brooke, J. 1996. SUS — A quick and dirty usability scale. *Usability Evaluation in Industry*, 89(194), pp. 4-7.

Brooke, J. 2013. SUS: A retrospective. *Journal of Usability Studies*, 8(2), pp. 29-40.

Buljan, I., Malički, M., Wager, E., Puljak, L., Hren, D., Kellie, F., West, H., Alfrević, Ž. and Marušić, A. 2018. No difference in knowledge obtained from infographic or plain language summary of a Cochrane systematic review: Three randomized controlled trials. *Journal of Clinical Epidemiology*, 97, pp. 86-94.

Burton-Roberts, N. 1986. *Analysing Sentences: An Introduction to English Syntax*. London and New York: Longman.

Buzzetti, E., Kalafateli, M., Thorburn, D., Davidson, B. R., Thiele, M., Gluud, L. L., Del Giovane, C., Askgaard, G., Krag, A., Tsochatzis, E. and Gurusamy, K. S. 2017. Pharmacological interventions for alcoholic liver disease (alcohol-related liver disease). *Cochrane Database of Systematic Reviews*, 3(CD011646). doi:10.1002/14651858.CD011646.pub2.

C

Cadwell, P., O'Brien, S. and Teixeira, C. 2018. Resistance and accommodation: Factors for the (non-)adoption of machine translation among professional translators. *Perspectives: Studies in Translatology*, 26(3), pp. 301-321.

Cairns, P. 2013. A commentary on short questionnaires for assessing usability. *Interacting with Computers*, 25(4), pp. 312-316.

Cambridge Dictionary 2018. Available at: <https://bit.ly/2zfTy2z> [Accessed 12 December 2018].

Campbell, D. 2018. New drive to encourage doctors to write to patients in plain English. *The Guardian*, 4 September. Available at: <https://bit.ly/2Q4cGW0> [Accessed 12 December 2018].

Campbell Collaboration 2016. *How to Write a Plain Language Summary for a Campbell Systematic Review*. Available at: <https://bit.ly/2BqeuTJ> [Accessed 12 December 2018].

Candido, A., Maziero, E., Gasperin, C., Pardo, T. A., Specia, L. and Aluísio, S. M. 2009. Supporting the adaptation of texts for poor literacy readers: A text simplification editor for Brazilian Portuguese. IN: Tetreault, J., Burstein, J. and Leacock, C. (eds.) *Proceedings of the Fourth Workshop on Innovative Use of NLP for Building Educational Applications (NAACL HLT 09)*. Boulder, Colorado, 5 June. Stroudsburg, Pennsylvania: Association for Computational Linguistics, pp. 34-42.

- Carbone, E. T. and Zoellner, J. M. 2012. Nutrition and health literacy: A systematic review to inform nutrition research and practice. *Journal of the Academy of Nutrition and Dietetics*, 112(2), pp. 254-265.
- Cardey, S., Greenfield, P. and Wu, X. 2004. Designing a controlled language for the machine translation of medical protocols: The case of English to Chinese. *IN: Frederking, R. E. and Taylor, K. B. (eds.) Machine Translation: From Real Users to Research*. Berlin and Heidelberg: Springer, pp. 37-47.
- Cardinal, R. and Aitken, M. 2013. *ANOVA for the Behavioral Sciences Researcher*. Hove, United Kingdom: Psychology Press.
- Carlisle, J. F. 1999. Free recall as a test of reading comprehension for students with learning disabilities. *Learning Disability Quarterly*, 22(1), pp. 11-22.
- Carrell, P. L. 1983. Three components of background knowledge in reading comprehension. *Language Learning*, 33(2), pp. 183-203.
- Carrell, P. L. 1985. Facilitating ESL reading by teaching text structure. *TESOL Quarterly*, 19(4), pp. 727-752.
- Carroll, J. M. 2002. *Human-Computer Interaction in the New Millennium*. New York: Addison-Wesley.
- Carter, M. 2018. *Configuring Guidelines that Regulate Paragraph and Sentence Length*. Available at: <https://bit.ly/2E3INTe> [Accessed 12 December 2018].
- Cascales, R. R. 2002. *A Specification and Validating Parser for Simplified Technical Spanish*. MSc thesis. University of Limerick.
- Castilho, S. 2016. *Measuring Acceptability of Machine Translated Enterprise Content*. PhD thesis. Dublin City University.
- Castilho, S., Doherty, S., Gaspari, F. and Moorkens, J. 2018a. Approaches to human and machine translation quality assessment. *IN: Moorkens, J., Castilho, S., Gaspari, F. and Doherty, S. (eds.) Translation Quality Assessment: From Principles to Practice*. Basel, Switzerland: Springer International Publishing, pp. 9-38.
- Castilho, S., Moorkens, J., Gaspari, F., Sennrich, R., Sosoni, V., Georgakopoulou, P., Lohar, P., Way, A., Miceli Barone, A. and Gialama, M. 2017. A comparative quality evaluation of PBSMT and NMT using professional translators. *IN: Kurohashi, S. and Fung, P. (eds.) Proceedings of Machine Translation Summit XVI, Vol. 1: MT Research Track*. Nagoya, Japan, 18-22 September. Available at: <https://bit.ly/2w6Kk4g> [Accessed 12 December 2018].

- Castilho, S., Moorkens, J., Gaspari, F., Sennrich, R., Way, A. and Georgakopoulou, P. 2018b. Evaluating MT for massive open online courses: A multifaceted comparison between PBSMT and NMT systems. *Machine Translation*, pp. 1-24.
- Cates, C. J. and Rowe, B. H. 2013. Vaccines for preventing influenza in people with asthma. *Cochrane Database of Systematic Reviews*, 2 (CD000364). doi:10.1002/14651858.CD000364.pub4.
- CDC (Centers for Disease Control and Prevention) 2016. *Everyday Words for Public Health Communication*. Available at: <https://bit.ly/2Du1bnK> [Accessed 12 December 2018].
- Central Statistics Office 2016. *Census of Population of Ireland*. Available at: <https://bit.ly/2P3Ztux> [Accessed 12 December 2018].
- Chall, J. and Dale, E. 1995. *Readability Revisited: The New Dale-Chall Readability Formula*. Cambridge, Massachusetts: Brookline Books.
- Chang, Y. F. 2006. On the use of the immediate recall task as a measure of second language reading comprehension. *Language Testing*, 23(4), pp. 520-543.
- Changyong, F., Hongyue, W., Naiji, L., Tian, C., Hua, H. and Ying, L. 2014. Log-transformation and its implications for data analysis. *Shanghai Archives of Psychiatry*, 26(2), pp. 105-109.
- Chaparro, B. S., Phan, M. H., Siu, C. and Jardina, J. R. 2014. User performance and satisfaction of tablet physical keyboards. *Journal of Usability Studies*, 9(2), pp. 70-80.
- Chen, X., Acosta, S. and Barry, A. E. 2016. Evaluating the accuracy of Google Translate for diabetes education material. *JMIR Diabetes*, 1(1). doi:10.2196/diabetes.5848.
- Chisholm, W. A. and Henry, S. L. 2005. Interdependent components of web accessibility. *IN: Proceedings of the 2005 International Cross-Disciplinary Workshop on Web Accessibility (W4A)*. Chiba, Japan, 10 May. New York: Association for Computing Machinery, pp. 31-37.
- Choi, B. and Pak, A. 2005. A catalog of biases in questionnaires. *Preventing Chronic Disease. Public Health Research, Practice, and Policy*, 2(1), A13.
- Christophersen, T. and Konradt, U. 2010. Reliability, validity and sensitivity of a single-item measure of online store usability. *International Journal of Human-Computer Studies*, 69(4), pp. 269-280.

Clarke, C., Friedman, S. M., Shi, K., Arenovich, T., Monzon, J. and Culligan, C. 2005. Emergency department discharge instructions comprehension and compliance study. *Canadian Journal of Emergency Medicine*, 7(1), pp. 5-11.

Clary, E. G., Snyder, M., Ridge, R. D., Copeland, J., Stukas, A. A., Haugen, J. and Miene, P. 1998. Understanding and assessing the motivations of volunteers: A functional approach. *Journal of Personality and Social Psychology*, 74(6), pp. 1516-1530.

CMS (Centers for Medicare and Medicaid Services) 2012. *Toolkit for Making Written Material Clear and Effective*. Available at: <https://go.cms.gov/2QtvkGD> [Accessed 12 December 2018].

Cochrane Blogshots Used Globally 2018. Available at: <https://bit.ly/2RMOO9z> [Accessed 12 December 2018].

Cochrane Collaboration Steering Group 2013. *Minutes of the Cochrane Collaboration Steering Group Meeting in Oxford, UK, 17 and 20 March 2013*. Available at: <https://bit.ly/2AvbdRi> [Accessed 12 December 2018].

Cochrane Colloquium Vienna 2015. *How Can we Ensure that Cochrane's Standards and Approaches to Plain Language Summaries Are Implemented? A Discussion Workshop*. Available at: <https://bit.ly/2AxayPI> [Accessed 12 December 2018].

Cochrane Database of Systematic Reviews 2018. Available at: <https://bit.ly/2EyNCVn> [Accessed 12 December 2018].

Cochrane Norway 2017. *How to Write a Plain Language Summary of a Cochrane Intervention Review*. Available at: <https://bit.ly/2BG9FWv> [Accessed 12 December 2018].

Cochrane Organisational Dashboard 2017. Available at: <https://bit.ly/2E0JD2A> [Accessed 12 December 2018].

Cochrane Review Groups 2018. Available at: <https://bit.ly/2OGZTM3> [Accessed 12 December 2018].

Cochrane Translation 2018. *Infographic*. Available at: <https://bit.ly/2D5UFUo> [Accessed 12 December 2018].

Collins English Dictionary 2018. Available at: <https://bit.ly/2kiPzHu> [Accessed 12 December 2018].

Collins-Thompson, K. 2014. Computational assessment of text readability: A survey of current and future research. *International Journal of Applied Linguistics*, 165(2), pp. 97-135.

Costa-jussà, M. R., Farrús, M. and Pons, J. S. 2012. A quality analysis of statistical machine translation in the medical domain. *IN: Proceedings of the Fifth Virtual Conference of Advanced Research in Scientific Areas (ARSA)*. Žilina, Slovakia, 3-7 December. EDIS - Publishing Institution of the University of Žilina, pp. 1995-1998.

Coster, W. and Kauchak, D. 2011. Simple English Wikipedia: A new text simplification task. *IN: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (Volume 2)*. Portland, Oregon, 19-24 June. Stroudsburg, Pennsylvania: Association for Computational Linguistics, pp. 665-669.

Council of Europe 2011. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Available at: <https://bit.ly/2w4iMwg> [Accessed 12 December 2018].

Creswell, J. W. and Plano Clark, V. L. 2018. *Designing and Conducting Mixed Methods Research* (3rd ed.). Los Angeles: Sage.

Crossley, S. A., Allen, D. and McNamara, D. S. 2011. Text readability and intuitive simplification: A comparison of readability formulas. *Reading in a Foreign Language*, 23(1), pp. 84-101.

Crossley, S. A., Allen, D. and McNamara, D. S. 2012. Text simplification and comprehensible input: A case for an intuitive approach. *Language Teaching Research*, 16(1), pp. 89-108.

Crossley, S. A., Dufty, D. F., McCarthy, P. M. and McNamara, D. S. 2007a. Toward a new readability: A mixed model approach. *IN: Proceedings of the 29th Annual Meeting of the Cognitive Science Society (CogSci 2007)*. Nashville, Tennessee, 1-4 August. New York: Lawrence Erlbaum Associates, pp. 197-202.

Crossley, S. A., Greenfield, J. and McNamara, D. S. 2008. Assessing text readability using cognitively based indices. *Tesol Quarterly*, 42(3), pp. 475-493.

Crossley, S. A., Louwrese, M. M., McCarthy, P. M. and McNamara, D. S. 2007b. A linguistic analysis of simplified and authentic texts. *The Modern Language Journal*, 91(1), pp. 15-30.

Crossley, S. A. and McNamara, D. S. 2008. Assessing L2 reading texts at the intermediate level: An approximate replication of Crossley, Louwrese, McCarthy and McNamara (2007). *Language Teaching*, 41(3), pp. 409-429.

Crossley, S. A. and McNamara, D. S. 2016. Text-based recall and extra-textual generations resulting from simplified and authentic texts. *Reading in a Foreign Language*, 28(1), pp. 1-19.

Crossley, S. A., Yang, H. S. and McNamara, D. S. 2014. What's so simple about simplified texts? A computational and psycholinguistic investigation of text comprehension and text processing. *Reading in a Foreign Language*, 26(1), pp. 92-113.

D

D'Alessandro, D., Kingsley, P. and Johnson-West, J. 2001. The readability of pediatric patient education materials on the World Wide Web. *Archives of Pediatrics & Adolescent Medicine*, 155(7), pp. 807-812.

Dağ, F., Durdu, L. and Gerdan, S. 2014. Evaluation of educational authoring tools for teachers stressing of perceived usability features. *Procedia Social and Behavioral Sciences*, 116, pp. 888-901.

Davis, T. C., Fredrickson, D. D., Arnold, C., Murphy, P. W., Herbst, M. and Bocchini, J. A. 1998. A polio immunization pamphlet with increased appeal and simplified language does not improve comprehension to an acceptable level. *Patient Education and Counseling*, 33(1), pp. 25-37.

Dew, K., Turner, A. M., Choi, Y. K., Bosold, A. and Kirchhoff, K. 2018. Development of machine translation technology for assisting health communication: A systematic review. *Journal of Biomedical Informatics*, 85, pp. 56-67.

Dew, K., Turner, A. M., Desai, L., Martin, N., Laurenzi, A. and Kirchhoff, K. 2015. PHAST: A collaborative machine translation and post-editing tool for public health. *IN: Proceedings of the AMIA Annual Symposium*. San Francisco, California, 14-18 November. Bethesda, Maryland: American Medical Informatics Association, pp. 492-501.

Di Marco, C., Bray, P., Covvey, H. D., Cowan, D. D., Di Ciccio, V., Hovy, E., Lipa, J. and Yang, C. 2006. Authoring and generation of individualized patient education materials. *IN: Proceedings of the AMIA Annual Symposium*. Washington, D.C., 11-15 November. Bethesda, Maryland: American Medical Informatics Association, pp. 195-199.

Diao, Y. and Sweller, J. 2007. Redundancy in foreign language reading comprehension instruction: Concurrent written and spoken presentations. *Learning and Instruction*, 17(1), pp. 78-88.

Dix, A., Finlay, J., Abowd, G. D. and Beale, R. 2004. *Human-Computer Interaction* (3rd ed.). Harlow, United Kingdom: Pearson.

Doherty, S. 2012. *Investigating the Effects of Controlled Language on the Reading and Comprehension of Machine Translated Texts: A Mixed-Methods Approach*. PhD thesis. Dublin City University.

- Doherty, S. 2017. Issues in human and automatic translation quality assessment. *IN: Kenny, D. (ed.) Human Issues in Translation Technology*. London: Routledge, pp. 131-148.
- Doherty, S. and O'Brien, S. 2012. A user-based usability assessment of raw machine translated technical instructions. *IN: Proceedings of the Tenth Biennial Conference of the Association for Machine Translation in the Americas (AMTA)*. San Diego, California, 28 October-1 November. Available at: <https://bit.ly/2TXxX51> [Accessed 12 December 2018].
- Doing-Harris, K., Patterson, O., Igo, S. and Hurdle, J. 2013. Document sublanguage clustering to detect medical specialty in cross-institutional clinical texts. *IN: Proceedings of the Seventh International Workshop on Data and Text Mining in Biomedical Informatics*. San Francisco, California, 27 October-1 November. New York: Association for Computing Machinery, pp. 9-12.
- Dowell, N. M., Graesser, A. C. and Cai, Z. 2016. Language and discourse analysis with Coh-Matrix: Applications from educational material to learning environments at scale. *Journal of Learning Analytics*, 3(3), pp. 72-95.
- Drugan, J. 2013. *Quality in Professional Translation: Assessment and Improvement*. London: Bloomsbury.
- DuBay, W. 2004. *The Principles of Readability*. Available at: <https://bit.ly/2SifbUM> [Accessed 12 December 2018].
- DuBay, W. 2007. *Smart Language: Readers, Readability, and the Grading of Text*. Available at: <https://bit.ly/2PZYVeq> [Accessed 12 December 2018].
- Dufty, D. F., Graesser, A. C., Louwerse, M. M. and McNamara, D. S. 2006. Assigning grade levels to textbooks: Is it just readability? *IN: Proceedings of the 28th Annual Conference of the Cognitive Science Society (CogSci 2006)*. Vancouver, Canada, 26-29 July. Merced, California: University of California, pp. 1251–1256.
- Duma, L., Chininthorn, P., Glaser, M. and Tucker, W. D. 2015. Usability of an authoring tool for generalised scenario creation for SignSupport. *IN: Otten, F. and Balmahoon, R. (eds.) Proceedings of the Southern Africa Telecommunications Networks and Applications Conference (SATNAC)*. Hermanus, South Africa, 6-9 September. Available at: <https://bit.ly/2TXvvfx> [Accessed 12 December 2018].
- Duran, N. D., Bellissens, C., Taylor, R. S. and McNamara, D. S. 2007. Quantifying text difficulty with automated indices of cohesion and semantics. *IN: Proceedings of the 29th Annual Meeting of the Cognitive Science Society (CogSci 2007)*. Nashville, Tennessee, 1-4 August. New York: Lawrence Erlbaum Associates, pp. 233-238.
- Dušek, O., Hajič, J., Hlaváčová, J., Novák, M., Pecina, P., Rosa, R., Tamchyna, A., Urešová, Z. and Zeman, D. 2014. Machine translation of medical texts in the Khresmoi

project. *IN: Proceedings of the Ninth Workshop on Statistical Machine Translation*. Baltimore, Maryland, 26-27 June. Stroudsburg, Pennsylvania: Association for Computational Linguistics, pp. 221-228.

Dyson, M. and Hannah, J. 1987. Toward a methodology for the evaluation of machine assisted translation systems. *Computers and Translation*, 2(3), pp. 163-176.

E

Editorial Team 2018. Available at: <https://bit.ly/2E7VhJ8> [Accessed 12 December 2018].

Ellis, R. and Barkhuizen, G. 2005. *Analysing Learner Language*. Oxford: Oxford University Press.

Eltorai, A. E., Naqvi, S. S., Ghanian, S., Ebersson, C. P., Weiss, A., Born, C. T. and Daniels, A. H. 2015. Readability of invasive procedure consent forms. *Clinical and Translational Science*, 8(6), pp. 830-833.

European Commission 2014. *Flash Eurobarometer 404. European Citizens' Digital Health Literacy*. Available at: <https://bit.ly/2FXpXia> [Accessed 12 December 2018].

F

Fang, M. C., Panguluri, P., Machtinger, E. L. and Schillinger, D. 2009. Language, literacy, and characterization of stroke among patients taking warfarin for stroke prevention: Implications for health communication. *Patient Education and Counseling*, 75(3), pp. 403-410.

Farndale, E. 2004. *The Intra-Organisational Power of the Personnel Department in Higher Education in the UK*. PhD thesis. Cranfield University.

Favreau, M. and Segalowitz, N. 1982. Second language reading in fluent bilinguals. *Applied Psycholinguistics*, 3(4), pp. 329-341.

