Tense and tendencies: a linguistic analysis of verbal data from people with low mood

Kate Sayers

Submitted in accordance with the requirements for the degree of doctorate of philosophy in English

University of Leeds School of English, February 2026

Intellectual property and publications statement

I confirm that the work submitted is my own and that appropriate credit has been given where reference has been made to the work of others.

This copy has been supplied on the understanding that it is copyright material and that no quotation from the thesis may be published without proper acknowledgement.

The right of Kate Sayers to be identified as Author of this work has been asserted by Kate Sayers in accordance with the Copyright, Designs and Patents Act 1988.

Acknowledgements

Whilst writing a thesis may ostensibly be a solitary occupation, the folks around me have ensured that I never felt alone. I would like to formally thank each of my friends, from those who listened, distracted, fed and supported, I am grateful to you all.

A wholehearted thank you to Dr Mel Evans for choosing to be my main supervisor and taking this journey with me from the start: from Leicester to Leeds. You have been intellectually inspiring in the best way, as well as articulate, patient and supportive throughout this process. I appreciate all you have done for me.

I began the journey for this thesis at University of Leicester, and thanks go to Dr Alice Welham for choosing to be my second supervisor at the beginning. Her perspective on mental health matters was invaluable. I would also like to thank Dr Mary Ann Lund for her patience and support for supervising me whilst I transferred to Leeds.

Dr Anna Madil kindly agreed to become my second supervisor once I moved to University of Leeds. Thank you for adding your psychologist's perspective, for asking the intellectually stimulating questions and for your encouragement.

Ending on a personal perspective, thank you to Luggage for giving me the confidence to embark upon this journey and bolstering me along the way. To my beloved Matilda, for ensuring I took breaks, left the house and always had plenty of affection. The final thanks go to my Dad, who has supported me in every way possible. I would not be the person I am today were it not for the strength and belief instilled in me by this amazing being, and I will always be grateful for his intellect, beneficence and humour.

The thesis was proof-read before submission by Stephen Sayers. The PGR confirms that the third-party proof-reading undertaken was in accordance with the Postgraduate Researcher Proof-reading Policy.

Abstract

My psychologically grounded linguistic analysis of verbal data investigates the link between language use and low mood. The focus of this study is low mood such as that experienced as part of mood disorders including depression. I address research question one: *What linguistic markers are prevalent in the verbal language of people experiencing low mood?* And research question two: *Do people with low mood focus on the past*?

Existing studies into language and mood disorders frequently rely on written language without demographic metadata. This obfuscates the potential for intersectionality, whilst popular analytical software LIWC is methodologically obtuse.

In this thesis, I cover the distinctive process of data generation including informant recruitment, interviews and transcription alongside categorisation. The focus of this study is the researcher-elicited verbal data from informants with low mood, which is compared to the data from the remaining interviews as a reference.

I analysed the language data for patterns potentially associated with low mood, including stylistic, thematic, syntactic and temporal preponderances. I primarily used corpus linguistic methodologies including keyword and part of speech analysis. I also combined the psychological framework of attribution theory with linguistic appraisal theory.

The findings for research question one include an increased focus on the self and a reduced sense of agency for people experiencing low mood when compared to the reference group. Further to this, low mood informants use a less varied vocabulary when compared to the reference group. The low mood informants are specific about negative things and talk generally about positive things, whilst the reference informants are the opposite. In response to research question two, low mood informants consider their current experience: important events, people and effects to be constant and unchanging. These findings have applications for linguistic approaches to mood data and for health professionals' usage in outreach and amelioration.

Contents

Title page ..... **i**
Intellectual property and publications statement ..... **ii**
Acknowledgements ..... **iii**
Abstract ..... **iv**
Contents ..... **v**
Chapter 1 Tense and tendencies introduction **10**
1.1 Thesis outline 13
Chapter 2 Literature review **16**
2.1 Overview 16
2.2 Applied linguistics 16
2.3 Corpus Linguistics 19
2.4 Psychology of low mood and mood disorders 25
2.5 Psychology and linguistics 32
2.6 Linguistics and psychology 39
2.7 Social characteristics 42
2.8 Temporal focus 45
2.9 Conclusion 50
Chapter 3 Data generation **51**
3.1 Overview 51
3.2 Ethics 51
3.2.1 Overview 51
3.2.2 Informant consent 52
3.2.3 Data protection 53
3.2.4 Detriment mitigation 53
3.2.5 Final points 54
3.3 Positioning the researcher 54
3.4 Pilot study 55
3.4.1 Overview 55
3.4.2 Recruitment and informants 57
3.4.3 Pilot data generation 58
3.4.4 Pilot learning points 61
3.5 Main study 62
3.5.1 Overview 62
3.5.2 Informant description and rationale 64
3.5.3 Informant recruitment 64