Feng, L., Elhadad, N. and Huenerfauth, M. 2009. Cognitively motivated features for readability assessment. *IN: Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*. Athens, Greece, 30 March-3 April. Stroudsburg, Pennsylvania: Association for Computational Linguistics, pp. 229-237.

Finstad, K. 2010. The usability metric for user experience. *Interacting with Computers*, 22(5), pp. 323-327.

Fisher, D. and Frey, N. 2015. Selecting texts and tasks for content area reading and learning. *The Reading Teacher*, 68(7), pp. 524-529.

Flesch, R. 1948. A new readability yardstick. *Journal of Applied Psychology*, 32(3), pp. 221-233.

Flodgren, G. 2016. Are Cochrane plain language summaries plain? *IN: Abstracts of the 24th Cochrane Colloquium. Challenges to Evidence-Based Health Care and Cochrane.* Seoul, Korea, 23-27 October. Available at: <https://bit.ly/2QoQTLz> [Accessed 12 December 2018].

Frey, L. R., Botan, C. H., Friedman, P. G. and Kreps, G. L. 1991. *Investigating Communication: An Introduction to Research Methods.* Upper Saddle River, New Jersey: Prentice Hall.

Friedman, D. B. and Hoffman-Goetz, L. 2006. A systematic review of readability and comprehension instruments used for print and web-based cancer information. *Health Education & Behavior*, 33(3), pp. 352-373.

Frøkjær, E., Hertzum, M. and Hornæk, K. 2000. Measuring usability: Are effectiveness, efficiency and satisfaction really correlated? *IN: Turner, T., Szwillus, G., Czerwinski, M., Paternò, F. and Pemberton, S. (eds.) Proceedings of the Conference of Human Factors in Computing Systems (CHI 2000).* The Hague, Netherlands, 1-6 April. New York: Association for Computing Machinery, pp. 345-352.

Fuchs, N. E., Kaljurand, K. and Kuhn, T. 2008. Attempto controlled English for knowledge representation. *IN: Baroglio, C., Bonatti, P. A., Małuszyński, J., Marchiori, M., Polleres, A. and Schaffert, S. (eds.) Reasoning Web.* Berlin and Heidelberg: Springer, pp. 104-124.

G

Gaeta, M., Loia, V., Mangione, G. R., Orciuoli, F., Ritrovato, P. and Salerno, S. 2014. A methodology and an authoring tool for creating Complex Learning Objects to support interactive storytelling. *Computers in Human Behavior*, 31, pp. 620-637.

Gallagher, T., Fazio, X. and Gunning, T. 2012. Varying readability of science-based text in elementary readers: Challenges for teachers. *Reading Improvement*, 49(3), pp. 93-112.

Gambrell, L. B. 2011. Seven rules of engagement: What's most important to know about motivation to read. *The Reading Teacher*, 65(3), pp. 172-178.

Gan, S. 2012. *Lost in Translation: How Much Is Translation Costing the NHS, and How Can We both Cut Costs and Improve Service Provision?* Available at: <https://bit.ly/2ygANcH> [Accessed 12 December 2018].

Gaspari, F., Almaghout, H. and Doherty, S. 2015. A survey of machine translation competences: Insights for translation technology educators and practitioners. *Perspectives*, 23(3), pp. 333-358.

Ghasemi, A. and Zahediasl, S. 2012. Normality tests for statistical analysis: A guide for non-statisticians. *International Journal of Endocrinology and Metabolism*, 10(2), pp. 486-489.

- Gigliotti, G. 2017. The quality of mercy: A corpus-based analysis of the quality of volunteer translations for non-profit organisations (NPOs). *New Voices in Translation Studies*, 17, pp. 52-81.
- Gilliver, S. 2015. Online plain English and readability resources. *Medical Writing*, 24(1), pp. 20-22.
- Glenton, C. 2017. *Assessing the Feasibility and Acceptability of Approaches for Improving the Quality of Plain Language Summaries in Cochrane Reviews: A Pilot Study*. Available at: <https://bit.ly/2qWr4Ee> [Accessed 12 December 2018].
- Glenton, C., Santesso, N., Rosenbaum, S., Nilsen, E. S., Rader, T., Ciapponi, A. and Dilkes, H. 2010. Presenting the results of Cochrane Systematic Reviews to a consumer audience: A qualitative study. *Medical Decision Making*, 30, pp. 566-577.
- Godden, K. 2000. The evolution of CASL controlled authoring at General Motors. *IN: Proceedings of the Third International Workshop on Controlled Language Applications (CLAW 2000)*. Seattle, Washington, 29-30 April. Available at: <https://bit.ly/2SRKRRS> [Accessed 12 December 2018].
- Gordillo, A., Barra, E. and Quemada, J. 2017. An easy to use open source authoring tool to create effective and reusable learning objects. *Computer Applications in Engineering Education*, 25(2), pp. 188-199.
- Graesser, A. C., McNamara, D. S., Cai, Z., Conley, M., Li, H. and Pennebaker, J. 2014. Coh-Metrix measures text characteristics at multiple levels of language and discourse. *The Elementary School Journal*, 115(2), pp. 210-229.
- Graesser, A. C., McNamara, D. S. and Kulikowich, J. M. 2011. Coh-Metrix: Providing multilevel analyses of text characteristics. *Educational Researcher*, 40(5), pp. 223-234.
- Graesser, A. C., McNamara, D. S., Louwrese, M. M. and Cai, Z. 2004. Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers*, 36(2), pp. 193-202.
- Green, S., Higgins, J., Alderson, P., Clarke, M., Mulrow, C. and Oxman, A. 2011. Introduction. *IN: Higgins, J. and Green, S. (eds.) Cochrane Handbook for Systematic Reviews of Interventions* (version 5.1.0). Available at: <https://bit.ly/2yM5INh> [Accessed 12 December 2018].
- Greene, M., Cleary, Y. and Marcus-Quinn, A. 2017. Use of plain-language guidelines to promote health literacy. *IEEE Transactions on Professional Communication*, 60(4), pp. 384-400.
- Group Health Research Institute 2009. *PRISM [Program for Readability in Science and Medicine]*. Available at: <https://bit.ly/2qDMLGu> [Accessed 12 December 2018].

Guillardeau, S. 2009. *Freie Translation Memory Systeme für die Übersetzungspraxis: Ein kritischer Vergleich*. MA dissertation. University of Vienna.

Gyte, G. and Struthers, C. 2015. Writing and commenting on plain language summaries (PLSs) to improve quality. *IN: Abstracts of the 23rd Cochrane Colloquium. Filtering the Information Overload for Better Decisions*. Vienna, Austria, 3-7 October. Available at: <https://bit.ly/2rcQ5LK> [Accessed 12 December 2018].

H

Hall, A. K., Stellefson, M. and Bernhardt, J. M. 2012. Healthy aging 2.0: The potential of new media and technology. *Preventing Chronic Disease*, 9(110241). doi:10.5888/pcd9.110241.

Hallgren, K. A. 2012. Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorials in Quantitative Methods for Psychology*, 8(1), pp. 23-34.

Hansen, C. L. 1978. Story retelling used with average and learning disabled readers as a measure of reading comprehension. *Learning Disability Quarterly*, 1(3), pp. 62-69.

Harniss, M., Witzel, J., Westbrook, J. and Starks, J. 2013. *Plain Language Summary Tool*. Austin, Texas: Center on Knowledge Translation for Disability and Rehabilitation Research; Seattle, Washington: University of Washington, Center for Technology and Disability Studies. Available at: <https://bit.ly/2DzyzbO> [Accessed 12 December 2018].

Harper, R. and Zimmerman, D. 2009. Exploring plain language guidelines. *IN: Proceedings of the IEEE International Professional Communication Conference (IPCC 2009)*. University of Twente, Enschede, The Netherlands, 19-22 July. Available at: <https://bit.ly/2OwZfeU> [Accessed 12 December 2018].

Harvey, L. A. 2018. Summaries of Cochrane Systematic Reviews: Making high-quality evidence accessible. *Spinal Cord*, 56(185). doi:10.1038/s41393-018-0071-5.

Heilman, J. and West, A. 2015. Wikipedia and medicine: Quantifying readership, editors, and the significance of natural language. *Journal of Medical Internet Research*, 17(3). doi:10.2196/jmir.4069.

Helsel, D. and Hirsch, R. 2002. *Statistical Methods in Water Resources* (Vol. 323). Reston, Virginia: US Geological Survey.

HHS (Department of Health and Human Services) 2018. *Plain Writing Act Compliance Report*. Available at: <https://bit.ly/2SWD2dJ> [Accessed 12 December 2018].

Hiebert, E. H. 2002. Standards, assessment, and text difficulty. *IN: Farstrup, A. E. and Samuels, S. J. (eds.) What Research Has to Say about Reading Instruction* (3rd ed.). Newark, Delaware: International Reading Association, pp. 337-369.

Higgins, J. and Green, S. (eds.) 2011. *Cochrane Handbook for Systematic Reviews of Interventions* (Version 5.1.0). Available at: <https://bit.ly/2t90rNM> [Accessed 12 December 2018].

HimL (Health in my Language) 2016. *HimL Project Demo*. Available at: <https://bit.ly/2NuZ9oZ> [Accessed 12 December 2018].

Hochheiser, H. and Lazar, J. 2007. HCI and societal issues: A framework for engagement. *International Journal of Human-Computer Interaction*, 23(3), pp. 339-374.

Hoonakker, P., Carayon, P., Gurses, A., Brown, R., McGuire, K., Khunlertkit, A. and Walker, J. M. 2011. Measuring workload of ICU nurses with a questionnaire survey: The NASA Task Load Index (TLX). *IIE Transactions on Healthcare Systems Engineering*, 1(2), pp. 131-143.

Hornbæk, K. 2006. Current practice in measuring usability: Challenges to usability studies and research. *International Journal of Human-Computer Studies*, 64, pp. 79-102.

Hornbæk, K. and Law, E. L. 2007. Meta-analysis of correlations among usability measures. *IN: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI 2007)*. San Jose, California, 30 April-3 May. New York: Association for Computing Machinery, pp. 617-626.

Horowitz, M. W. and Newman, J. B. 1964. Spoken and written expression: An experimental analysis. *The Journal of Abnormal and Social Psychology*, 68(6), pp. 640-647.

Houts, P. S., Doak, C. C., Doak, L. G. and Loscalzo, M. J. 2006. The role of pictures in improving health communication: A review of research on attention, comprehension, recall, and adherence. *Patient Education and Counseling*, 61, pp. 173-190.

HSE (Health Service Executive) 2017. *Guidelines for Communicating Clearly Using Plain English with Our Patients and Service Users*. Available at: <https://bit.ly/2OyucQ2> [Accessed 12 December 2018].

Hughes, B., Joshi, I. and Wareham, J. 2008. Health 2.0 and Medicine 2.0: Tensions and controversies in the field. *Journal of Medical Internet Research*, 10(3). doi:10.2196/jmir.1056.

Huijsen, W. O. 1998. Controlled language: An introduction. *IN: Mitamura, T. (ed.) Proceedings of the Second International Workshop on Controlled Language Applications (CLAW)*. Pittsburgh, Pennsylvania: Language Technologies Institute, Carnegie Mellon University, pp. 1-15.

Huitema, B. 2011. *The Analysis of Covariance and Alternatives: Statistical Methods for Experiments, Quasi-Experiments, and Single-Case studies* (2nd ed.). Hoboken, New Jersey: Wiley.

I

Introducing Kiswahili for Microsoft Translator 2015. Available at: <https://bit.ly/2vq9E5A> [Accessed 12 December 2018].

Introduction to RevMan Web 2018. Available at: <https://bit.ly/2LeLB0b> [Accessed 12 December 2018].

ISO (International Organization for Standardization) 1998. *ISO 9241-11:1998. Ergonomic Requirements for Office Work with Visual Display Terminals (VDTs) - Part 11: Guidance on Usability*. Preview available at: <https://bit.ly/166ufwn> [Accessed 12 December 2018].

ISO (International Organization for Standardization) 2018. *ISO 9241-11:2018. Ergonomics of Human-System Interaction - Part 11: Usability: Definitions and Concepts*. Preview available at: <https://bit.ly/2MM4Ay4> [Accessed 12 December 2018].

Issa, T. and Isaias, P. 2015. *Sustainable Design: HCI, Usability and Environmental Concerns*. London: Springer-Verlag.

Ive, J. 2017. *Towards a Better Human-Machine Collaboration in Statistical Translation: Example of Systematic Medical Reviews*. PhD thesis. University of Paris-Saclay.

Izumi, E., Uchimoto, K. and Isahara, H. 2006. Measuring intelligibility of Japanese learner English. *IN: Salakoski, T., Ginter, F., Pyysalo, S. and Pahikkala, T. (eds.) Advances in Natural Language Processing*. Berlin and Heidelberg: Springer, pp. 476-487.

J

Jackson, G. T., Allen, L. K. and McNamara, D. S. 2016. Common Core TERA: Text ease and readability assessor. *IN: Crossley, S. A. and McNamara, D. S. (eds.) Adaptive Educational Technologies for Literacy Instruction*. New York and London: Routledge, pp. 49-68.

Johnston, P. 1981. *Implications of Basic Research for the Assessment of Reading Comprehension* (Technical Report No. 2016). Washington, D.C.: National Institute of Education.

Joyce, E. and Kraut, R. E. 2006. Predicting continued participation in newsgroups. *Journal of Computer-Mediated Communication*, 11, pp. 723-747.

K

Kadic, A. J., Fidahic, M., Vujcic, M., Saric, F., Propadalo, I., Marelja, I., Dosenovic, S. and Puljak, L. 2016. Cochrane plain language summaries are highly heterogeneous with

low adherence to the standards. *BMC Medical Research Methodology*, 16(61). doi:10.1186/s12874-016-0162-y.

Kajzer-Wietrzny, M., Whyatt, B. and Stachowiak, K. 2016. Simplification in inter-and intralingual translation — combining corpus linguistics, key logging and eye-tracking. *Poznan Studies in Contemporary Linguistics*, 52(2), pp. 235-267.

Kammerbauer, M. and Minnery, J. 2018. Risk communication and risk perception: Lessons from the 2011 floods in Brisbane, Australia. *Disasters*, 43(1), pp. 110-134.

Kandula, S. and Zeng-Treitler, Q. 2008. Creating a gold standard for the readability measurement of health texts. *IN: Proceedings of the AMIA Annual Symposium*. Washington, D.C., 8-12 November. Bethesda, Maryland: American Medical Informatics Association, pp. 353-357.

Kao, L. and Green, C. 2008. Analysis of variance: Is there a difference in means and what does it mean? *Journal of Surgical Research*, 144(1), pp. 158-170.

Karačić, J., Buljan, I., Hren, D., Dondi, P. and Marušić, A. 2017. Readability of different formats of information about Cochrane Systematic Reviews: A cross sectional study. *IN: Abstracts of the Ninth Croatian Cochrane Symposium*. Split, Croatia, 9-10 June. Available at: <https://bit.ly/2ShUmJa> [Accessed 12 December 2018].

Keenan, J. M., Betjemann, R. S. and Olson, R. K. 2008. Reading comprehension tests vary in the skills they assess: Differential dependence on decoding and oral comprehension. *Scientific Studies of Reading*, 12(3), pp. 281-300.

Khodambashi, S. and Nytrø, Ø. 2017. A systematic literature review on evaluation of digital tools for authoring evidence-based clinical guidelines. *IN: Ryan, A., Schaper, L. K. and Whetton, S. (eds.) Integrating and Connecting Care*. Amsterdam: IOS Press, pp. 48-54.

Kim, H., Zeng-Treitler, Q., Goryachev, S., Keselman, A., Slaughter, L. and Arnott, C. 2007. Text characteristics of clinical reports and their implications for the readability of personal health records. *IN: Kuhn, K. A., Warren, J. R. and Leong, T. (eds.) Proceedings of the 12th World Congress on Health (Medical) Informatics; Building Sustainable Health Systems*. Brisbane, Australia, 20-24 August. Amsterdam: IOS Press, pp. 1117-1121.

Kindig, D. A., Panzer, A. M. and Nielsen-Bohlman, L. 2004. *Health Literacy: A Prescription to End Confusion*. Washington, D.C.: National Academies Press.

Kintsch, W. 1988. The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review*, 95(2), pp. 163-182.

Kintsch, W. 1998. *Comprehension: A Paradigm for Cognition*. Cambridge, United Kingdom: Cambridge University Press.

Kintsch, W. 2004. The construction-integration model of text comprehension and its implication for instruction. *Theoretical Models and Processes of Reading*, 5, pp. 1270-1328.

Kirakowski, J. 1994. *The Use of Questionnaire Methods for Usability Assessment*. Available at: <https://bit.ly/2E13Jd2> [Accessed 12 December 2018].

Kirakowski, J., Claridge, N. and Whitehand, R. 1998. Human centered measures of success in web site design. IN: *Proceedings of the Fourth Conference on Human Factors and the Web*. Basking Ridge, New Jersey, 5 June. Available at: <https://bit.ly/2KFZc09> [Accessed 12 December 2018].

Kirakowski, J. and Corbett, M. 1993. SUMI: the Software Usability Measurement Inventory. *British Journal of Educational Technology*, 24(3), pp. 210-212.

Kirchhoff, K., Capurro, D. and Turner, A. 2012. Evaluating user preferences in machine translation using conjoint analysis. IN: *Proceedings of the 16th Annual Conference of the European Association for Machine Translation*. Trento, Italy, 28-30 May. European Association for Machine Translation, pp. 119-126.

Kirchhoff, K., Capurro, D. and Turner, A. 2014. A conjoint analysis framework for evaluating user preferences in machine translation. *Machine Translation*, 28(1), pp. 1-17.

Kirchhoff, K., Turner, A. M., Axelrod, A. and Saavedra, F. 2011. Application of statistical machine translation to public health information: A feasibility study. *Journal of the American Medical Informatics Association*, 18(4), pp. 473-478.

Klare, G. R. 1974. Assessing readability. *Reading Research Quarterly*, 10(1), pp. 62-102.

Kline, P. 2000. *A Psychometrics Primer*. London: Free Association Books.

Knowledge Translation in Multi-Languages 2018. Available at: <https://bit.ly/2uhZ6VB> [Accessed 12 December 2018].

Koehn, P. and Monz, C. 2006. Manual and automatic evaluation of machine translation between European languages. IN: *Proceedings of the Workshop on Statistical Machine Translation (HLT-NAACL 06)*. New York, 8-9 June. Stroudsburg, Pennsylvania: Association for Computational Linguistics, pp. 102-121.

Koo, T. K. and Li, M. Y. 2016. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, 15, pp. 155-163.

Kools, M. 2007. A focus on the usability of health education materials. *Patient Education and Counseling*, 65, pp. 275-276.

Kools, M., Ruiter, R. A., Van de Wiel, M. W. and Kok, G. 2004. Increasing readers' comprehension of health education brochures: A qualitative study into how professional writers make texts coherent. *Health Education & Behavior*, 31(6), pp. 720-740.

Koponen, M. 2010. Assessing machine translation quality with error analysis. *IN: Electronic Proceedings of the KäTu Symposium on Translation and Interpreting Studies*. Available at: <https://bit.ly/2Ni6c7l> [Accessed 12 December 2018].

Kortum, P. and Acemyan, C. Z. 2013. How low can you go? Is the System Usability Scale range restricted? *Journal of Usability Studies*, 9(1), pp. 14-24.

Krug, S. 2014. *Don't Make Me Think, Revisited. A Common Sense Approach to Web and Mobile Usability* (3rd ed.). San Francisco: New Riders.

Krüger, R. 2016. Contextualising computer-assisted translation tools and modelling their usability. *Trans-Kom - Journal of Translation and Technical Communication Research*, 9(1), pp. 114-148.

Kuhn, T. 2014. A survey and classification of controlled natural languages. *Computational Linguistics*, 40(1), pp. 121-170.

Kurtzman, E. T. and Greene, J. 2016. Effective presentation of health care performance information for consumer decision making: A systematic review. *Patient Education and Counseling*, 99(1), pp. 36-43.

L

La Biblioteca Cochrane Plus 2018. Available at: <https://bit.ly/2L8NOJH> [Accessed 12 December 2018].

Lachance, C. R., Erby, L. A., Ford, B. M., Allen, V. C. and Kaphingst, K. A. 2010. Informational content, literacy demands, and usability of websites offering health-related genetic tests directly to consumers. *Genetics in Medicine*, 12(5), pp. 304-312.

Landauer, T. K. 1997. Behavioral research methods in human-computer interaction. *IN: Helander, M. G., Landauer, T. K. and Prabhu, P. V. (eds.) Handbook of Human-Computer Interaction* (2nd ed.). Amsterdam: Elsevier, pp. 203-227.

Langendam, M. W., Akl, E. A., Dahm, P., Glasziou, P., Guyatt, G. and Schünemann, H. J. 2013. Assessing and presenting summaries of evidence in Cochrane Reviews. *Systematic Reviews*, 2(81). doi:10.1186/2046-4053-2-81.

Läubli, S., Sennrich, R. and Volk, M. 2018. Has machine translation achieved human parity? A case for document-level evaluation. Available at: <https://bit.ly/2M4cumf> [Accessed 12 December 2018].

- Lazar, J., Feng, J. H. and Hochheiser, H. 2010. *Research Methods in Human-Computer Interaction*. Chichester, England: John Wiley & Sons Ltd.
- Lee, H. J., Lee, M. K., Jeong, Y. S., Han, S. K. and Joung, S. T. 2005. Design and implementation of multimedia content authoring system for healthcare information. *IN: Proceedings of Seventh International Workshop on Enterprise Networking and Computing in Healthcare Industry (HEALTHCOM 2005)*. Busan, South Korea, 23-25 June. IEEE, pp. 311-315.
- Lee, J. E and Ballman, T. L. 1987. FL learners' ability to recall and rate the important ideas of an expository text. *IN: VanPatten, B., Dvorak, T. and Lee, J. E. (eds.) Foreign Language Learning: A Research Perspective*. Cambridge, Massachusetts: Newbury House, pp. 108-117.
- Lee, M. C. 2000. Knowledge, barriers, and motivators related to cervical cancer screening among Korean-American women: A focus group approach. *Cancer Nursing*, 23(3), pp. 168-175.
- Leroy, G. and Endicott, J. E. 2011. Term familiarity to indicate perceived and actual difficulty of text in medical digital libraries. *IN: Xing, C., Crestani, F. and Rauber, A. (eds.) 13th International Conference on Asian-Pacific Digital Libraries (ICADL 2011)*. Beijing, China, 24-27 October. Berlin: Springer, pp. 307-310.
- Leroy, G., Eryilmaz, E. and Laroya, B. T. 2006. Health information text characteristics. *IN: Proceedings of the AMIA Annual Symposium*. Washington, D.C., 11-15 November. Bethesda, Maryland: American Medical Informatics Association, pp. 479-483.
- Leroy, G., Kauchak, D. and Mouradi, O. 2013. A user-study measuring the effects of lexical simplification and coherence enhancement on perceived and actual text difficulty. *International Journal of Medical Informatics*, 82(8), pp. 717-730.
- Levine, T. R. and Hullett, C. R. 2002. Eta squared, partial eta squared, and misreporting of effect size in communication research. *Human Communication Research*, 28(4), pp. 612-625.
- Lewis, J. R. and Sauro, J. 2009. The factor structure of the system usability scale. *IN: Jacko, J. A., Stephanidis, C., Harris, D., Schmorow, D. D., Grootjen, M., Karsh, B. T., Shumaker, R., Zaphiris, P., Ozok, A. A., Duffy, V. G., Kurosu, M., Smith, M. J., Salvendy, G., Aykin, N. and Estabrooke, I. V. (eds.) Proceedings of the 13th International Conference on Human-Computer Interaction (HCI 2009)*. San Diego, California, 19-24 July. New York: Springer, pp. 94-103.
- Lewis, J. R., Utesch, B. S. and Maher, D. E. 2013. UMUX-LITE – When there is no time for the SUS. *IN: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI 2013)*. Paris, France, 27 April-2 May. Paris: Association for Computing Machinery, pp. 2099–2102.

Leys, C., Ley, C., Klein, O., Bernard, P. and Licata, L. 2013. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology*, 49(4), pp. 764-766.

Lieberman, M. and Cunningham, W. 2009. Type I and Type II error concerns in fMRI research: Re-balancing the scale. *Social, Cognitive and Affective Neuroscience*, 4(4), pp. 423-428.

Lindgaard, G. and Kirakowski, J. 2013. Introduction to the special issue: The tricky landscape of developing rating scales in HCI. *Interacting with Computers*, 25, pp. 271-277.

Ling, K., Beenen, G., Ludford, P., Wang, X., Chang, K., Li, X., Cosley, D., Frankowski, D., Terveen, L., Rashid, A. M. and Resnick, P. 2005. Using social psychology to motivate contributions to online communities. *Journal of Computer-Mediated Communication*, 10(4). doi:10.1111/j.1083-6101.2005.tb00273.x.

Linguistic Data Consortium 2002. *Linguistic Data Annotation Specification: Assessment of Fluency and Adequacy in Arabic-English and Chinese-English Translations*. Available at: <https://bit.ly/2zyCrXJ> [Accessed 12 December 2018].

Liu, C. J., Kemper, S. and Bovaird, J. A. 2009. Comprehension of health-related written materials by older adults. *Educational Gerontology*, 35(7), pp. 653-668.

Liu, C. J. and Rawl, S. M. 2012. Effects of text cohesion on comprehension and retention of colorectal cancer screening information: A preliminary study. *Journal of Health Communication*, 17(Sup3), pp. 222-240.

Lonsdale, M. 2014. Typographic features of text: Outcomes from research and practice. *Visible Language*, 48(3), pp. 29-67.

Lund Research Ltd 2013. *Friedman Test in SPSS Statistics*. Available at: <https://bit.ly/2QsRyeV> [Accessed 12 December 2018].

M

MacGinitie, W. and MacGinitie, R. 1989. *Gates-MacGinitie Reading Test* (3rd ed.). Rolling Meadows, Illinois: Riverside Publishing.

MacKenzie, I. S. 2013. *Human-Computer Interaction: An Empirical Research Perspective*. Burlington, Massachusetts: Morgan Kaufmann.

Maguire, L. K. and Clarke, M. 2014. How much do you need: A randomised experiment of whether readers can understand the key messages from summaries of Cochrane Reviews without reading the full review. *Journal of the Royal Society of Medicine*, 107(11), pp. 444-449.

- Malloy-Weir, L. J., Charles, C., Gafni, A. and Entwistle, V. 2016. A review of health literacy: Definitions, interpretations, and implications for policy initiatives. *Journal of Public Health Policy*, 37(3), pp. 334-352.
- Martikainen, H. 2018. A functional approach to translation quality assessment: Categorizing sources of translational distortion in medical abstracts. *Linguistica Antverpiensia, New Series: Themes in Translation Studies*, 14, pp. 106-121.
- Matthews, B. and Ross, L. 2010. *Research Methods: A Practical Guide for the Social Sciences*. Edinburgh: Pearson Education Ltd.
- Max, A. 2006. Writing for language-impaired readers. *IN: Gelbukh, A. (ed.) Computational Linguistics and Intelligent Text Processing*. Berlin and Heidelberg: Springer, pp. 567-570.
- Mc Laughlin, G. H. 1969. SMOG grading - A new readability formula. *Journal of Reading*, 12(8), pp. 639-646.
- McDonald, J. 2009. *Handbook of Biological Statistics* (2nd ed.). Baltimore, Maryland: Sparky House Publishing.
- McGraw, K. O. and Wong, S. P. 1996. Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, 1(1), pp. 30-46.
- McIlwain, C. and Tovey, D. 2018. *PubMed Health and Cochrane's Plain Language Summaries*. Available at: <https://bit.ly/2DPcdTU> [Accessed 12 December 2018].
- McNamara, D. S. and Graesser, A. C. 2012. Coh-Metrix: An automated tool for theoretical and applied natural language processing. *IN: McCarthy, P. and Boonthum-Denecke, C. (eds.) Applied Natural Language Processing: Identification, Investigation, and Resolution*. Hershey, Pennsylvania: IGI Global, pp. 188-205.
- McNamara, D. S., Graesser, A. C., Cai, Z. and Kulikowich, J. 2011. Coh-Metrix easability components: Aligning text difficulty with theories of text comprehension. *IN: Annual Meeting of the American Educational Research Association (AERA)*. Available at: <https://bit.ly/2Rp2Tdo> [Accessed 12 December 2018].
- McNamara, D. S., Graesser, A. C., McCarthy, P. and Cai, Z. 2014. *Automated Evaluation of Text and Discourse with Coh-Metrix*. Cambridge, Massachusetts: Cambridge University Press.
- McNamara, D. S., Louwerse, M. M., McCarthy, P. M. and Graesser, A. C. 2010. Coh-Metrix: Capturing linguistic features of cohesion. *Discourse Processes*, 47(4), pp. 292-330.

- McNamara, D. S. and Magliano, J. 2009. Toward a comprehensive model of comprehension. *IN: Ross, B. (ed.) Psychology of Learning and Motivation* (vol. 51). Burlington, New Jersey: Academic Press, pp. 297-384.
- McNamara, D., Ozuru, Y. and Floyd, R. 2011. Comprehension challenges in the fourth grade: The roles of text cohesion, text genre, and readers' prior knowledge. *International Electronic Journal of Elementary Education*, 4(1), pp. 229-257.
- Mellinger, C. D. and Hanson, T. A. 2017. *Quantitative Research Methods in Translation and Interpreting Studies*. London and New York: Routledge.
- Meppelink, C. S., Smit, E. G., Buurman, B. M. and Van Weert, J. C. 2015. Should we be afraid of simple messages? The effects of text difficulty and illustrations in people with low or high health literacy. *Health Communication*, 30(12), pp. 1181-1189.
- Meyers, L. S., Gamst, G. C. and Guarino, A. J. 2013. *Performing Data Analysis Using IBM SPSS*. Hoboken, New Jersey: Wiley.
- Mićić, S. 2013. Languages of medicine — Present and future. *JAHHR*, 4(7), pp. 217-233.
- Miller, J. 1991. Short report. Reaction time analysis with outlier exclusion: Bias varies with sample size. *The Quarterly Journal of Experimental Psychology Section A*, 43(4), pp. 907-912.
- Millette, V. and Gagné, M. 2008. Designing volunteers' tasks to maximize motivation, satisfaction and performance: The impact of job characteristics on volunteer engagement. *Motivation and Emotion*, 32, pp. 11-22.
- Mitamura, T. and Nyberg, E. 2001. Automatic rewriting for controlled language translation. *IN: Proceedings of the Sixth Natural Language Processing Pacific Rim Symposium (NLPRS 2001). Workshop on Automatic Paraphrasing: Theories and Applications*. Tokyo, Japan, 27-30 November. Available at: <https://bit.ly/2Q62trv> [Accessed 12 December 2018].
- Mitchell, L. 2015. *Community Post-Editing of Machine-Translated User-Generated Content*. PhD thesis. Dublin City University.
- Miyata, R., Hartley, A., Kageura, K. and Paris, C. 2017. Evaluating the usability of a controlled language authoring assistant. *The Prague Bulletin of Mathematical Linguistics*, 108(1), pp. 147-158.
- Murray, T. 2016. Coordinating the complexity of tools, tasks, and users: On theory-based approaches to authoring tool usability. *International Journal of Artificial Intelligence in Education*, 26(1), pp. 37-71.

N

Nguyen-Lu, N., Reide, P. and Yentis, S. M. 2010. 'Do you have a stick in your mouth?' Use of Google Translate as an aid to anaesthetic pre-assessment. *Anaesthesia*, 65(1), pp. 96-97.

Nielsen, J. 2012. *Usability 101: Introduction to Usability*. Available at: <https://bit.ly/1OOHO8T> [Accessed 12 December 2018].

Nilsson, N. L. 2008. A critical analysis of eight informal reading inventories. *The Reading Teacher*, 61(7), pp. 526-536.

Nov, O. and Rao, B. 2008. Technology-facilitated 'Give according to your abilities, receive according to your needs'. *Communications of the ACM*, 51(5), pp. 83-87.

Novillo-Ortiz, D., Hernández-Pérez, T. and Saigí-Rubió, F. 2017. Availability of information in Public Health on the Internet: An analysis of national health authorities in the Spanish-speaking Latin American and Caribbean countries. *International Journal of Medical Informatics*, 100, pp. 46-55.

Nyberg, E., Mitamura, T. and Huijsen, W. 2003. Controlled language for authoring and translation. IN: Somers, H. (ed.) *Computers and Translation: A Translator's Guide*. Amsterdam: John Benjamins, pp. 245-281.

O

O'Brien, S. 2003. Controlling controlled English: An analysis of several controlled language rule sets. IN: *Proceedings of the Eighth International Workshop of the European Association for Machine Translation and the Fourth Controlled Language Applications Workshop (EAMT-CLAW)*. Dublin City University, Ireland, 15-17 May. European Association for Machine Translation, pp. 105-114.

O'Brien, S. and Cadwell, P. 2017. Translation facilitates comprehension of health-related crisis information: Kenya as an example. *The Journal of Specialised Translation*, 28, pp. 23-51.

O'Brien, S., Ehrensberger-Dow, M., Hasler, M. and Connolly, M. 2017. Irritating CAT tool features that matter to translators. *Hermes Journal of Language and Communication in Business*, 56, pp. 145-162.

O'Brien, S. and Roturier, J. 2007. How portable are controlled language rules? A comparison of two empirical MT studies. IN: Maegaard, B. (ed.) *Proceedings of the Machine Translation Summit XI*. Copenhagen, Denmark, 10-14 September. European Association for Machine Translation, pp. 345-352.

O'Brien, S. and Schäler, R. 2010. Next generation translation and localization: Users are taking charge. *Translating and the Computer 32*. London, 18-19 November. Available at: <https://bit.ly/2RnXqUb> [Accessed 12 December 2018].

Ogden, C. K. 1930. *Basic English: A General Introduction with Rules and Grammar*. London: Paul Treber and Co., Ltd.

Ojala, M. 2013. Translating health information so people can understand it. *Online Searcher*, 37(3), pp. 58-59.

Olohan, M. 2014. Why do you translate? Motivation to volunteer and TED translation. *Translation Studies*, 7(1), pp. 17-33.

Olohan, M. 2018. Technology themes in translation studies research. *ITI Research Network E-Book 2018. The Human and the Machine*. Available at: <https://bit.ly/2Upvj8T> [Accessed 12 December 2018].

Orfanou, K., Tselios, N. and Katsanos, C. 2015. Perceived usability of learning management systems: Empirical evaluation of the System Usability Scale. *The International Review of Research in Open and Distributed Learning*, 16(2), pp. 227-246.

Orwell, G. 1968. Politics and the English language. IN: Orwell, S. and Angos, I. (eds.) *The Collected Essays, Journalism and Letters of George Orwell*. New York: Harcourt Brace Javanovich, pp. 127-140.

Our Work 2018. Available at: <https://bit.ly/2KBSawq> [Accessed 12 December 2018].

Ozuru, Y., Dempsey, K. and McNamara, D. S. 2009. Prior knowledge, reading skill, and text cohesion in the comprehension of science texts. *Learning and Instruction*, 19(3), pp. 228-242.

P

Pandya, C., McHugh, M. and Batalova, J. 2011. *Limited English Proficiency Individuals in the United States: Number, Share, Growth, and Linguistic Diversity*. Washington, D.C.: Migration Policy Institute.

Papineni, K., Roukos, S., Ward, T. and Zhu, W. J. 2002. BLEU: A method for automatic evaluation of machine translation. IN: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Philadelphia, Pennsylvania, 6-12 July. Stroudsburg, Pennsylvania: Association for Computational Linguistics, pp. 311-318.

Park, E., Cho, M. and Ki, C. 2009. Correct use of repeated measures analysis of variance. *The Korean Journal of Laboratory Medicine*, 29(1), pp. 1-9.

Parker, R. M., Baker, D. W., Williams, M. V. and Nurss, J. R. 1995. The test of functional health literacy in adults. *Journal of General Internal Medicine*, 10(10), pp. 537-541.

Parra Escartín, C., O'Brien, S., Goulet, M. and Simard, M. 2017. Machine translation as an academic writing aid for medical practitioners. *IN: Kurohashi, S. and Fung, P. (eds.) Proceedings of Machine Translation Summit XVI, vol. 1.* Nagoya, Japan, 18-22 September. Available at: <https://bit.ly/2w6Kk4g> [Accessed 12 December 2018].

Patel, V. L. and Kaufman, D. R. 2006. Cognitive science and biomedical informatics. *IN: Shortliffe, E. H. and Cimino, J. J. (eds.) Biomedical Informatics: Computer Applications in Health Care and Biomedicine (3rd ed.).* New York: Springer, pp. 133-185.

Perrier, L., Kealey, M. R. and Straus, S. E. 2014. A usability study of two formats of a shortened systematic review for clinicians. *BMJ Open*, 4(12). Available at: <https://bit.ly/2Qo7vDc> [Accessed 12 December 2018].

Petersen, S. E. and Ostendorf, M. 2007. Text simplification for language learners: A corpus analysis. *IN: Proceedings of the Workshop on Speech and Language Technology in Education (SLaTE2007).* Farmington, Pennsylvania, 1-3 October. Available at: <https://bit.ly/2zxn8hf> [Accessed 12 December 2018].

Petrie, H. and Kheir, O. 2007. The relationship between accessibility and usability of websites. *IN: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI 2007).* San Jose, California, 30 April-3 May. New York: Association for Computing Machinery, pp. 397-406.

Pevalin, D. and Robson, K. 2009. *The Stata Survival Manual.* New York: McGraw-Hill Education.

Peveryly, S. T., Ramaswamy, V., Brown, C., Sumowski, J., Alidoost, M. and Garner, J. 2007. What predicts skill in lecture note taking? *Journal of Educational Psychology*, 99(1), pp. 167-180.

Plain English Campaign 2011. *How to Write Medical Information in Plain English.* Available at: <https://bit.ly/2DtdY9P> [Accessed 12 December 2018].

Q

Quispersaravia, A., Perez, W., Sobrevilla, M. and Alva-Manchengo, F. 2016. Coh-Matrix-Esp: A complexity analysis tool for documents written in Spanish. *IN: Calzolari, N., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J. and Piperidis, S. (eds.) Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016).* Portorož, Slovenia, 23-28 May. Paris, France: European Language Resources Association, pp. 4694- 4698.