3.5.4 Data categorisation 65
3.5.5 Interview schedule 66
3.5.6 Transcription 67
3.6 Reflexivity 69
Chapter 4 Data summary 71
4.1 Overview 71
4.2 Corpus metadata 72
4.3 Demographics 74
4.4 Topics 82
4.5 Narrative 85
4.6 Vocabulary 88
4.7 Data summary review 89
Chapter 5 Analytical methodology 90
5.1 Analytical methodology introduction 90
5.2 Corpus linguistics 90
5.2.1 Keywords 91
5.2.2 Part of Speech tagging 92
5.2.3 Prototypical texts 92
5.3 Theoretical frameworks 96
5.3.1 Attribution theory 96
5.3.2 Appraisal theory 98
5.4 Analytical methodology summary 101
Chapter 6 Word count, content and diversity analysis 102
6.1 Introduction 102
6.2 Word counts 102
6.2.1 Hypothesis testing for word counts 105
6.3 Topics 109
6.4 Lexical diversity 111
6.4.1 Type to token ratio 112
6.4.2 Automated Lexical Diversity analysis 113
6.5 Keywords direct comparison 114
6.6 Keywords BNC comparison 120
6.7 Keywords and social background 124
6.8 Word count, content and diversity analysis discussion 130
Chapter 7 Part of Speech tagging analysis 132
7.1 Introduction 132
7.2 Parts of Speech tagging 133

7.3 Prepositions 137
7.4 General adjectives alongside *be* 140
7.5 Personal pronouns 143
7.6 Modal verbs 150
7.7 Part of Speech tagging analysis discussion 154
Chapter 8 Language and temporal focus 156
8.1 Introduction 156
8.2 Main study CLAWS tense based analysis 156
8.3 Pilot CLAWS tense based analysis 158
8.4 Pilot study tense tracking 160
8.5 Main study tense tracking 164
8.6 Attribution theory and temporal focus 166
8.7 Language and temporal focus discussion 170
Chapter 9 Attribution theory meets appraisal theory 171
9.1 Introduction 171
9.2 Attribution analysis 173
9.3 Appraisal of attributes 179
9.3.1 TLM case study: informant 95 181
9.3.2 RLM case study: informant 128 184
9.4 Attribution meets appraisal discussion 185
Chapter 10 Tense and tendencies discussion 187
10.1 Introduction 187
10.2 The self, agency and power 188
10.3 Lexical diversity 192
10.4 Specificity 194
10.5 Temporal focus 198
10.6 Strengths and limitations 199
10.7 Conclusion 204
References 207
Appendix A 231
Appendix B 238
Appendix C 246
Appendix D 250
Appendix E 253