R

- Rada, G., Pérez, D. and Capurro, D. 2013. Epistemonikos: A free, relational, collaborative, multilingual database of health evidence. *Studies in Health Technology and Informatics*, 192, pp. 486-490.
- Rai, M. K., Loschky, L. C. and Harris, R. J. 2014. The effects of stress on reading: A comparison of first-language versus intermediate second-language reading comprehension. *Journal of Educational Psychology*, 107(2), pp. 348-363.
- Rathbone, J., Hoffmann, T. and Glasziou, P. 2015. Faster title and abstract screening? Evaluating Abstrackr, a semi-automated online screening program for systematic reviewers. *Systematic Reviews*, 4(80). doi:10.1186/s13643-015-0067-6.
- Reed, D. K. and Vaughn, S. 2012. Retell as an indicator of reading comprehension. *Scientific Studies of Reading*, 16(3), pp. 187-217.
- Reeder, R. W., Karat, C. M., Karat, J. and Brodie, C. 2007. Usability challenges in security and privacy policy-authoring interfaces. IN: Baranauskas, C., Palanque, P., Abascal, J. and Barbosa, S. D. (eds.) *Proceedings of the 11th IFIP TC 13 International Conference on Human-Computer Interaction*. Rio de Janeiro, Brazil, 10-14 September. Berlin: Springer, pp. 141-155.
- Rello, L., Baeza-Yates, R., Dempere-Marco, L. and Saggion, H. 2013. Frequent words improve readability and short words improve understandability for people with dyslexia. IN: Kotzé, P., Marsden, G., Lindgaard, G., Wesson, J. and Winckler, M. (eds.) *Proceedings of the 14th IFIP TC 13 International Conference on Human-Computer Interaction*. Cape Town, South Africa, 2-6 September. Berlin: Springer, pp. 203-219.
- Rello, L., Saggion, H., Baeza-Yates, R. and Graells, E. 2012. Graphical schemes may improve readability but not understandability for people with dyslexia. IN: Williams, S., Siddharthan, A. and Nenkova, A. (eds.) *Proceedings of the First Workshop on Predicting and Improving Text Readability for Target Reader Populations*. Montréal, Canada, 7 June. Stroudsburg, Pennsylvania: Association for Computational Linguistics, pp. 25-32.
- Renahy, E., Parizot, I. and Chauvin, P. 2008. Health information seeking on the Internet: A double divide? Results from a representative survey in the Paris metropolitan area, France, 2005-2006. *BMC Public Health*, 8(69). doi:10.1186/1471-2458-8-69.
- Renahy, J., Devitre, D., Thomas, I. and Dziadkiewicz, A. 2009. Controlled language norms for the redaction of security protocols: Finding the median between system needs and user acceptability. IN: *Proceedings of the 11th International Symposium on Social Communication (ISSC 2009)*. Santiago de Cuba, Cuba, 19-23 January. Available at: <https://bit.ly/2QqyC1f> [Accessed 12 December 2018].

- Renahy, J., Vuitton, D. A., Rath, B., Thomas, I., De Grivel, V. and Cardey, S. 2015. Controlled language and information on vaccines: Application to package inserts. *Current Drug Safety*, 10(1), pp. 41-48.
- Reuther, U. and Schmidt-Wigger, A. 2000. Designing a multi-purpose CL application. *IN: Proceedings of the Third International Workshop on Controlled Language Applications (CLAW 2000)*. Seattle, Washington, 29-30 April. Available at: <https://bit.ly/2QtPEKU> [Accessed 12 December 2018].
- Richards, T., Peverly, S., Wolf, A., Abbott, R., Tanimoto, S., Thompson, R., Nagy, W. and Berninger, V. 2016. Idea units in notes and summaries for read texts by keyboard and pencil in middle childhood students with specific learning disabilities: Cognitive and brain findings. *Trends in Neuroscience and Education*, 5(3), pp. 146-155.
- Riley, G. L. and Lee, J. F. 1996. A comparison of recall and summary protocols as measures of second language reading comprehension. *Language Testing*, 13(2), pp. 173-189.
- Risku, H., Milošević, J. and Pein-Weber, C. 2016. Writing vs. translating: Dimensions of text production in comparison. *IN: Muñoz Martín, R. (ed.) Reembedding Translation Process Research*. Amsterdam: John Benjamins, pp. 47-68.
- Rodríguez Vázquez, S. 2016. *Assuring Accessibility during Web Localisation: An Empirical Investigation on the Achievement of Appropriate Text Alternatives for Images*. PhD thesis. University of Geneva.
- Roediger, H. L. and Marsh, E. J. 2005. The positive and negative consequences of multiple-choice testing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31(5), pp. 1155-1159.
- Roseblat, G., Logan, R., Tse, T. and Graham, L. 2006. Text features and readability: Expert evaluation of consumer health text. *IN: Proceedings of the 11th World Congress on Internet in Medicine (MEDNET '06)*. Toronto, Canada, 13-20 October. Available at: <https://bit.ly/2DRz8xG> [Accessed 12 December 2018].
- Roturier, J. 2006. *An Investigation into the Impact of Controlled English Rules on the Comprehensibility, Usefulness and Acceptability of Machine-translated Technical Documentation for French and German Users*. PhD thesis. Dublin City University.
- Roturier, J. 2009. Deploying novel MT technology to raise the bar for quality: A review of key advantages and challenges. *IN: Proceedings of the Machine Translation Summit XII*. Ottawa, Canada, 26-30 August. Available at: <https://bit.ly/2yx2bDh> [Accessed 12 December 2018].
- Rubin, J. and Chisnell, D. 2008. *Handbook of Usability Testing: How to Plan, Design and Conduct Effective Tests* (2nd ed.). Indianapolis, Indiana: Wiley Publishing, Inc.

Rudd, R. E., Kaphingst, K., Colton, T., Gregoire, J. and Hyde, J. 2004. Rewriting public health information in plain language. *Journal of Health Communication*, 9(3), pp. 195-206.

Rusko, E., Van der Waarde, K. and Heiniö, R. L. 2012. Challenges to read and understand information on pharmaceutical packages. *IN: Singh, J. (ed.) Proceedings of the 18th IAPRI World Packaging Conference*. California Polytechnic State University, California, 17-21 June. Lancaster, Pennsylvania: DEStech Publication, pp. 79-85.

Ryan, R. M., Patrick, H., Deci, E. L. and Williams, G. C. 2008. Facilitating health behaviour change and its maintenance: Interventions based on self-determination theory. *European Health Psychologist*, 10, pp. 2-5.

Rychtycky, N. 2002. An assessment of machine translation for vehicle assembly process planning at Ford Motor company. *IN: Richardson, S. D. (ed.) Machine Translation: From Research to Real Users*. Berlin and Heidelberg: Springer, pp. 207-215.

S

Sakai, Y. 2013. The role of readability in effective health communication: An experiment using a Japanese health information text on chronic suppurative otitis media. *Health Information & Libraries Journal*, 30(3), pp. 220-231.

Saldanha, G. 2009. Principles of corpus linguistics and their application to translation studies research. *Tradumàtica: Traducció i Tecnologies de la Informació i la Comunicació*, (7).

Saldanha, G. and O'Brien, S. 2013. *Research Methodologies in Translation Studies*. Manchester and Kinderhook (New York): St. Jerome Publishing.

Santesso, N., Rader, T., Nilsen, E. S., Glenton, C., Rosenbaum, S., Ciapponi, A., Moja, L., Pardo, J. P., Zhou, Q. and Schünemann, H. J. 2015. A summary to communicate evidence from systematic reviews to the public improved understanding and accessibility of information: A randomized controlled trial. *Journal of Clinical Epidemiology*, 68(2), pp. 182-190.

Sato, T. 1996. Type I and Type II error in multiple comparisons. *The Journal of Psychology*, 130(3), pp. 293-302.

Sauro, J. 2011. *A Practical Guide to the System Usability Scale: Background, Benchmarks, & Best Practices*. Denver: Measuring Usability LLC.

Sauro, J. and Lewis, J. R. 2009. Correlations among prototypical usability metrics: Evidence for the construct of usability. *IN: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI 2009)*. Boston, Massachusetts, 4-9 April. New York: Association for Computing Machinery, pp. 1609-1618.

- Sauro, J. and Lewis, J. R. 2011. When designing usability questionnaires, does it hurt to be positive? *IN: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI 2011)*. Vancouver, Canada, 7-12 May. New York: Association for Computing Machinery, pp. 2215-2224.
- Scarton, C., De Oliveira, M., Candido, A., Gasperin, C. and Aluísio, S. M. 2010. SIMPLIFICA: A tool for authoring simplified texts in Brazilian Portuguese guided by readability assessments. *IN: Proceedings of the 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL HLT 2010), Demonstration Session*. Los Angeles, California, 2-4 June. Stroudsburg, Pennsylvania: Association for Computational Linguistics, pp. 41-44.
- Schiefele, U. and Krapp, A. 1996. Topic interest and free recall of expository text. *Learning and Individual Differences*, 8(2), pp. 141-160.
- Schillinger, D., Grumbach, K., Piette, J., Wang, F., Osmond, D., Daher, C., Palacios, J., Sullivan, G. D. and Bindman, A. B. 2002. Association of health literacy with diabetes outcomes. *Journal of the American Medical Association*, 288(4), pp. 475-482.
- Schotter, E. R., Tran, R. and Rayner, K. 2014. Don't believe what you read (only once): Comprehension is supported by regressions during reading. *Psychological Science*, 25(6), pp. 1218-1226.
- Schriver, K. A. 2017. Plain language in the US gains momentum: 1940-2015. *IEEE Transactions on Professional Communication*, 60(4), pp. 343-366.
- Schwitzer, R. 2015. Controlled language. *IN: Sin-wai, C. (ed.) The Routledge Encyclopedia of Translation Technology*. London and New York: Routledge, pp. 450-464.
- Shafee, T., Masukume, G., Kipersztok, L., Das, D., Häggström, M. and Heilman, J. 2017. Evolution of Wikipedia's medical content: Past, present and future. *Journal of Epidemiology and Community Health*, 71, pp. 1122-1129.
- Shapiro, A. 2004. How including prior knowledge as a subject variable may change outcomes of learning research. *American Educational Research Journal*, 41(1), pp. 159-189.
- Shardlow, M. 2014. A survey of automated text simplification. *International Journal of Advanced Computer Science and Applications, Special Issue on Natural Language Processing*, 4(1), pp. 58-70.
- Sheldon, M., Fillyaw, M. and Thompson, D. 1996. The use and interpretation of the Friedman test in the analysis of ordinal-scale data in repeated measures designs. *Physiotherapy Research International*, 1(4), pp. 221-228.
- Shiffman, R. N., Michel, G., Rosenfeld, R. M. and Davidson, C. 2012. Building better guidelines with BRIDGE-Wiz: Development and evaluation of a software assistant to

promote clarity, transparency, and implementability, *Journal of the American Medical Informatics Association*, 19(1), pp. 94-101.

Shin, S., Lidster, R., Sabraw, S. and Yeager, R. 2016. The effects of L2 proficiency differences in pairs on idea units in a collaborative text reconstruction task. *Language Teaching Research*, 20(3), pp. 366-386.

Shiu-Thornton, S., Balabis, J., Senturia, K., Tamayo, A. and Oberle, M. 2007. Disaster preparedness for limited English proficient communities: Medical interpreters as cultural brokers and gatekeepers. *Public Health Reports*, 122(4), pp. 466-471.

Siddharthan, A. 2014. A survey of research on text simplification. *International Journal of Applied Linguistics*, 165(2), pp. 259-298.

Siegel, P., Martin, E. and Bruno, R. 2001. Language use and linguistic isolation: Historical data and methodological issues. *IN: Statistical Policy Working Paper 32: 2000 Seminar on Integrating Federal Statistical Information and Processes*. Washington, D.C.: Federal Committee on Statistical Methodology, Office of Management and Budget, pp. 167-190.

Slotmaker, A., Hummel, H. and Koper, R. 2017. Evaluating the usability of authoring environments for serious games. *Simulation & Gaming*, 48(4), pp. 553-578.

Smith, C. A., Hetzel, S., Dalrymple, P. and Keselman, A. 2011. Beyond readability: Investigating coherence of clinical text for consumers. *Journal of Medical Internet Research*, 13(4). doi:10.2196/jmir.1842.

Smith, H. K. 2018. Review of PubMed Health. *Journal of Consumer Health on the Internet*, 22(1), pp. 80-87.

Smith, M. and Taffler, R. J. 1992. Readability and understandability: Different measures of the textual complexity of accounting narrative. *Accounting, Auditing, and Accountability Journal*, 5(4), pp. 84-98.

Smith, R. 2013. The Cochrane Collaboration at 20: Much has been achieved, but much remains to be done. *BMJ Editorials*, 347(f7383). doi:10.1136/bmj.f7383.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L. and Makhoul, J. 2006. A study of translation edit rate with targeted human annotation. *IN: Proceedings of the Seventh Conference of the Association for Machine Translation in the Americas (AMTA)*. Cambridge, Massachusetts, 8-12 August. The Association for Machine Translation in the Americas, pp. 223-231.

Snow, C. 2002. *Reading for Understanding: Toward a Research and Development Program in Reading Comprehension*. Santa Monica, California: RAND.

Soergel, D., Tse, T. and Slaughter, L. A. 2004. Helping healthcare consumers understand: An “interpretive layer” for finding and making sense of medical information. *Studies in Health Technology and Informatics*, 107(Pt 2), pp. 931-935.

Spaggiari, L., Beaujard, F. and Cannesson, E. 2003. A controlled language at Airbus. *IN: Proceedings of the Eighth International Workshop of the European Association for Machine Translation and the Fourth Controlled Language Applications Workshop (EAMT-CLAW)*. Dublin City University, Ireland, 15-17 May. European Association for Machine Translation, pp. 151-159.

Spaulding, S. 1951. Two formulas for estimating the reading difficulty of Spanish. *Educational Research Bulletin*, 30, pp. 117-124.

Spilich, G., Vesonder, G., Chiesi, H. and Voss, J. 1979. Text processing of domain-related information for individuals with high and low domain knowledge. *Journal of Verbal Learning and Verbal Behavior*, 18(3), pp. 275-290.

SPSS Tutorials: Independent Samples t Test 2018. Available at: <https://bit.ly/2FwvmvU> [Accessed 12 December 2018].

Stableford, S. and Mettger, W. 2007. Plain language: A strategic response to the health literacy challenge. *Journal of Public Health Policy*, 28(1), pp. 71-93.

Štajner, S., Mitkov, R. and Saggion, H. 2014. One step closer to automatic evaluation of text simplification systems. *IN: Williams, S., Siddharthan, A. and Nenkova, A. (eds.) Proceedings of the Third Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)*. Gothenburg, Sweden, 27 April. Stroudsburg, Pennsylvania: Association for Computational Linguistics, pp. 1-10.

Štajner, S. and Popović, M. 2016. Can text simplification help machine translation? *Baltic Journal of Modern Computing*, 4(2), pp. 230-242.

STAR Reading 2008. *Definitions*. Available at: <https://bit.ly/2TTCrKq> [Accessed 12 December 2018].

Stellefson, M. L., Shuster, J. J., Chaney, B. H., Paige, S. R., Alber, J. M., Chaney, J. D. and Sriram, P. S. 2018. Web-based health information seeking and eHealth literacy among patients living with chronic obstructive pulmonary disease (COPD). *Health Communication*, 33(12), pp. 1410-1424.

Strategy to 2020 2013. Available at: <https://bit.ly/2u6Qukm> [Accessed 12 December 2018].

Streiner, D. L. 2016. Control or overcontrol for covariates? *Evidence-Based Mental Health*, 19(1), pp. 4-5.

Struthers, C., Lyddiatt, A. and McIlwain, C. 2014. Commenting on plain language summaries. *IN: Abstracts of the 22nd Cochrane Colloquium. Evidence-Informed Public*

Health: Opportunities and Challenges. Hyderabad, India, 21-26 September. Available at: <https://bit.ly/2BGzndK> [Accessed 12 December 2018].

Stymne, S. 2013. Using a grammar checker and its error typology for annotation of statistical machine translation errors. *IN: Proceedings of the 24th Scandinavian Conference of Linguistics (24SCL)*. Joensuu, Finland, 25-27 August. Available at: <https://bit.ly/2x5HRbr> [Accessed 12 December 2018].

Sullivan, G. and Artino, A. 2013. Analyzing and interpreting data from Likert-type scales. *Journal of Graduate Medical Education*, 5(4), pp. 541-542.

T

Taylor, B. M. and Samuels, S. J. 1983. Children's use of text structure in the recall of expository material. *American Educational Research Journal*, 20(4), pp. 517-528.

Teixeira, C. S. and O'Brien, S. 2017. Investigating the cognitive ergonomic aspects of translation tools in a workplace setting. *Translation Spaces*, 6(1), pp. 79-103.

Temnikova, I. 2012. *Text Complexity and Text Simplification in the Crisis Management Domain*. PhD thesis. University of Wolverhampton.

Terranova, G., Ferro, M., Carpeggiani, C., Recchia, V., Braga, L., Semelka, R. C. and Picano, E. 2012. Low quality and lack of clarity of current informed consent forms in cardiology: How to improve them. *JACC: Cardiovascular Imaging*, 5(6), pp. 649-655.

Tharp, J. B. 1939. The measurement of vocabulary difficulty. *Modern Language Journal*, 24, pp. 169-178.

Thatcher, J., Waddell, C. D., Henry, S. L., Swierenga, S., Urban, M. D., Burks, M., Regan, B. and Bohman, P. 2003. *Constructing Accessible Web Sites*. San Francisco: Glasshaus.

The Cochrane Collaboration 2013. *Standards for the Reporting of Plain Language Summaries in New Cochrane Intervention Reviews (PLEACS)*. Available at: <https://bit.ly/2QsT2pv> [Accessed 12 December 2018].

The Cochrane Collaboration 2016. *Cochrane Style Manual*. Available at: <https://bit.ly/2re0wP8> [Accessed 12 December 2018].

The Translation Strategy Working Group 2014. *Cochrane Translation Strategy and Business Proposal*. Available at: <https://bit.ly/2L0qGQJ> [Accessed 12 December 2018].

Thomas, I., Laroche, L., Plaisantin-Alecu, B., Betbeder, M. L., Seillès, E., Renahy, J., Blagosklonov, O. and Vuitton, D. A. 2015. Computerization of a 'controlled language' to write medical standard operating procedures (SOPs). *Procedia Computer Science*, 64, pp. 95-102.

- Todd, L. and Hoffman-Goetz, L. 2011. Predicting health literacy among English-as-a-second-language older Chinese immigrant women to Canada: Comprehension of colon cancer prevention information. *Journal of Cancer Education*, 26(2), pp. 326-332.
- Tonelli, S., Manh, K. T. and Pianta, E. 2012. Making readability indices readable. *IN: Williams, S., Siddharthan, A. and Nenkova, A. (eds.) Proceedings of the First Workshop on Predicting and Improving Text Readability for Target Reader Populations*. Montréal, Canada, 7 June. Stroudsburg, Pennsylvania: Association for Computational Linguistics, pp. 40-48.
- Tovey, D. and Dellavalle, R. 2010. *Cochrane in the United States of America*. Available at: <https://bit.ly/2KGhOgH> [Accessed 12 December 2018].
- Translated Cochrane Evidence* 2018. Available at: <https://bit.ly/2jZF0Is> [Accessed 12 December 2018].
- Translation at Cochrane: An Introduction* 2016. Available at: <https://bit.ly/2L6ZvDT> [Accessed 12 December 2018].
- Tsafnat, G., Glasziou, P., Choong, M. K., Dunn, A., Galgani, F. and Coiera, E. 2014. Systematic review automation technologies. *Systematic Reviews*, 3(74). doi:10.1186/2046-4053-3-74.
- Tullis, T. S. and Albert, B. 2013. *Measuring the User Experience: Collecting, Analyzing and Presenting Usability Metrics* (2nd ed.). San Francisco: Morgan Kaufmann.
- Tullis, T. S. and Stetson, J. N. 2004. A comparison of questionnaires for assessing website usability. *IN: Proceedings of the 13th Annual Usability Professionals' Association Conference (UPA 2004) — Connecting Communities*. Minneapolis, Minnesota, 7-11 June. Available at: <https://bit.ly/2HlnQ3n> [Accessed 12 December 2018].
- Turchi, M., Negri, M. and Federico, M. 2014. Data-driven annotation of binary MT quality estimation corpora based on human post-editions. *Machine Translation*, 28(3), pp. 281-308.
- Turian, J. P., Shen, L. and Melamed, I. D. 2003. Evaluation of machine translation and its evaluation. *IN: Proceedings of the Machine Translation Summit IX*. New Orleans, Louisiana: 23-27 September. Available at: <https://bit.ly/2OPvaZq> [Accessed 12 December 2018].
- Turner, A. M., Bergman, M., Brownstein, M., Cole, K. and Kirchoff, K. 2014. A comparison of human and machine translation of health promotion materials for public health practice: Time, costs, and quality. *Journal of Public Health Management and Practice*, 20(5), pp. 523-529.

Turner, A. M., Brownstein, M. K., Cole, K., Karasz, H. and Kirchhoff, K. 2015a. Modeling workflow to design machine translation applications for public health practice. *Journal of Biomedical Informatics*, 53, pp. 136-146.

Turner, A. M., Dew, K. N., Desai, L., Martin, N. and Kirchhoff, K. 2015b. Machine translation of public health materials from English to Chinese: A feasibility study. *JMIR Public Health and Surveillance*, 1(2). doi:10.2196/publichealth.4779.

Turner, A. M., Mandel, H. and Capurro, D. 2013. Local health department translation processes: Potential of machine translation technologies to help meet needs. *IN: Proceedings of the AMIA Annual Symposium*. Washington, D.C., 16-20 November. Bethesda, Maryland: American Medical Informatics Association, pp. 1378-1385.

V

Vasey, M. and Thayer, J. 1987. The continuing problem of false positives in repeated measures ANOVA in Psychophysiology: A multivariate solution. *Psychophysiology*, 24(4), pp. 479-486.

Vickstrom, E., Shin, H. B., Collazo, S. G. and Bauman, K. J. 2015. *How Well—Still Good? Assessing the Validity of the American Community Survey English-Ability Question*. Available at: <https://bit.ly/2vv00Az> [Accessed 12 December 2018].

Von Elm, E., Ravaud, P., MacLehose, H., Mbuagbaw, L., Garner, P., Ried, J. and Bonfill, X. 2013. Translating Cochrane reviews to ensure that healthcare decision-making is informed by high-quality research evidence. *PLOS Medicine*, 10(9). doi:10.1371/journal.pmed.1001516.

Vygotsky, L. S. 1978. *Mind in Society: The Development of Higher Psychological Processes*. Cambridge, Massachusetts: Harvard University Press.

W

Warde, F., Papadakos, J., Papadakos, T., Rodin, D., Salhia, M. and Giuliani, M. 2018. Plain language communication as a priority competency for medical professionals in a globalized world. *Canadian Medical Education Journal*, 9(2), pp. e52-e59.

Way, A. 2018. Quality expectations of machine translation. *IN: Moorkens, J., Castilho, S., Gaspari, F. and Doherty, S. (eds.) Translation Quality Assessment: From Principles to Practice*. Basel, Switzerland: Springer International Publishing, pp. 159-178.

Weinstein, N. D., Kwitel, A., McCaul, K. D., Magnan, R. E., Gerrard, M. and Gibbons, F. X. 2007. Risk perceptions: Assessment and relationship to influenza vaccination. *Health Psychology*, 26(2), pp. 146-151.

Weiss, B. D. 2007. *Health Literacy and Patient Safety: Help Patients Understand. Manual for Clinicians* (2nd ed.). American Medical Association Foundation. Available at: <https://bit.ly/2zFZyjl> [Accessed 12 December 2018].

White, J., O'Connell, T. and O'Mara, F. 1994. The ARPA MT evaluation methodologies: Evolution, lessons, and future approaches. *IN: Proceedings of the First Conference of the Association for Machine Translation in the Americas (AMTA)*. Columbia, Maryland, 5-8 October. Available at: <https://bit.ly/2MrSQll> [Accessed 12 December 2018].

Williams, J. and Chesterman, A. 2002. *The Map: A Beginner's Guide to Doing Research in Translation Studies*. Manchester: St. Jerome Publishing.

Wilson, E. A. and Wolf, M. S. 2009. Working memory and the design of health materials: A cognitive factors perspective. *Patient Education and Counseling*, 74(3), pp. 318-322.

Wolk, K. and Marasek, K. 2015. Neural-based machine translation for medical text domain. Based on European Medicines Agency leaflet texts. *Procedia Computer Science*, 64, pp. 2-9.

Wu, C., Xia, F., Deleger, L. and Solti, I. 2011. Statistical machine translation for biomedical text: Are we there yet? *IN: Proceedings of the AMIA Annual Symposium*. Chicago, Illinois, 3-7 November. Bethesda, Maryland: American Medical Informatics Association, pp. 1290-1299.

Wu, X., Li, L., Du, J. and Way, A. 2016. ProphetMT: Controlled language authoring aid system description. *IN: Proceedings of the Sixth International Workshop on Controlled Language Applications (CLAW 2016)*. Portorož, Slovenia, 28 May. Available at: <https://bit.ly/2Qgan5S> [Accessed 12 December 2018].

Y

Yaneva, V. 2015. Easy-read documents as a gold standard for the evaluation of text simplification output. *IN: Proceedings of the Student Research Workshop Associated with the Tenth International Conference on Recent Advances in Natural Language Processing (RANLP 2015)*. Hissar, Bulgaria, 7-9 September. Available at: <https://bit.ly/2E5ftvd> [Accessed 12 December 2018].

Yano, Y., Long, M. H. and Ross, S. 1994. The effects of simplified and elaborated texts on foreign language reading comprehension. *Language Learning*, 44(2), pp. 189-219.

Yatim, M. H. 2008. Usability and fun evaluation of a game authoring tool. *IN: Luca, J. and Weippl, E. R. (eds.) Proceedings of the World Conference on Educational Media and Technology (ED-Media 2008)*. Vienna, Austria, 3 June. Waynesville, North Carolina: Association for the Advancement of Computing in Education, pp. 1504-1511.

Yesilada, Y., Brajnik, G., Vigo, M. and Harper, S. 2012. Understanding web accessibility and its drivers. *IN: Proceedings of the International Cross-Disciplinary Conference on Web Accessibility (W4A 2012)*. Lyon, France, 16-17 April. New York: Association for Computing Machinery. doi:10.1145/2207016.2207027.

Yimam, S. M. and Biemann, C. 2018. Par4Sim — Adaptive Paraphrasing for Text Simplification. *IN: Bender, E. M., Derczynski, L. and Isabelle, P. (eds.) Proceedings of the 27th International Conference on Computational Linguistics (COLING 2018)*. Santa Fe, New Mexico, 20-26 August. Stroudsburg, Pennsylvania: Association for Computational Linguistics, pp. 331-342.

Z

Zamanian, M. and Heydari, P. 2012. Readability of texts: State of the art. *Theory and Practice in Language Studies*, 2(1), pp. 43-53.

Zarcadoolas, C. 2010. The simplicity complex: Exploring simplified health messages in a complex world. *Health Promotion International*, 26(3), pp. 338-350.

Zeng-Treitler, Q., Kim, H., Roseblat, G. and Keselman, A. 2010. Can multilingual machine translation help make medical record content more comprehensible to patients? *Studies in Health Technology and Informatics*, 160(Pt 1), pp. 73-77.

Zhelev, Z., Garside, R. and Hyde, C. 2013. A qualitative study into the difficulties experienced by healthcare decision makers when reading a Cochrane diagnostic test accuracy review. *Systematic Reviews*, 2(32). doi:10.1186/2046-4053-2-32.

Zipf, G. K. 1949. *Human Behavior and the Principle of Least Effort*. Cambridge, Massachusetts: Addison-Wesley Press.

Zwaan, R. A. 1994. Effect of genre expectations on text comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 20(4), pp. 920-933.

Zwaan, R. A. and Radvansky, G. A. 1998. Situation models in language comprehension and memory. *Psychological Bulletin*, 123, pp. 162-185.

Appendices

Table of contents

Appendix A: Research Ethics Committee Letter of Approval for Experiment on Cochrane Authors' Satisfaction

Appendix B: Call for Participation Targeting Cochrane Authors of Plain Language Summaries

Appendix C: Plain Language Statement, Informed Consent Form, and Pre-Task Questionnaire for Cochrane Authors

Appendix D: Questionnaire for Cochrane Authors on Their Typical Interaction and Satisfaction with the Non-Automated Simplification Approach

Appendix E: Instructions for Cochrane Authors on the Installation of TeamViewer

Appendix F: Instructions for Cochrane Authors on Main Editing Task with Acrolinx

Appendix G: Instructions for Cochrane Authors on Warm-Up Task with Acrolinx

Appendix H: Post-Session Questionnaire for Cochrane Authors on Satisfaction Associated with Acrolinx and Future Editing Preferences

Appendix I: Ethical Approval for Reading Comprehension Study Received from Research Ethics Committee (Dublin City University) and Institutional Review Board (Arizona State University)

Appendix J: Informed Consent Form and Background Questionnaire for Arizona State University Students Involved in Reading Comprehension Study

Appendix K: Instructions and Questions Submitted to Arizona State University Students Involved in Reading Comprehension Study

Appendix L: Prior Knowledge Questions Asked to Arizona State University Students after Reading Comprehension Study

Appendix M: Research Ethics Committee Letter of Approval for Experiment on Evaluation of Spanish Machine Translation Output

Appendix N: Call for Participation Targeting Cochrane Health Professionals (Native Speakers of Spanish)

Appendix O: Plain Language Statement, Informed Consent Form, and Background Questionnaire for Cochrane Machine Translation Evaluators

Appendix P: Instructions and Questions for Machine Translation Evaluators

Appendix Q: Follow-Up Email Sent to Cochrane Machine Translation Evaluators

**Appendix A: Research Ethics Committee Letter of Approval for Experiment on
Cochrane Authors' Satisfaction**

Ollscoil Chathair Bhaile Átha Cliath
Dublin City University



Alessandra Rossetti
School of Applied Language and Intercultural Studies
Centre for Translation and Textual Studies

11th October 2016

REC Reference: DCUREC/2016/155
Proposal Title: Usability of editing scenarios for medical texts
Applicant(s): Alessandra Rossetti, Dr Sharon O'Brien

Dear Alessandra,

Further to expedited review, the DCU Research Ethics Committee approves this research proposal.

Materials used to recruit participants should note that ethical approval for this project has been obtained from the Dublin City University Research Ethics Committee.

Should substantial modifications to the research protocol be required at a later stage, a further amendment submission should be made to the REC.

Yours sincerely,

A handwritten signature in blue ink that reads 'Dónal O'Gorman'.

Dr Dónal O'Gorman
Chairperson
DCU Research Ethics Committee



Taighde & Nuálaíocht Tacaíocht
Ollscoil Chathair Bhaile Átha Cliath,
Baile Átha Cliath, Éire

Research & Innovation Support
Dublin City University,
Dublin 9, Ireland

T +353 1 700 8000
F +353 1 700 8002
E research@dcu.ie
www.dcu.ie

Appendix B: Call for Participation Targeting Cochrane Authors of Plain Language Summaries

Opportunity for Cochrane Authors to participate in a Plain Language Summaries project

To all Cochrane authors:

We would like to invite you to take part in a study on the editing process of Cochrane Plain Language Summaries. This study is being conducted by Alessandra Rossetti and Dr Silvia Rodríguez Vázquez, under the supervision of Dr Sharon O'Brien (Dublin City University, Ireland) and within the framework of the European Project INTERACT (International Network on Crisis Translation), in which Cochrane UK is also involved.

We are looking for volunteer health professionals who have experience in collaborating with Cochrane for the production of Plain Language Summaries of Systematic Reviews.

If you accept to participate, you will be asked to check a Plain Language Summary of a Cochrane Systematic Review using Acrolinx, an authoring support tool that will be provided to you along with instructions on how to use it. You will also be asked to complete a short pre-task questionnaire and two short post-task questionnaires.

Your participation in the entire study would take approximately 2 hours of your time. If you decide to participate, you will be given the possibility to take part in this study either remotely or onsite (at Cochrane UK in Oxford).

Taking part in this research study is voluntary, and you may withdraw from the study at any point without repercussion. All the data collected during the study will be treated confidentially.

If you are interested in participating or have any questions, please send us an email within two weeks at either of the addresses below.

Your help would be very much appreciated! Thank you very much!

Alessandra Rossetti (alessandra.rossetti2@mail.dcu.ie)
Silvia Rodríguez Vázquez (silvia.rodriguezvazquez@dcu.ie)

Appendix C: Plain Language Statement, Informed Consent Form, and Pre-Task Questionnaire for Cochrane Authors

Plain Language Statement

This study is being carried out by Alessandra Rossetti (alessandra.rossetti2@mail.dcu.ie) and Dr Silvia Rodríguez Vázquez (silvia.rodriguezvazquez@dcu.ie) under the supervision of Dr Sharon O'Brien. All researchers are affiliated with Dublin City University (Ireland).

As a participant, you will be asked to check a Plain Language Summary of a Cochrane Systematic Review using Acrolinx, an authoring support tool that will be provided to you along with instructions on how to use it. You will also be asked to complete a short pre-task questionnaire (below) and two short post-task questionnaires. Your entire participation will take no longer than 2 hours of your time.

You will be given the possibility to take part in this study either remotely or onsite (at Cochrane UK in Oxford).

We are required by Dublin City University's Research Ethics Committee to provide you with the following additional information concerning your participation in the study:

We anticipate no potential risks to you from involvement in this research study and we anticipate that the collected data cannot be damaging to you in any way. We will make all the necessary arrangements to protect the anonymity and confidentiality of the data during their analysis, dissemination and disposal. Nonetheless, you are advised that the confidentiality of the information provided can only be protected within the limitations of the law.

We anticipate that you might benefit indirectly from this study as our aim is to research possible ways in which Cochrane editing scenarios can be improved, thus facilitating the authors' work and increasing the readability and translatability of Plain Language Summaries.

During this study, the data will be handled exclusively by the two researchers named in this Plain Language Statement. The study is scheduled to be completed by August 2019. You will have the option to have a report of the results on direct request to the researchers.

Your involvement in this research study is voluntary and you may withdraw from the study at any point without repercussion. If you have further questions, please do not hesitate to contact the researchers by sending an email to the addresses provided above.

If you have any concerns about this study and wish to contact an independent person, please contact:

The Secretary, Dublin City University Research Ethics Committee, c/o Research and Innovation Support, Dublin City University, Dublin 9. Tel +353 1 7008000

Thank you in advance,
Alessandra Rossetti
Dr Silvia Rodríguez Vázquez

Informed consent form *

Tick all that apply.

I have read the Plain Language Statement above and I have understood the information provided in it. I have been given the opportunity to ask questions to the researchers by email, and my questions and concerns have been answered by the researchers. I am aware that I will be asked to complete a pre-task questionnaire (below) and two post-task questionnaires. I am also aware that I will be asked to check a Cochrane Plain Language Summary using an authoring support tool called Acrolinx. I am aware that my entire participation will take no more than 2 hours of my time. I am aware that I can participate in this study either remotely or onsite. I confirm that my involvement in the study is voluntary and I am aware that I may withdraw from this study at any point without repercussion. I am aware that my answers are confidential, and I understand that confidentiality of the information provided is subject to legal limitations. I accept that as an individual participant I will not receive any financial compensation. I am aware that the text that I will simplify will be analysed by the researchers in terms of readability and translatability. By ticking this box, you confirm that you have read and understood the information in this section and you consent to take part in this research project.

1. Native language: *

2. Job: *

3. Have you ever produced Plain Language Summaries of Cochrane Systematic Reviews? *

Yes

No (*Skip to question 8*)

4. Could you please write the titles of the Systematic Review(s) for which you have produced Plain Language Summaries? Alternatively, you can copy and paste their URLs. *

5. In which editing environment do you usually produce Plain Language Summaries? *

Microsoft Word

Google Docs

Other: _____

6. When was the last time you produced a Cochrane Plain Language Summary? Please indicate month and year (e.g. January 2017). *

7. Please write your email address below: *

Please note that only the researchers will have access to your email address and this will not be passed on to anyone else for any reason. Please also note that by providing an email address you are waiving your anonymity in the survey, but your data will be processed confidentially and the research will be reported anonymously.

8. Is there anything else that you would like to add?

Appendix D: Questionnaire for Cochrane Authors on Their Typical Interaction and Satisfaction with the Non-Automated Simplification Approach

PLEASE NOTE: For the purposes of this study, by ‘Cochrane Plain Language Summaries (PLS) guidance’ we understand any instructions, recommendations or guidelines you have been provided with by Cochrane regarding the authoring of PLS.

1. Participant ID *

2. Please indicate which type of guidance you have been provided with by Cochrane in order to create PLS of Cochrane’s Systematic Reviews. You can choose more than one option. *

- Cochrane Style Manual (The Cochrane Collaboration)
- PLEACS - Standards for the reporting of Plain language summaries in new Cochrane Intervention Reviews (The Cochrane Collaboration)
- How to write a plain language summary of a Cochrane intervention review (Cochrane Norway)
- Cochrane Handbook for Systematic Reviews of Interventions (The Cochrane Collaboration)
- Cochrane Style Manual section: Simple and accessible English (The Cochrane Collaboration)
- None of the above
- Other: _____

3. Which of the following statements best describes your PLS authoring procedure in terms of use of Cochrane PLS guidance? *

- I never check Cochrane PLS guidance when producing a PLS
- I check Cochrane PLS guidance only once, before starting writing the PLS
- As I write the PLS, I check Cochrane PLS guidance only a couple of times
- As I write the PLS, I check Cochrane PLS guidance multiple times

I first write the PLS and then, once I finished, I check Cochrane PLS guidance to verify that I have not contravened any of the guidelines

Other: _____

4. Do you check PLS guidance every time that you need to write a PLS? *

Yes, for each PLS that I wrote, I checked Cochrane PLS guidance

No, I have only checked Cochrane PLS guidance for the first PLS that I wrote

No, I have checked Cochrane PLS guidance during the production of more than one PLS, but not for all of them

Other: _____

5. Please indicate your level of agreement with the statements below. Please give an answer to all the statements. If you feel that you cannot respond to a specific statement, mark the centre point (i.e. “3”). Record your immediate response to each statement, rather than pondering it for a long time. *

	Strongly disagree (1)	(2)	(3)	(4)	Strongly agree (5)
1. I think that I would like to use Cochrane PLS guidance frequently.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
2. I found Cochrane PLS guidance to be simple.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
3. I thought Cochrane PLS guidance was easy to use.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
4. I think that I could use Cochrane PLS guidance without the support of a technical person.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
5. I found the various documents of Cochrane PLS guidance were well integrated.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
6. I thought there was a lot of consistency in Cochrane PLS guidance.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

7. I would imagine that most people would learn to use Cochrane PLS guidance very quickly.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
8. I found Cochrane PLS guidance very intuitive.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
9. I felt very confident using Cochrane PLS guidance.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
10. I could use Cochrane PLS guidance without having to learn anything new.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

6. Please indicate your level of agreement with the statements below. *

	Strongly disagree (1)	(2)	(3)	(4)	Strongly agree (5)
Cochrane PLS guidance provides enough indications on the type of content to be included in PLS.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Cochrane PLS guidance provides enough indications on the writing style to be followed in PLS.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

7. Is there anything else that you would like to add?

Appendix E: Instructions for Cochrane Authors on the Installation of Team Viewer

TeamViewer – INSTALLATION

TeamViewer is a free piece of software that can be used to remotely access a computer. You will be asked to download TeamViewer and to use it in order to access our computer, where you will find all the materials that you need for the editing study on Cochrane Plain Language Summaries.