List of tables

Table 2.1. Key and attributory factors in depression diagnosis comparison 27
Table 2.2. Psychological linguistic features associated with depression 34
Table 2.3. TTD based tags 48
Table 2.4. Summary of linguistic features associated with depression 50
Table 3.1. Pilot study informant classification table 55
Table 4.1. Main study identification of depression by corpus 73
Table 4.2. Comparative summary of main and pilot study informants 75
Table 4.3. Main study informant demographic summary 77
Table 4.4. TLM intersectional informant demographics 80
Table 4.5. RLM intersectional informant demographics 81
Table 4.6. TLM and RLM Topics first, second and third response 83
Table 5.1. Keywords from sampled TLM and RLM against spoken BNC 94
Table 5.2. Appraisal categories 99
Table 6.1 Word count summary for main and pilot study 102
Table 6.2. Chi-squared test for significance of mood against word count 108
Table 6.3. TLM vs RLM keywords 115
Table 6.4. TLM vs spoken BNC and RLM vs spoken BNC keywords 121
Table 6.5. TLM vs BNC keywords not also in RLM vs BNC 122
Table 6.6. Word counts according to demographics 125
Table 6.7. Keywords by demographic; general keywords from full comparison 126
Table 7.1. Key parts of speech TLM compared to RLM 134
Table 7.2. Prepositions from TLM and RLM normalised per 1,000 words 138
Table 7.3. Collocates of PoS prepositions in TLM and RLM 139
Table 7.4. Collocates of be from both TLM and RLM 142
Table 7.5. Collocate in context of be and able across TLM and RLM 143
Table 7.6. Proportional 3rd person pronoun use in TLM and RLM 150
Table 7.7. Modal verbs in TLM and RLM normalised per thousand words 151
Table 7.8. Would epistemic and boulomaic examples in context 152
Table 8.1. Tense based positive and negative key PoS, normalised 157
Table 8.2. TTD notation examples from target corpus 161
Table 8.3. Pilot temporal focus tags percentages by informant 162
Table 8.4. Pilot tense and temporal focus breakdown 163
Table 8.5. TTD for main study prototypical texts 164
Table 8.6. Breakdown of TLM and RLM temporal focus based on TTD 165
Table 8.7. Temporal focus for first response in main study corpora 169
Table 9.1. Visual summary of Attributional codings 174
Table 10.1 Tendencies of low mood psycholect 205

List of figures

Figure 2.1. Concordances for discharge during patient interviews 22
Figure 2.2. Tense tracking diagram 47
Figure 2.3 TTD in use for *The Children of Men* extract 48
Figure 2.4. TTD amended to show my tags in *The Children of Men* extract 48
Figure 3.1. Pilot study data generation flowchart 56
Figure 3.2. Main study data generation flowchart 63
Figure 4.1. Sexuality wordcloud 78
Figure 4.2. TLM and RLM grouped Topics 4th-30th responses 84
Figure 6.1. Main study grouped word count by percentage of informants 103
Figure 6.2. Pilot study grouped word count by percentage of informants 103
Figure 6.3. Main study mood score vs word count 106
Figure 6.4. Word count TLM by gender- male, female & Non-binary 107
Figure 6.5. TLM vs RLM topics of importance 110
Figure 7.1. Proportion of personal pronoun use across TLM or RLM 144
Figure 7.2. First person single and plural pronoun use in TLM and RLM 145
Figure 7.3. Second person pronoun usage within TLM and RLM 148
Figure 7.4. Third person pronoun usage in TLM and RLM 149
Figure 7.5. Collocates of should in TLM and RLM 153
Figure 9.1. Reliability attributional coding for prototypical informants 175
Figure 9.2. Scope of outcome attributional coding for prototypical informants 175
Figure 9.3. Area of origin attributional coding for prototypical informants 176
Figure 9.4. Distinctiveness attributional coding for prototypical informants 177
Figure 9.5. Controllability attributional coding for prototypical informants 177
Figure 9.6. Scope of outcome attribution for all informants' first response 178

Chapter 1 Tense and tendencies introduction

> I had to find a critical framework in which to justify instinct... Then,
>
> afterwards, you look for a way of describing what you've done
>
> (Barker & Brown, 2011, p.154)

This PhD arose from the drive to use insights from linguistics and psychology to help people experiencing depression, alongside my anecdotal observation that friends with depression seemed to focus on the past. The critical frameworks I have chosen are primarily the tools of a linguist, since they offer distinctive and effective insight. As my understanding of the subject matter has improved, the project itself has evolved: I investigate *low mood* rather than *depression*. I develop my methodological approaches, as well as providing new insights into the language use of people with low mood, focussing on linguistic patterns and evidence of temporal focus. The research questions that I address during this thesis are: *What linguistic markers are prevalent in the verbal language of people experiencing low mood?* And *Do people with low mood focus on the past?*

I have chosen to use the term *low mood* in consensus with the concerns of professionals in psychology-related fields including academics around the psychological label of *depression* (as outlined by Kamens et al., 2018). Depression is a condition that is diagnosed based on the presentation of a collection of symptoms over a prolonged period of time (WHO, 2024a). The experience of depression can be debilitating, affecting a person's capacity to maintain relationships, to work and to take part in hobbies. Since depression is a loosely defined collection of symptoms that are often comorbid with other conditions, there is a movement within psychology to cease using this diagnostic label. In line with this emerging viewpoint, I have chosen not to link my work with this *depression* label. Instead, I use the term *low mood*, which is one of the symptoms of depression as it is currently diagnosed (WHO 2024a). The term *low mood* will ensure utility and impact for my work and is generally a more inclusive label than depression.