Installation

- 1) Please click on the following link to download TeamViewer for free: <https://www.teamviewer.com/en/credentials/free-for-personal-use/>
- 2) Scroll down the page and click on *Download TeamViewer* (see Figure 1).

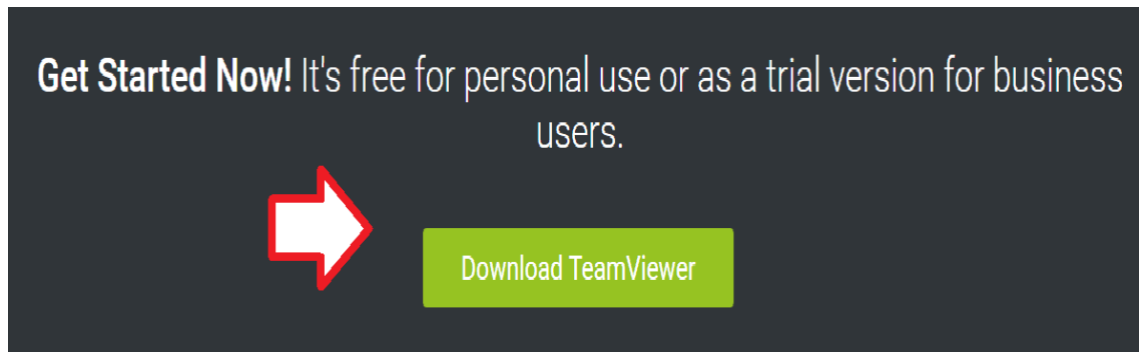


Figure 1: Link for TeamViewer installation

- 3) Once you click on *Download TeamViewer*, you will get an executable file (TeamViewer_Setup.exe).
- 4) Run the .exe file to start the installation wizard. Please select *Basic Installation* and *Personal/Non-commercial use*. Then click on *Accept – finish* (see Figure 2).

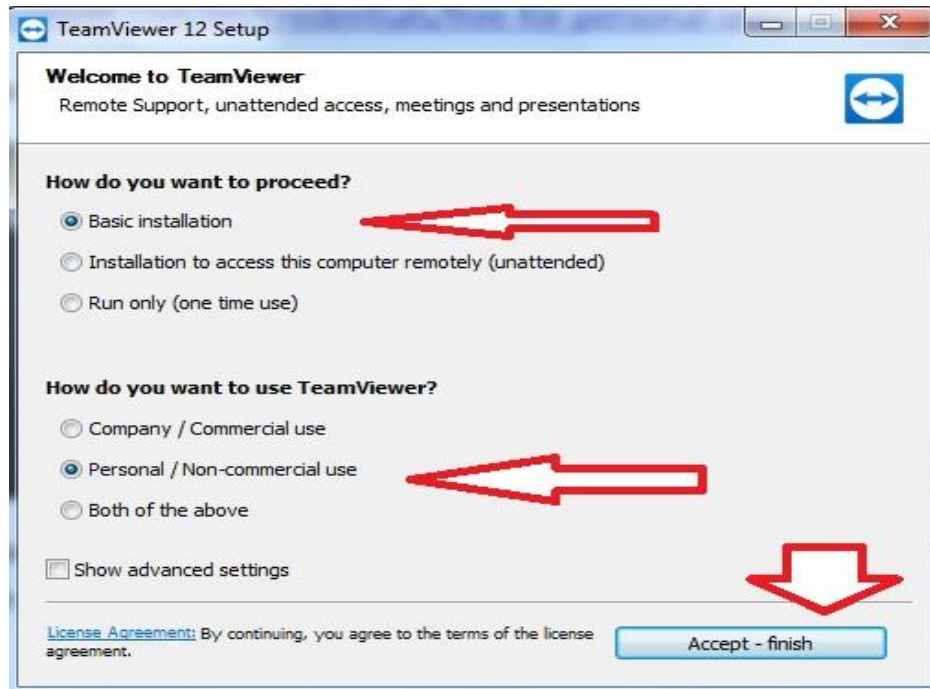


Figure 2: TeamViewer installation wizard

5) Once the installation is complete, the window in Figure 3 will open. Before the main editing study, we will provide you with the Partner ID and the password that you will need in order to access our computer. Now you can close the application.

6) In order to open TeamViewer again, you will just need to click on its icon and the window in Figure 3 will appear on your desktop.

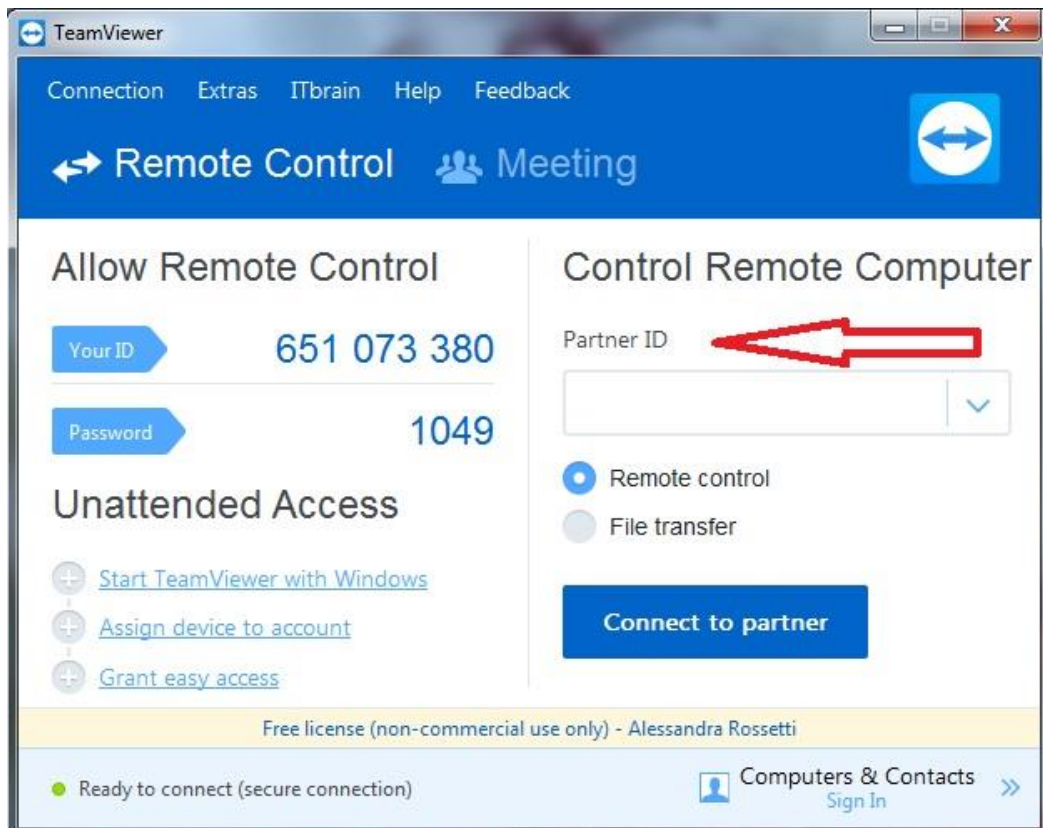


Figure 3: Main TeamViewer window

7) If you have any problems installing TeamViewer or if you have any questions, please send us an email at: alessandra.rossetti2@mail.dcu.ie or silvia.rodriguezvazquez@dcu.ie

8) If possible, please conduct the study using a computer with a screen size of 17 inches or higher. If you have a screen of a smaller size, once you access our computer via TeamViewer, please select: *View > Scale > Original* in order to improve the quality of the visualisation.

Thank you very much!

Appendix F: Instructions for Cochrane Authors on Main Editing Task with Acrolinx

MAIN EDITING TASK - INSTRUCTIONS

For this task, you will be asked to use Acrolinx to:

- 1) check the Plain Language Summary provided for readability; and
 - 2) modify the Plain Language Summary by following Acrolinx suggestions, when applicable
- Please open the Word file called *YourParticipantID_Cochrane_PLS*.
 - You can take your time to read the Plain Language Summary before starting the editing task.
 - When you are ready to start editing, please click on *Review > Acrolinx > Check* (as indicated during the warm up task).
 - Feel free to look at the readability issues that appear in the sidebar for as long as you want. Also, remember that you can open the Scorecard (as shown during the warm up task) to get an overview of the text quality level in terms of readability. However, please do **not** change any of the Acrolinx settings.

Below we remind you of some of the indications provided in the warm-up session:

- Use the *Show actions* button (circled in Figure 1) to see the editing options available.

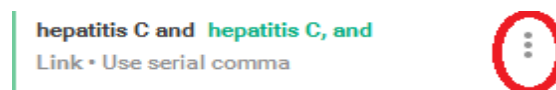


Figure 1: 'Show actions' button

- If you want to see where an issue is located within the text, click on the short explanation below the name of the rule (see Figure 2).



Figure 2: Name of the rule and short explanation

- Try to use the sidebar options to edit the issues (e.g. by automatically replacing

words or ignoring issues). Avoid editing issues directly in the text. If you edit issues directly in the text, the sidebar might not immediately reflect your changes and you will see the flag faded (see Figure 3). If you have to edit the issues directly in the text, make sure that you run another check so that the sidebar reflects your latest changes.

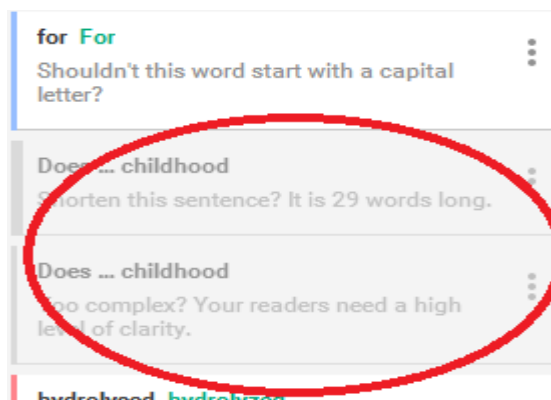


Figure 3: Faded readability flag

- You can run as many checks as you deem necessary by clicking on the button *Check* until you are happy with your new PLS version.
- Feel free to follow the order that you prefer when solving readability issues.
- Always use your common sense in deciding whether to apply a change recommended by Acrolinx or not. As a rule of thumb, refrain from applying Acrolinx suggestions when you believe they would lead to a distortion of meaning and/or to an unnatural style in English.
- While editing the text, feel free to also use any grammar or spelling suggestions that Microsoft Word may provide you with.
- You can work at your own pace – you do not have time limits.
- Once you start the editing task, please avoid any interruptions (e.g. phone calls, email checks, etc.).
- When you finish editing the Plain Language Summary, please **save** the changes that you have made and complete the **short post-session questionnaire** that we sent you in the email. You can now close the TeamViewer session.


Thank you very much for your time and collaboration!

Appendix G: Instructions for Cochrane Authors on Warm-Up Task with Acrolinx

WARM UP TASK - INSTRUCTIONS

Acrolinx is a tool which allows editors to create content that is easier to read. It flags text characteristics that are detrimental to text readability and, when applicable, provides suggestions on how to modify them. It also provides help files with more detailed information on each specific readability issue flagged. This is the tool that you will be asked to use during this session to revise the Plain Language Summary (PLS) you recently produced.

If you have any questions, please do not hesitate to ask the researcher. We will use

Notepad++ as a chat window (). You can use the Notepad file that is already open to type in your queries. The researcher will see your questions on the screen and answer them right away.

- Please open the Word file called *Text_WarmUp_Task*
- Click on the tab *Review* circled in Figure 1.

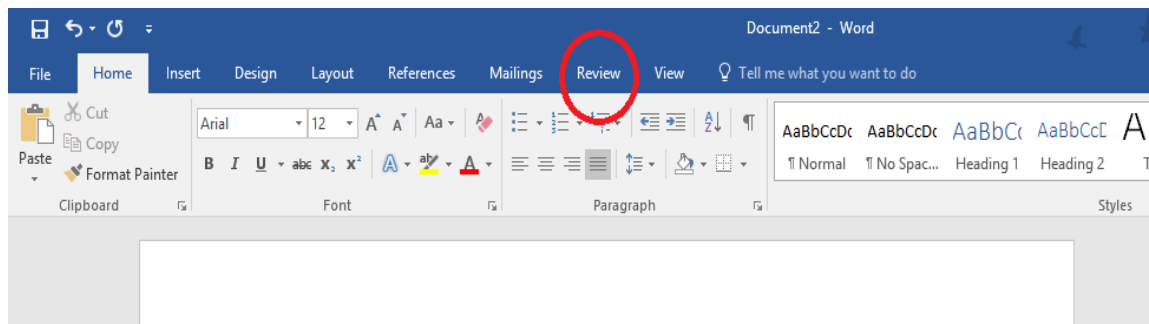


Figure 1: Review tab in Microsoft Word

- Click on *Acrolinx*, as shown in Figure 2.

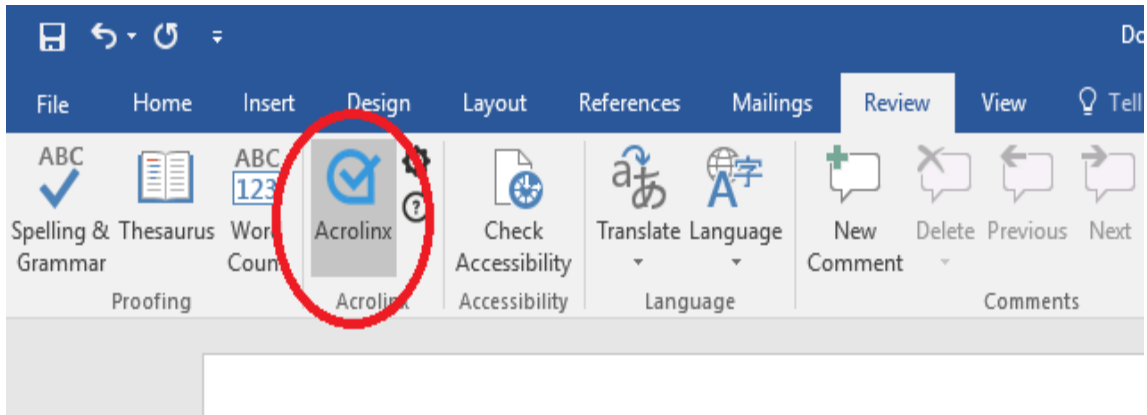


Figure 2: Acrolinx button in Microsoft Word

- A sidebar will appear on the right side of the screen (see Figure 3). Please click on *Check* and wait a few seconds for Acrolinx to check the text for readability issues.

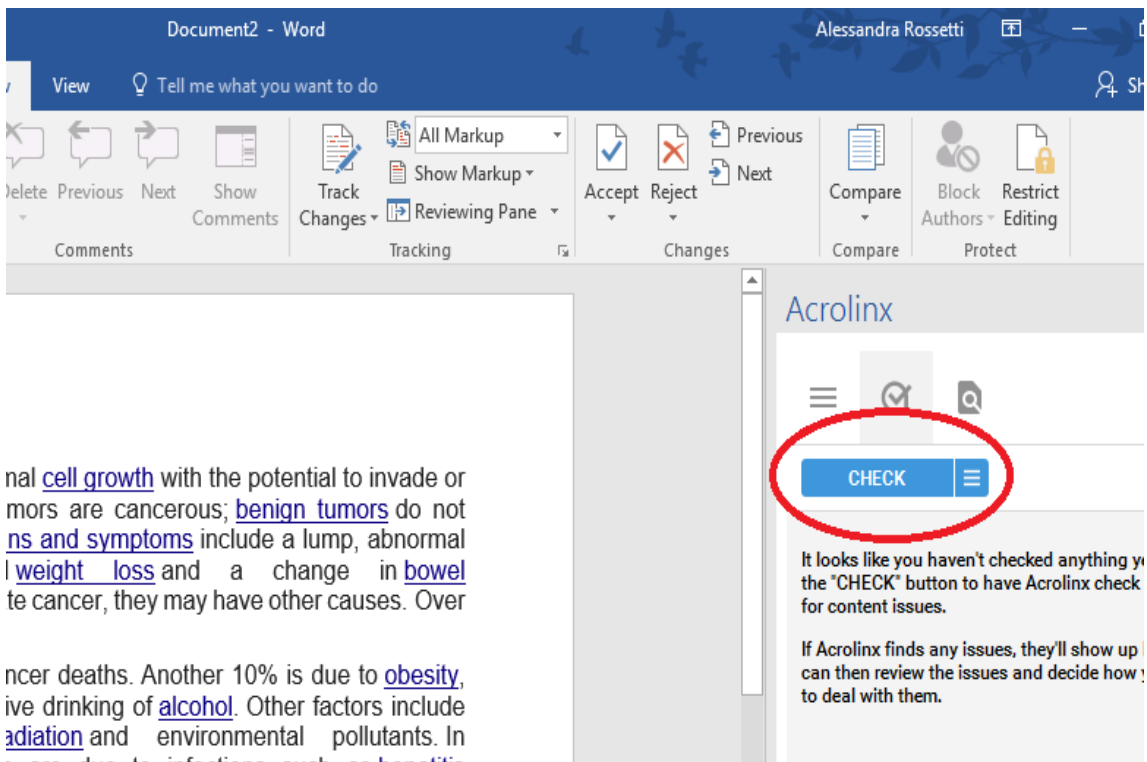


Figure 3: Acrolinx sidebar

- A list of readability issues will appear in the sidebar, under the Results tab, as shown in Figure 4.

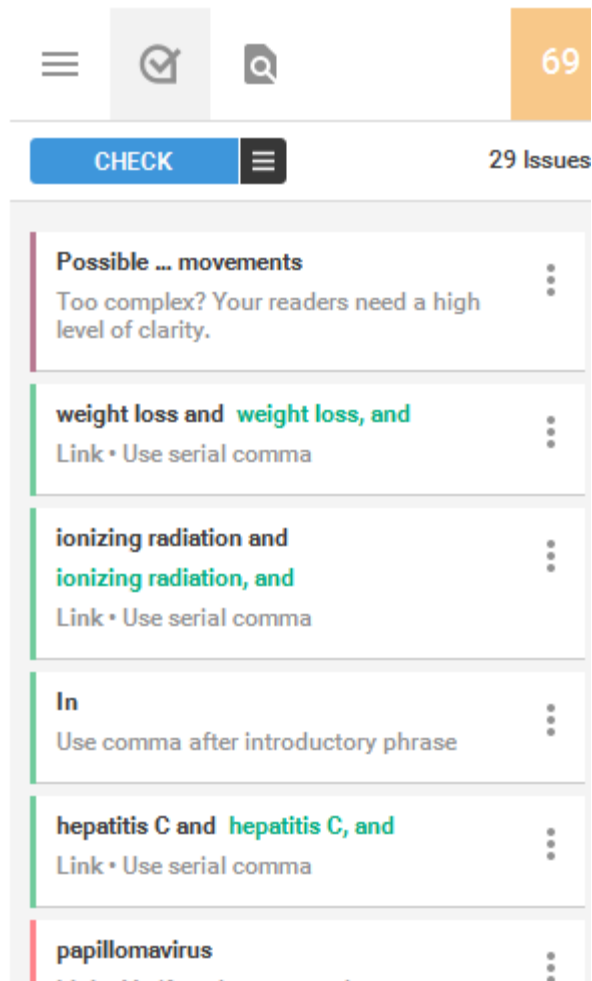


Figure 4: Examples of readability issues flagged by Acrolinx

- If you want to see the categories to which the readability issues flagged belong to, you can do so by clicking on the button circled in Figure 5. Please do **not** change any of the checking settings.