The title for this thesis is: *Tense and tendencies, a linguistic analysis of verbal data from people with low mood*. This project includes eliciting original verbal data from informants who were then divided,

based on diagnostic profiling (discussed below; Appendix D), into a target group who were experiencing low mood and a reference group. The reference group is the informants who were not experiencing low mood at the time of the interview. I use corpus linguistics (CL) as the primary basis for linguistic analysis in combination with other frameworks. Specifically, I have also used a tense-tracking diagram designed by Svalberg (2019) for second language pedagogy to examine temporal focus. I also used the linguistic framework of appraisal theory (Martin & White, 2007) and the psychological framework of attribution theory (Stratton, 1988).

My first research question is: *What linguistic markers are prevalent in the verbal language of people experiencing low mood?* I have already discussed the choice of the term low mood, as well as the nature of the verbal data. One of the foci of this project was linguistic pattern recognition in order to enable outreach and amelioration techniques. Linguistic markers that indicate low mood, identified through comparison with a control group (the reference corpus) are an additional tool for mental health professionals to use in recognising the kind of distress their patients are experiencing (e.g. Neuman et al., 2012). Linguistic markers indicative of low mood also give insight into the embodied experience of low mood, and can therefore inform therapeutic techniques.

I use CL techniques, particularly keyword and part of speech tagging analysis to search for linguistic markers. CL software enables one to examine and compare large volumes of linguistic data. One of these comparisons is the capacity to search for *keyness*, which denotes items that are statistically more likely to appear in one set of text when compared to another (Groom, 2022). These key items can be individual words, or grammatical categories. A keyword is a word that is characteristically more likely to occur in one set of linguistic data when compared with another. A key part of speech tag is a grammatical category (e.g. *general prepositions* such as *from* or *upon*; *locative adverbs* such as *alongside* or *forward)* that is characteristically more likely to occur in one set of linguistic data when compared to another. Key words or part of speech tags in one set of linguistic data when compared to the other can provide insight into thematic and linguistic patterns and preoccupations. In my data, I use them to contribute towards addressing the first research question: *What linguistic*

markers are prevalent in the verbal language of people experiencing low mood? I do this by comparing the language of the low mood informants with the reference informants.

In this study, I discuss methodological advances based on my data collection and analytics. As part of the investigation addressing my research questions, I contribute insight into the experience of low mood. The firsthand knowledge of low mood is complex, yet language is how we as humans process experience. Halliday and Matthiessen (2004, p.76) tell us that “"language is the reflection of human mind interpreting human experiences". The window into the human mind is therefore the language used to express ourselves. I examine the themes that preoccupy people experiencing low mood, based on their keywords. In addressing my first research question: *What linguistic markers are prevalent in the verbal language of people experiencing low mood?* I compare the verbal data of the low mood informants against the verbal data of the reference group. I find that my low mood informants are more self-focussed and that they do not feel agency over their lives, when compared to the reference informants. In the comparison of the language of low mood and reference informants, I also found that the low mood informants were likely to employ a less varied vocabulary. Further to this, the low mood informants used general language when talking about positive things that are important to them, and specific language when talking about negative things.

I designed my second research question: *Do people with low mood focus on the past?* around anecdotes from friends and published literature written by people with experience of low mood. Whilst the wording of the research question may seem to indicate a binary yes or no answer, I use this as a prompt to linguistically explore temporal focus. Amongst the published literature on low mood is the autobiographical and biographical collection of embodied experiences of depression: *The Noonday Demon: An Atlas of Depression* (Solomon, 2001). A key quote from this book that led to my second research question is: “The worst of depression lies in a present moment that cannot escape the past it idealises or deplores” (Solomon, 2001, p.99). An inescapable idealised or deplored past seemed to fit with my lay person’s idea of low mood. As I will show in the analysis chapters, my data does not support the hypothesis. I find that my low mood informants do not focus

on the past, instead the low mood informants were more likely to consider objects, people and situations to be temporally constant.