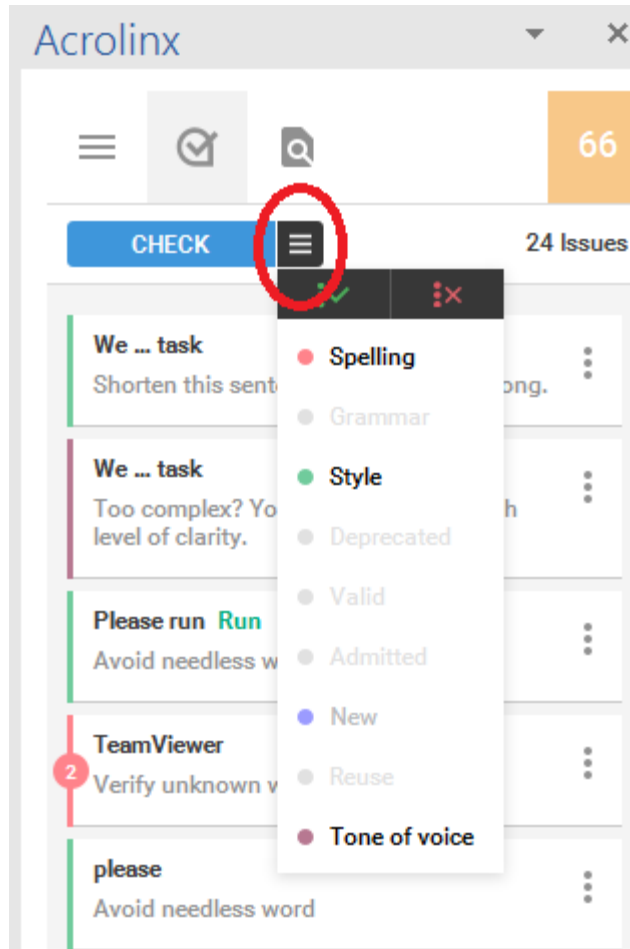


Figure 5: Categories of readability issues

- In the sidebar, use your mouse to move through the issues. To see the editing options for each issue flagged, click on or hover your mouse over the *Show actions* button (circled in Figure 6).

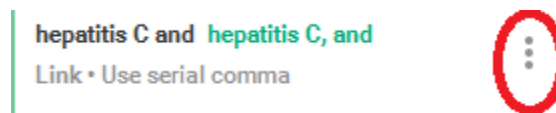


Figure 6: 'Show actions' button

- You will then see different editing options. Please refer to Figure 7.1, Figure 7.2 and the key below to learn about the different editing options available.

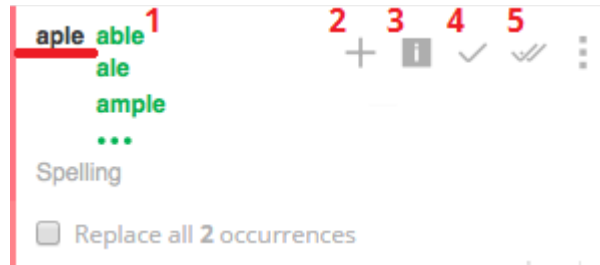


Figure 7.1: Examples of editing options (1)

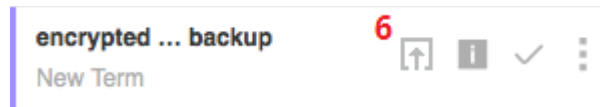


Figure 7.2: Examples of editing options (2)

1 - If you click on any of the suggestions provided in green, you will replace the occurrence with the green word chosen.

2 - If you click on the plus icon, you will add the highlighted word to one of several custom dictionaries (**please ignore this option in this task**).

3 - If you click on the 'i' icon, you will see help information. A window will open which contains a more detailed description of the issue along with relevant examples on how to correct it.

4 - If you click on the single check mark icon, you will ignore the issue. You can also ignore the issue by clicking on the title of the box (word in black underlined in red in the image above).

5 - If you click on the double check mark icon, you will ignore all occurrences of the issue. **Important:** Note that when you ignore an issue, you only have a few seconds to undo this editing action.

6 - If you click on this icon, you will add the term flagged to the terminology database (**please ignore this option in this task**).

- If you want to see where an issue is located within the text, click on the short explanation below the name of the rule (see Figure 8).

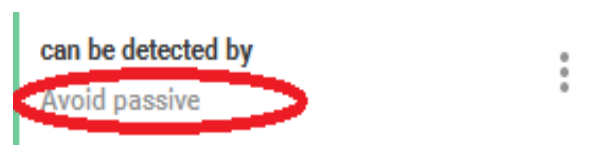


Figure 8: Name of the rule and short explanation

- Whenever possible, try to use the sidebar options to edit the issues. Avoid

editing issues directly in the text. If you edit issues directly in the text, the sidebar might not reflect your changes and you will see the flag faded (see Figure 9). If you have to edit the issues directly in the text, make sure that you run another check so that the sidebar reflects your latest changes.

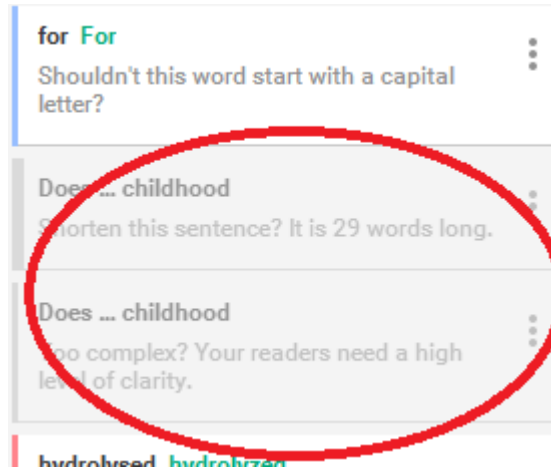


Figure 9: Faded readability flag

- You can always get an overview of the readability level of the text. To do so, open the slide-out menu by clicking on the button circled in Figure 10 and then on *Scorecard* (see Figure 11). The Acrolinx Scorecard contains different sections (e.g. spelling, grammar, style, tone) and provides you with different scores. It also offers a summary of all the issues flagged and improvement suggestions, when available.

Important: The Acrolinx Score expresses the average of all category scores as a standardized score out of 100 that is easier for most users to understand. The higher the score, the higher the quality of the content.

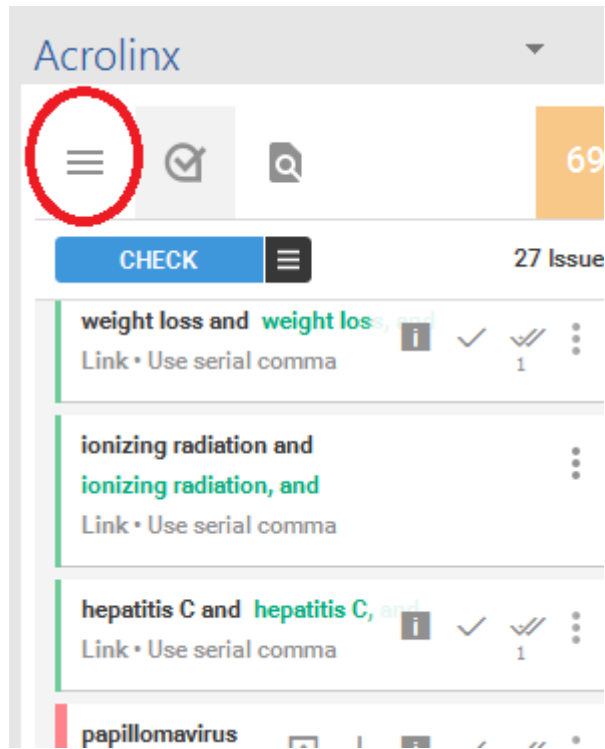


Figure 10: Button to open the slide-out menu

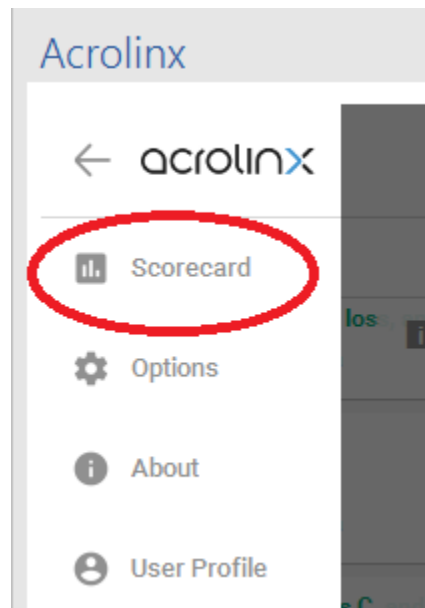


Figure 11: Button to access Acrolinx Scorecard

- Feel free to consult the other menus (i.e. *Options*, *About*, and *User Profile*), but please do **not** change any of the settings neither during the warm-up session nor during the main editing task. The tool has been specifically configured for this task and changing the settings can have a negative impact on the validity of the study.

- Feel free to spend as much time as you need to familiarise yourself with Acrolinx.

Once you are ready, please go to the folder *Main_PLS_Task*.

Appendix H: Post-Session Questionnaire for Cochrane Authors on Satisfaction Associated with Acrolinx and Future Editing Preferences

1. Participant ID *

2. Please give an answer to all the statements. If you feel that you cannot respond to a specific statement, mark the centre point (i.e. “3”). Record your immediate response to each statement, rather than pondering it for a long time. *

	Strongly disagree (1)	(2)	(3)	(4)	Strongly agree (5)
1. I think that I would like to use Acrolinx frequently.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
2. I found Acrolinx to be simple.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
3. I thought Acrolinx was easy to use.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
4. I think that I could use Acrolinx without the support of a technical person.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
5. I found the various functions in Acrolinx were well integrated.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
6. I thought there was a lot of consistency in Acrolinx.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
7. I would imagine that most people would learn to use Acrolinx very quickly.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
8. I found Acrolinx very intuitive.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
9. I felt very confident using Acrolinx.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
10. I could use Acrolinx without having to learn anything new.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

3. If you were to produce a new PLS in the future, which type of authoring support would you use to check the text readability? *

I would use Cochrane PLS guidance only

- I would use Acrolinx only
- I would use both Cochrane PLS guidance and Acrolinx
- I would not use any authoring support material
- Other: _____

4. Please explain the reason(s) for your answer to the previous question *

5. Is there anything else that you would like to add?

Appendix I: Ethical Approval for Reading Comprehension Study Received from Research Ethics Committee (Dublin City University) and Institutional Review Board (Arizona State University)

Ollscoil Chathair Bhaile Átha Cliath
Dublin City University



Ms Alessandra Rossetti
School of Applied Language and Intercultural Studies

03 April 2017

REC Reference: DCUREC/2017/066
Proposal Title: Reading comprehension of health content
Applicant(s): Ms Alessandra Rossetti, Dr Sharon O'Brien

Dear Alessandra,

Further to expedited review, the DCU Research Ethics Committee approves this research proposal.

Materials used to recruit participants should note that ethical approval for this project has been obtained from the Dublin City University Research Ethics Committee.

Should substantial modifications to the research protocol be required at a later stage, a further amendment submission should be made to the REC.

Yours sincerely,

A handwritten signature in blue ink that reads 'Dónal O'Gorman'.

Dr Dónal O'Gorman
Chairperson
DCU Research Ethics Committee



Taighde & Nuálaíocht Tacaíocht
Ollscoil Chathair Bhaile Átha Cliath,
Baile Átha Cliath, Éire

Research & Innovation Support
Dublin City University,
Dublin 9, Ireland

T +353 1 700 8000
F +353 1 700 8002
E research@dcu.ie
www.dcu.ie



EXEMPTION GRANTED

Danielle McNamara
Science of Teaching and Learning, Institute for the (ISTL)
480/727-5690
dsmcnama@asu.edu

Dear Danielle McNamara:

On 7/21/2017 the ASU IRB reviewed the following protocol:

Type of Review:	Initial Study
Title:	Reading comprehension of health content
Investigator:	Danielle McNamara
IRB ID:	STUDY00006514
Funding:	None
Grant Title:	None
Grant ID:	None
Documents Reviewed:	<ul style="list-style-type: none">• INTERACT_InformedConsentOtherParticipants.pdf, Category: Consent Form;• INTERACT_Appendices_7_12_17.pdf, Category: Measures (Survey questions/Interview questions /interview guides/focus group questions);• INTERACT_InformedConsentSubjectPool.pdf, Category: Consent Form;• INTERACT_RecruitingScript.pdf, Category: Recruitment Materials;• INTERACT_IRBProtocol, Category: IRB Protocol;

The IRB determined that the protocol is considered exempt pursuant to Federal Regulations 45CFR46 (2) Tests, surveys, interviews, or observation on 7/21/2017.

In conducting this protocol you are required to follow the requirements listed in the INVESTIGATOR MANUAL (HRP-103).

Appendix J: Informed Consent Form and Background Questionnaire for Arizona State University Students Involved in Reading Comprehension Study

Title: Reading comprehension of English health-related texts

INFORMED CONSENT FORM

Investigators: Alessandra Rossetti, Dr. Sharon O'Brien, and Dr. Danielle McNamara

Contact: arosset3@asu.edu

Department of Psychology – Institute for the Science of Teaching and Learning,
Arizona State University & School of Applied Language and Intercultural Studies,
Dublin City University

I hereby agree to participate as a volunteer in the above named research project.

The purpose of this study is to investigate the impact of different text simplification strategies on the reading comprehension of English health-related texts.

I am aware that I will be asked: 1) to complete a short background questionnaire; 2) to take a reading skills test; 3) to read and answer questions on three health-related texts in English; and 4) to answer some questions on my prior knowledge of health-related topics.

I am aware that this study consists of one session and that my entire participation will take around 2 hours of my time.

I understand that as an individual participant, upon completion of the study I will receive 2 credits from the ASU (Sona Systems) Psychology subject pool.

I confirm that my involvement in the study is voluntary and I am aware that I may withdraw from this study at any point without repercussion. I also understand that if I behave inappropriately or disrupt the research project, I will be asked to leave.

I understand that all information obtained in this study that could identify me will be kept confidential within the limits allowed by law. The information will be kept in a locked filing cabinet and in secure computer files. The specific results of my participation will not be provided to me or to any other persons or institutions.

The information obtained in this study may be published in scientific journals or presented at scientific meetings, but my name and identity will never be included with this information.

I am aware that I will not be at risk of psychological or physical discomfort or harm during the completion of this research. I am aware that Arizona State University does not have any funds budgeted for compensation for injury, damages, or other expenses.

I confirm that I have been given the opportunity to ask questions to the researchers, and that my questions and concerns have been answered by the researchers.

I confirm that I have read and understood the information in this section. I understand that by continuing on to the next page I am agreeing to participate in this research and by doing so, I do not waive any of my legal rights.

*If you have any questions about this project, please do not hesitate to contact the principal investigators, Alessandra Rossetti at arosset3@asu.edu or +1-209-254-2310, or Dr. Danielle S. McNamara at dsmcnamara1@gmail.com or 480-727-5690.

*If you have any questions regarding research participants' rights please contact the Chair of the Committee for the Protection of Human Research Participants at (480) 965-6788.

1. Please write the participant ID that has been provided to you. *

Please answer the following questions as completely and honestly as possible. All of your responses will be confidential.

2. I am a... *

- Male
- Female
- I prefer not to disclose

3. What is your year of birth? *

4. What is the name of the ASU college or school you currently attend? *

5. I am in... *

- 1st year of college
- 2nd year of college
- 3rd year of college
- 4th year or higher of college

6. Is English your first language? *

- Yes (*Skip to question 11*)
- No

7. What is your native language? *

8. What language do you speak at home? *

9. How well do you speak English? *

Very well

Well

Not well

Not at all

10. How many years have you been speaking English? *

Less than 1 year

1 year

2 years

3 years

4 years

5 years

6 years

7 or more years

11. What types of texts do you generally read in English? Please check all that apply. *

E-mails

Letters

Notes

Essays

Research papers

Reports

Stories

Other: _____

Appendix K: Instructions and Questions Submitted to Arizona State University Students Involved in Reading Comprehension Study

Note: The instructions and the free recall questions were the same for all the texts, across all groups of participants. Therefore, they were reported only once in this appendix. The cued recall questions and the rating questions slightly varied depending on the text to be read.

Instructions before reading the text:

In this section, you will be asked to read the [first/second/last] of three texts and answer some questions about it. The text is the summary of a Cochrane Systematic Review. Cochrane Systematic Reviews collate health-related studies on the effects of interventions for prevention, treatment and rehabilitation.

*Please note that you are **not** allowed to take notes on the content of the text while reading.*

Start by clicking on the next button below to access the text. You can spend as much time as you need reading the text.

Instructions after reading the text:

*Once you finish reading the text, click on the next button below to access the comprehension questions. Please answer the comprehension questions as accurately as you possibly can. You are **not** allowed to consult the Internet or any other resource to answer the comprehension questions.*

*Please note that, once you click on the next button below, you will **not** be able to go back to the text.*

Free recall question:

In the box below, please write everything you can remember about the text you just read. Write as much as possible and do not worry about spelling mistakes. You will have four minutes to write. After four minutes, the survey will automatically progress to the next question.

[BOX]

Cued recall questions (Group A):

Abstract:

In the box below, please write everything you can remember about the objectives of the review summarized in the text you just read. Write as much as possible and do not worry

about spelling mistakes. You will have one and a half minute to write. After one and a half minute, the survey will automatically progress to the next question.

[BOX]

In the box below, please write everything you can remember about the authors' conclusions in the text you just read. Write as much as possible and do not worry about spelling mistakes. You will have one and a half minute to write. After one and a half minute, the survey will automatically progress to the next question.

[BOX]

Semi-automated PLS:

In the box below, please write everything you can remember about the quality of the studies described in the text you just read. Write as much as possible and do not worry about spelling mistakes. You will have one and a half minute to write. After one and a half minute, the survey will automatically progress to the next question.

[BOX]

In the box below, please write everything you can remember about the results reported in the text you just read. Write as much as possible and do not worry about spelling mistakes. You will have one and a half minute to write. After one and a half minute, the survey will automatically progress to the next question.

[BOX]

Non-automated PLS:

In the box below, please write everything you can remember about the key results reported in the text you just read. Write as much as possible and do not worry about spelling mistakes. You will have one and a half minute to write. After one and a half minute, the survey will automatically progress to the next question.

[BOX]

In the box below, please write everything you can remember about the quality of the evidence described in the text you just read. Write as much as possible and do not worry about spelling mistakes. You will have one and a half minute to write. After one and a half minute, the survey will automatically progress to the next question.