Please note that from a linguistic perspective I have chosen to use the gender neutral pronouns *they* and *them* throughout this thesis. I use gender neutral pronouns when referring to everybody, from quoted academics through to my informants, unless gender is directly relevant. The only point at which gender is relevant is when analysing data according to their demographic categories. Whilst I have anecdotally noticed more academics using gender neutral pronouns, I wanted to make it explicit that this is a choice within this work.

1.1 Thesis outline

This first chapter has thus far featured an introduction to the ideas covered in this thesis as well as my motivational underpinnings and key findings. The remainder of this section is an outline of the rest of the thesis, broken down according to section and chapter.

Chapter 2 is a literature review, setting the academic context for this work and revealing the building blocks for the project itself. I begin the literature review with an overview of the project, to set expectations on why the following sections are relevant. I provide context on applied linguistics including what it means and the parts of the discipline particularly relevant to my work. I lay out details about the specific analytical method of corpus linguistics, including how it was developed and what it entails. Having provided the linguistic focus of this project, I develop the psychological grounding. In this section I outline what psychology is, as well as mood disorders and low mood as academics in the discipline comprehend them. I continue by explaining how psychologists have taken a linguistic approach to their subject and what their findings are. I then contrast this with linguists who have taken an interest in psychology and their findings. The literature review is further built with an idea of what social characteristics are, how to define them and why they are important to this project. Finally, I look at temporal focus, to explain both what it is and how it can be studied.

Chapter 3 is the data generation chapter, in which I detail how I generated the data that I use in this study. This chapter includes an outline of the ethical approval process as well as how I am positioned

within the work. I recount my pilot study and how it provided the opportunity to trial my data generation methodology. I build on the pilot study data generation by depicting the main study data generation, which includes the entire process, from informant recruitment through to transcription. The chapter ends with a section on reflexivity, where I reflect on the data generation process and my part in it.

Chapter 4 is a summary of the data generated in chapter 3, particularly the main study data that I use in order to address my research questions. This includes a quantitative statistical and demographic based breakdown of the data. I also provide a qualitative overview of the topical content of my informants' responses to my interview prompts. I finally consider how the informants responded to my interview prompts, in terms of genre and vocabulary

Chapter 5 is a review of the methodology used for data analysis throughout the thesis. This includes detailed information on the practical use of corpus linguistics, particularly keyword searches, part of speech tagging and finding prototypical texts. I complete the analytical methodology review by disclosing the practical uses of the theoretical frameworks of appraisal and attribution.

Chapter 6 is where the analysis begins, with a chapter devoted to exploring the data. I explore the relevance of the number of words used by each informant and potential effects of the word count on subsequent analytical approaches. I also include a comparative topical breakdown and look at the difference in lexical diversity between the low mood and reference informants. I continue by talking through keywords based on a comparison of the low mood and reference corpora, as well as based on a comparison of each corpora with a publicly available external corpora. I then examine the impact of social background on keywords, and therefore the impact of demographics on language patterns.

Chapter 7 is devoted to the analysis of patterns, particularly those based around parts of speech. Displaying the results of this part of speech tagging analysis provides an opportunity to examine in detail those that are most salient. I concentrate on the use of *prepositions*, *general adjectives*, the base form of the word *be*, *personal pronouns*, and *modal verbs*. I describe and display how each of these part of speech tags is used within my data as well as the wider ramifications of each.

Chapter 8 is an analysis of the temporal focus of the verbal utterances I have elicited. The results of the automated tense based part of speech tags search are provided for both the main and pilot study data. I also use a tense tracking diagram to examine the pilot study data, and build upon this to examine a sample of the main study data. The temporal focus chapter concludes with the inclusion of an exploration using the psychological theory of attribution.

Chapter 9 is the introduction of a new analytic approach, the combination of a psychological and a linguistic framework. I examine my prototypical texts using the psychological attribution theory, then concentrate on a part of the theory that is relevant to my findings thus far. I then perform a case study on the verbal language of a low mood informant and a comparable reference informant to explore a combination of sections of appraisal and attribution theory.

Chapter 10 is a discussion of the entire thesis, drawing attention to the important insights gained from this project. I highlight the salient parts of my analysis which both corroborate existing studies and build upon new areas of interest. I particularly concentrate on: the self, agency and power; lexical diversity; specificity; the past. As part of each of these analytical insights I also examine the methodological choices that enabled me to explore these, and the potential impact for other researchers. I devote some time to strengths and limitations of this study before adding a conclusion for the project in its entirety.

Chapter 2 Literature review

2.1 Overview

This is the background to my research questions and the analytic approach chosen to address them. I begin with an introduction to applied linguistics as a discipline and a way of approaching data. This includes information on linguistics generally, what applied linguistics is, as well as information on discourse analysis. From there I outline corpus linguistics as the specific method chosen. In order to ensure this study is grounded in psychological theory, I cover the concept and frameworks for understanding low mood such as that experienced as part of depression. I continue with an overview of where the two disciplines have previously met, beginning with psychological analysis of linguistic trends, with a focus on low mood. From there, I offer a summary of previous corpus linguistic investigations into health phenomena, particularly those related to mental health. One of my motivational underpinnings is an intersectional approach whereby a range of social characteristics are represented, and so I include an outline of the reasons for this. My second research question involves an examination of temporal focus, so I continue by considering the psychological understanding of temporal focus. This literature review is completed with a section on how temporal focus can be linguistically analysed.

2.2 Applied linguistics

Linguistics is defined in the Oxford English Dictionary (2024) as “The scientific study of language and its structure”. The science of linguistics encompasses a range of aspects: from speech production to natural language processing, etymology, grammar and language acquisition. Linguistic theory may involve cross-disciplinary elements, and it is for this reason that it is well suited to an investigation of the language of low mood. An example of a cross-disciplinary area of research is psycholinguistics which includes research into speech production. Speech production within this context is considering not purely what is said (utterances), but the mechanics of speech including

phenomena like words that are stuck on the tip of the tongue (Traxler, 2012, p.45). Psycholinguistics is an intersection of disciplines whilst psychology and linguistics can also be studied within the context of applied linguistics. In this study, I take an applied linguistic approach with a psychological grounding, maintaining separation of these disciplines whilst drawing on elements of both.

Applied linguistics is a subfield of linguistics with a practical focus: research with real world applications (BAAL, 2001, p.1). This practicality may take disparate forms: Simpson (2011) guides us through these varied forms and their meanings which can include language as social identity, forensic linguistics, language pedagogy, multi-cultural correspondence and health communication (Simpson, 2011, p.1). The field of health communication is of particular relevance to this work. This is any type of linguistics related to healthcare, illness and medicine (Collins & Brookes, 2023, p.3). This can be exemplified by Semino et al. (2018) examining metaphors used by patients and healthcare practitioners around cancer. In analysing experiential usage, Semino and their team have developed a ‘metaphor menu’. This booklet empowers patients to choose the terminology that feels most relevant to them after a cancer diagnosis. Taking ownership of language enables patients to define the illness in their own terms, and encourages healthcare professionals to use a vernacular that enfranchises and comforts their patients.

The methodology for my thesis is an example of discourse analysis, which is “the interface between the real world and the inner world of language” (Sinclair, 2004, p.59). To clarify, discourse is the means of communicating our thoughts with the external world. An example of discourse analysis is the analysis of metaphor use in speech as utilised by Semino et al. (2018) in a health communication context. Sinclair (2004, p.19) informs us that the historical definition of discourse incorporated speech acts only. A speech act is “a verbal utterance” (*OED*, 2024), particularly an utterance associated with activity. Over time the concept of discourse has become wider, and also includes means of communication such as writing and sign language. *Discourse* is now a polysemous term used by disciplines outside of linguistics such as psychology (Sims-Schouten & Riley, 2019)

Critical discourse analysis involves examining the “relationship between power and ideology” (Demjén, 2020, p.5). Critical discourse analysis in the context of applied linguistics can take many

forms. Political rhetoric is an example of a discourse type that can be examined from an applied linguistic perspective. This examination may take the form of pattern recognition, such as looking for the context of specific forms of grammar such as modal verbs. This can enable the linguist to examine intended agency and urgency for audience members, as an example (Gage & Perelman, 2011, p.140). The technique for analysis of power construal in political speeches is transferable to my own data. Specifically, I analyse agency and empowerment in the verbal language of my informants.