[BOX]

Cued recall questions (Group B):

Semi-automated PLS:

In the box below, please write everything you can remember about the aim of the review summarized in the text you just read. Write as much as possible and do not worry about spelling mistakes. You will have one and a half minute to write. After one and a half minute, the survey will automatically progress to the next question.

[BOX]

In the box below, please write everything you can remember about the main results presented in the text you just read. Write as much as possible and do not worry about spelling mistakes. You will have one and a half minute to write. After one and a half minute, the survey will automatically progress to the next question.

[BOX]

Non-automated PLS:

In the box below, please write everything you can remember about the quality of the studies described in the text you just read. Write as much as possible and do not worry about spelling mistakes. You will have one and a half minute to write. After one and a half minute, the survey will automatically progress to the next question.

[BOX]

In the box below, please write everything you can remember about the results reported in the text you just read. Write as much as possible and do not worry about spelling mistakes. You will have one and a half minute to write. After one and a half minute, the survey will automatically progress to the next question.

[BOX]

Abstract:

In the box below, please write everything you can remember about the objectives of the review summarized in the text you just read. Write as much as possible and do not worry about spelling mistakes. You will have one and a half minute to write. After one and a half minute, the survey will automatically progress to the next question.

[BOX]

In the box below, please write everything you can remember about the authors' conclusions in the text you just read. Write as much as possible and do not worry about spelling mistakes. You will have one and a half minute to write. After one and a half minute, the survey will automatically progress to the next question.

[BOX]

Cued recall questions (Group C):

Non-automated PLS:

In the box below, please write everything you can remember about the aim of the review summarized in the text you just read. Write as much as possible and do not worry about spelling mistakes. You will have one and a half minute to write. After one and a half minute, the survey will automatically progress to the next question.

[BOX]

In the box below, please write everything you can remember about the main results presented in the text you just read. Write as much as possible and do not worry about spelling mistakes. You will have one and a half minute to write. After one and a half minute, the survey will automatically progress to the next question.

[BOX]

Abstract:

In the box below, please write everything you can remember about the objectives of the review summarized in the text you just read. Write as much as possible and do not worry about spelling mistakes. You will have one and a half minute to write. After one and a half minute, the survey will automatically progress to the next question.

[BOX]

In the box below, please write everything you can remember about the authors' conclusions in the text you just read. Write as much as possible and do not worry about spelling mistakes. You will have one and a half minute to write. After one and a half minute, the survey will automatically progress to the next question.

[BOX]

Semi-automated PLS:

In the box below, please write everything you can remember about the key results reported in the text you just read. Write as much as possible and do not worry about spelling mistakes. You will have one and a half minute to write. After one and a half minute, the survey will automatically progress to the next question.

[BOX]

In the box below, please write everything you can remember about the quality of the evidence described in the text you just read. Write as much as possible and do not worry about spelling mistakes. You will have one and a half minute to write. After one and a half minute, the survey will automatically progress to the next question.

[BOX]

Cued recall questions (Group D):

Semi-automated PLS:

In the box below, please write everything you can remember about the review question summarized in the text you just read. Write as much as possible and do not worry about spelling mistakes. You will have one and a half minute to write. After one and a half minute, the survey will automatically progress to the next question.

[BOX]

In the box below, please write everything you can remember about the conclusions reported in the text you just read. Write as much as possible and do not worry about spelling mistakes. You will have one and a half minute to write. After one and a half minute, the survey will automatically progress to the next question.

[BOX]

Non-automated PLS:

In the box below, please write everything you can remember about the background of the review summarized in the text you just read. Write as much as possible and do not worry about spelling mistakes. You will have one and a half minute to write. After one and a half minute, the survey will automatically progress to the next question.

[BOX]

In the box below, please write everything you can remember about the quality of the evidence reported in the text you just read. Write as much as possible and do not worry about spelling mistakes. You will have one and a half minute to write. After one and a half minute, the survey will automatically progress to the next question.

[BOX]

Abstract:

In the box below, please write everything you can remember about the objectives of the review summarized in the text you just read. Write as much as possible and do not worry about spelling mistakes. You will have one and a half minute to write. After one and a half minute, the survey will automatically progress to the next question.

[BOX]

In the box below, please write everything you can remember about the authors' conclusions in the text you just read. Write as much as possible and do not worry about spelling mistakes. You will have one and a half minute to write. After one and a half minute, the survey will automatically progress to the next question.

[BOX]

Cued recall questions (Group E):

Non-automated PLS:

In the box below, please write everything you can remember about the review question summarized in the text you just read. Write as much as possible and do not worry about spelling mistakes. You will have one and a half minute to write. After one and a half minute, the survey will automatically progress to the next question.

[BOX]

In the box below, please write everything you can remember about the conclusions reported in the text you just read. Write as much as possible and do not worry about spelling mistakes. You will have one and a half minute to write. After one and a half minute, the survey will automatically progress to the next question.

[BOX]

Abstract:

In the box below, please write everything you can remember about the objectives of the review summarized in the text you just read. Write as much as possible and do not worry about spelling mistakes. You will have one and a half minute to write. After one and a half minute, the survey will automatically progress to the next question.

[BOX]

In the box below, please write everything you can remember about the authors' conclusions in the text you just read. Write as much as possible and do not worry about spelling mistakes. You will have one and a half minute to write. After one and a half minute, the survey will automatically progress to the next question.

[BOX]

Semi-automated PLS:

In the box below, please write everything you can remember about the importance of the review summarized in the text you just read. Write as much as possible and do not worry about spelling mistakes. You will have one and a half minute to write. After one and a half minute, the survey will automatically progress to the next question.

[BOX]

In the box below, please write everything you can remember about the evidence from the review summarized in the text you just read. Write as much as possible and do not worry about spelling mistakes. You will have one and a half minute to write. After one and a half minute, the survey will automatically progress to the next question.

[BOX]

Cued recall questions (Group F):

Abstract:

In the box below, please write everything you can remember about the objectives of the review summarized in the text you just read. Write as much as possible and do not worry about spelling mistakes. You will have one and a half minute to write. After one and a half minute, the survey will automatically progress to the next question.

[BOX]

In the box below, please write everything you can remember about the authors' conclusions in the text you just read. Write as much as possible and do not worry about spelling mistakes. You will have one and a half minute to write. After one and a half minute, the survey will automatically progress to the next question.

[BOX]

Semi-automated PLS:

In the box below, please write everything you can remember about the background of the review summarized in the text you just read. Write as much as possible and do not worry about spelling mistakes. You will have one and a half minute to write. After one and a half minute, the survey will automatically progress to the next question.

[BOX]

In the box below, please write everything you can remember about the quality of the evidence reported in the text you just read. Write as much as possible and do not worry about spelling mistakes. You will have one and a half minute to write. After one and a half minute, the survey will automatically progress to the next question.

[BOX]

Non-automated PLS:

In the box below, please write everything you can remember about the importance of the review summarized in the text you just read. Write as much as possible and do not worry

about spelling mistakes. You will have one and a half minute to write. After one and a half minute, the survey will automatically progress to the next question.

[BOX]

In the box below, please write everything you can remember about the evidence from the review summarized in the text you just read. Write as much as possible and do not worry about spelling mistakes. You will have one and a half minute to write. After one and a half minute, the survey will automatically progress to the next question.

[BOX]

Rating question:

Please indicate how strongly you agree or disagree with the statement below:

	Strongly disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Strongly agree
I found the text [TITLE OF THE TEXT] easy to read	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Appendix L: Prior Knowledge Questions Asked to Arizona State University Students after Reading Comprehension Study

You will now answer questions that assess your prior knowledge on a variety of health-related topics. We anticipate that you will not have enough experience with the topics to answer all of the questions correctly. Therefore, simply answer the questions as accurately as you possibly can.

(Groups A, B and C)

Botulinum toxin

- is used as a nerve blocker
- is transmitted from person to person
- is produced by a bacterium called “Escherichia coli”

Cerebrolysin

- might improve behavioral performance
- might decrease cognitive performance
- has not been tested in animal studies

The spinal cord

- is not affected by autoimmune diseases
- is protected by the vertebrae
- is composed of optical fibers

Strabismus

- can be treated by patching the “weaker” eye
- cannot have a genetic component
- can be caused by brain damage

A stroke

- rarely results in the death of neural tissue
- may cause paralysis
- is less common in men than in women under the age of 85

Brain stimulation

- always involves invasive techniques
- is not used for the treatment of mental disorders
- may involve the use of electrodes

Ptosis

- consists in a drooping of the lower eyelid
- can exist from birth
- does not affect vision

Serious adverse events

- are adverse experiences resulting from drug use
- do not include hospitalization
- are never life threatening

Chronic pain

- is often defined as any pain lasting more than two weeks
- may be caused by an infection
- cannot be treated with surgery

(Groups D, E and F)

Patient reported outcomes

- are reported exclusively orally by the patient
- require confirmation from an external observer
- help determine if a patient is eligible for a clinical trial

Aneurysms

- can only occur in the aorta
- do not affect breathing
- might be caused by infections

Hydroxyurea

- may cause a decrease in the number of blood cells in the bone marrow
- may increase the need for blood transfusions in patients with sickle cell anemia
- is not used to treat leukemia

Mental health disorders

- do not include addictive behaviors
- may manifest themselves as physical problems
- are not caused by traumatic events

The aorta

- is connected to the pericardium through the iliac arteries
- carries blood from the heart to the rest of the body
- is the second largest artery in the human body

Sickle cell disease

- is contagious
- may cause chronic pain
- results in a higher number of red blood cells

Depression

- may have a limited number of causes
- is more prevalent in men than in women
- could be influenced by the quantity of serotonin in the brain

Antiplatelet drugs

- prevent the development of blood clots
- are not prescribed to patients who have had a heart attack
- do not include aspirin

Hemoglobin

- transports carbon dioxide from the lungs to the tissues
- is a protein
- does not contain iron

Appendix M: Research Ethics Committee Letter of Approval for Experiment on Evaluation of Spanish Machine Translation Output

Ollscoil Chathair Bhaile Átha Cliath
Dublin City University



Ms Alessandra Rossetti

School of Applied Language and Intercultural Studies

5 October 2017

REC Reference: DCUREC/2017/149

Proposal Title: Machine translatability of health-related texts

Applicant(s): Ms Alessandra Rossetti, Dr Sharon O'Brien

Dear Alessandra,

This research proposal qualifies under our Notification Procedure, as a low risk social research project. Therefore, the DCU Research Ethics Committee approves this project.

Materials used to recruit participants should state that ethical approval for this project has been obtained from the Dublin City University Research Ethics Committee.

Should substantial modifications to the research protocol be required at a later stage, a further amendment submission should be made to the REC.

Yours sincerely,

A handwritten signature in blue ink that reads 'Dónal O'Gorman'.

Dr Dónal O'Gorman
Chairperson
DCU Research Ethics Committee



Taighde & Nuálaíocht Tacalocht
Ollscoil Chathair Bhaile Átha Cliath,
Baile Átha Cliath, Éire

Research & Innovation Support
Dublin City University,
Dublin 9, Ireland

Appendix N: Call for Participation Targeting Cochrane Health Professionals (Native Speakers of Spanish)

Opportunity for Cochrane volunteers

Dear Sir, Madam,

We would like to invite you to take part in a study on the machine translatability of Cochrane plain language summaries. This study is being conducted by Alessandra Rossetti, under the supervision of Dr Sharon O'Brien (Dublin City University, Ireland) and within the framework of the European Project INTERACT (International Network on Crisis Translation), in which Cochrane is also involved.

We are looking for volunteer **health professionals** who are **native speakers of Spanish**.

If you accept to participate, you will be asked to evaluate the fluency and content of two Cochrane plain language summaries that have been machine translated into Spanish. You will also be asked to complete a short background questionnaire.

Your entire participation in this study would take approximately 1 hour of your time. If you decide to participate, you will conduct the task by means of an online survey, on the day and at the time that suit you the most.

This study has been approved by Dublin City University's Research Ethics Committee (DCUREC2017_149). Taking part in this research study is voluntary, and you may withdraw from the study at any point without repercussion. All the data collected during the study will be treated confidentially.

If you are interested in participating or have any questions, please send us an email at: alessandra.rossetti2@mail.dcu.ie

Your help would be very much appreciated! Thank you very much!

Alessandra Rossetti and Dr Sharon O'Brien

Appendix O: Plain Language Statement, Informed Consent Form, and Background Questionnaire for Cochrane Machine Translation Evaluators

Plain Language Statement

Institution: School of Applied Language and Intercultural Studies, Dublin City University

Principal investigators: Alessandra Rossetti and Dr Sharon O'Brien

Purpose of the research: To determine if using a controlled language checker when producing Cochrane plain language summaries increases their machine translatability

This study is being carried out by Alessandra Rossetti (alessandra.rossetti2@mail.dcu.ie) under the supervision of Dr Sharon O'Brien and is part of the H2020 INTERACT project (grant agreement No 734211), in which Cochrane is also involved.

This study will collect data on the quality of the machine translation outputs of Cochrane plain language summaries.

As a participant, you will be asked to complete a short background questionnaire and to read two short Cochrane plain language summaries and their Spanish machine translated versions. Subsequently, you will be asked to evaluate the Spanish machine translated versions.

As a participant, you can take part in this study remotely since the texts will be provided to you by means of an online survey. Your entire participation will take around 1 hour of your time.

We are required by DCU's Research Ethics Committee to provide you with the following additional information concerning your participation in the study:

We anticipate no potential risks to you from involvement in this research study and we will make all the necessary arrangements to protect the anonymity and confidentiality of the data. Each participant will be assigned a number before we start to process the data, so that your identity will never be visible during the analysis and dissemination of results. Data will be disposed of by both researchers in a manner that protects the security and confidentiality of the data five years after the collection of the data has taken place. Nonetheless, you are advised that confidentiality of information provided is subject to legal limitations. It is possible for data to be subject to subpoena, freedom of information claim or mandated reporting by some professions. As we are not assessing your abilities or competencies, but rather the machine translatability of different texts, we anticipate that the collected data cannot be damaging to you in any way.

We anticipate that you might benefit from this study as our aim is to research possible ways in which controlled language checkers and machine translation can be integrated into the workflow of volunteer translators, thus speeding up and facilitating their work.

During this study, the data will be handled exclusively by the two researchers named in this invitation to participate. The study is scheduled to be completed by August 2019. You will have the option to have a detailed, plain language report on direct request to the researchers. Your involvement in this research study is voluntary and you may withdraw from the study at any point without repercussion.

If you have further questions, please do not hesitate to contact the researcher by sending an email to alessandra.rossetti2@mail.dcu.ie

If you have any concerns about this study and wish to contact an independent person, please contact:

The Secretary, Dublin City University Research Ethics Committee, c/o Research and Innovation Support, Dublin City University, Dublin 9. Tel +353 1 7008000

Thank you in advance,

Alessandra Rossetti
Sharon O'Brien

Informed consent form *

Tick all that apply.

- I have read the description of the study and I have understood the information provided in it.
- I have been given the opportunity to ask questions to the researchers by email, and my questions and concerns have been answered by the researchers.
- I am aware that I will be asked to complete a short online background questionnaire.
- I am aware that I will be asked to read and assign scores to the Spanish machine translated outputs of two Cochrane plain language summaries.
- I am aware that I will carry out the scoring of the two texts by means of an online survey.
- I am aware that my entire participation will take around 1 hour of my time.
- I confirm that my involvement in the study is voluntary and I am aware that I may withdraw from this study at any point without repercussion.
- I am aware that my answers are confidential, and I understand that confidentiality of information provided is subject to legal limitations. It is possible for data to be subject to subpoena, freedom of information claim or mandated reporting by some professions.
- I accept that as an individual participant I will not receive any financial compensation.

By ticking this box, I confirm that I have read and understood the information in this section and I consent to take part in this research project.

Please write the participant ID (e.g. E01) that has been provided to you: *

1. Is Spanish your first language? *

Yes

No (*Skip to the end of the survey: Unfortunately, you do not meet the requirements for participating in this study.*)

2. Do you work or are you training to work in the health field? *

Yes

No

3. What is your job? *

4. Do you read medical texts in English? *

Yes

No (*Skip to question 7*)

5. For how many years have you been reading medical texts in English? If less than one year, please indicate how many months (e.g. 8 months). *

6. On average, how many hours per month do you spend reading medical texts in English? Please provide an estimate that is as accurate as possible. *

7. How well do you speak English? *

Very well

Well

Not well

Not at all

8. Please take the short Cambridge English test that is available at <http://www.cambridgeenglish.org/test-your-english/adult-learners/> and report your final score below. It will take you about 5 minutes to complete the Cambridge English test. *

9. If a piece of information is in a language that you do not know, do you ever use machine translation systems such as Google Translate to understand its meaning? *

Yes

No

10. How often do you use machine translation systems? *

Always

Frequently

Rarely

Never

Appendix P: Instructions and Questions for Machine Translation Evaluators

Scoring task 1

You will now be presented with a Cochrane plain language summary in English (source text) and its Spanish machine translated version (target text). The source and the target text will appear segmented at the sentence level.

Instructions: each pair of source sentence-target sentence will be followed by two questions. Please answer both questions for each pair of sentences by assigning scores from 1 to 4. One question will deal with the degree to which the Spanish translation contains the same information that is in the English source sentence, while the other question will deal with the extent to which the Spanish target sentence is correct in Spanish.

Feel free to take as much time as you need to complete this task. You do not have any time limit.

Please note that, in the scoring task, SS indicates the source sentence in English, and TS indicates the target sentence in Spanish. Click on the Next button below to begin the scoring task.

Question on adequacy	None of it (1)	(2)	(3)	All of it (4)
How much of the information contained in the English source sentence (SS) appears in the Spanish target sentence (TS)?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Question on fluency	Incorrect and disfluent (1)	(2)	(3)	Correct and fluent (4)
Indicate the extent to which the Spanish target sentence (TS) is in grammatically well-formed and fluent Spanish.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Scoring task 2

For this last scoring task, you will again be presented with a Cochrane plain language summary in English (source text) and its Spanish machine translated version (target text). The source and the target text will appear segmented at the sentence level.

Instructions: each pair of source sentence-target sentence will be followed by two questions. Please answer both questions for each pair of sentences by assigning scores

from 1 to 4. One question will deal with the degree to which the Spanish translation contains the same information that is in the English source sentence, while the other question will deal with the extent to which the Spanish target sentence is correct in Spanish.

Feel free to take as much time as you need to complete this task. You do not have any time limit.

Please note that, in the scoring task, SS indicates the source sentence in English, and TS indicates the target sentence in Spanish. Click on the Next button below to begin the scoring task.

Question on adequacy	None of it (1)	(2)	(3)	All of it (4)
How much of the information contained in the English source sentence (SS) appears in the Spanish target sentence (TS)?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Question on fluency	Incorrect and disfluent (1)	(2)	(3)	Correct and fluent (4)
Indicate the extent to which the Spanish target sentence (TS) is in grammatically well-formed and fluent Spanish.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Appendix Q: Follow-Up Email Sent to Cochrane Machine Translation Evaluators

Dear [participant's name],

Many thanks again for participating in the study on the machine translatability of Cochrane Plain Language Summaries. I just have one final follow-up question. Do you have any comments on the quality of the machine translated texts? On the variety of Spanish? Or on any other aspect of the task? Any feedback that you might have would be much appreciated.

Thank you very much.

*Best wishes,
Alessandra*