Systemic Functional linguistics is another branch of discourse analysis, examining how grammar and meaning are associated with discourse (Halliday & Matthiessen, 2013, p.xiii). A subset of systemic functional linguistics is appraisal theory, which is a key linguistic framework that I apply in my investigation of low mood. Appraisal theory, first developed by Martin and White (2007) enables the linguist to scientifically analyse how emotions are linguistically constructed (Bednarek, 2008, p.12) . Grundlingh (2018) outlines that there are three constituent areas of focus for appraisal theory- *Engagement*, *attitude* and *graduation*. To expand, *engagement* is an examination of the source of feelings, this source can be personal or due to external factors. *Attitude* looks at the impact of the emotion, which can be positive or negative and is considered according to *affect*, *judgement* and *appreciation*. Judgement as a sub group of attitude may be personal or modal and includes consideration for morality and ethics. Lastly, *graduation* looks at the intensity or the specificity of these feelings (Martin & Rose, 2003, p.54). By clearly separating out the constituent components that form emotion, this analytical technique is a useful approach to looking at emotions within both fictional and non-fictional texts. In this thesis I will use appraisal theory to examine the construction of emotion in the spoken language of my corpora.

Discourse analysis of speech patterns in particular may involve the examination and categorisation of ways of speaking (Cook, 2011, p.432). There have been studies based on discourse analysis for speech in the context of health communication. One example is Jaworska and Ryan (2018), who examined the way that chronically and terminally ill patients spoke about their pain. The discourses were categorised according to gender, to examine if there were differences between how patients

of different genders spoke about the physical sensations associated with their illness. These researchers purely discuss cis (cisgender i.e. someone who matches their gender at birth (*OED*, 2023)) male and female patients' experiences but did find a contrastive difference in how cis men and women talk about pain. The study found that cis women employ a greater vocabulary when referring to their pain, as well as a greater use of factual terms with regard to pain categorisation when compared to cis men. Jaworska and Ryan (2018, p.107) also found that cis men use more solution focused language, particularly with regards to pain killers, when compared to the cis women in the study.

Discourse analysis can include looking at speech patterns distinct to an individual, known as an idiolect, or those distinct to a geographic area, known as a dialect. It can also involve looking at variations of speech to categorise social groupings, and this is known as a sociolect (Chambers, 1995). This social grouping can include education level, economic background or sexuality. The suffix *lect* in each of these terms comes from the Greek word for "to speak" which is *legein* (Danesi, 1981, p.319). Within this work, I will be examining if it is possible to group discourse according to psychological parameters, and suggest that this could be referred to as a *psycholect*. At the beginning of the 1980's, Danesi (1981, p. 321) suggested that this term could be used to refer to language designed to psychologically influence others. I could find no other references to the word *psycholect* and it certainly has yet to come into popular usage. Therefore I propose that the term *psycholect* instead be used to refer to speech specific to psychological parameters. In this study, I will be searching for linguistic elements that can be used to distinguish between discourses. If there is a distinct psycholect for people experiencing low mood, this could enable outreach, diagnosis or treatment.

2.3 Corpus Linguistics

This section of the literature review will include an overview of what corpus linguistics is and potential applications. Corpus linguistics involves taking a collection of texts, known as a corpus (plural corpora), and undertaking a statistical analysis of language within those texts (Crawford et al., 2017,

p.75). Corpus linguistic techniques were originally developed for use in second language acquisition pedagogy and lexicography, using written and spoken language data (McEnery & Hardie, 2012, p.71). Due to the applicability and utility of these techniques, uptake has since expanded into almost all areas of linguistics as well as disciplines across the humanities and social sciences (Brookes & Collins, 2023, p.15).

Applications for corpus linguistics include, but are not limited to letters sent by Tudor monarchs (Evans, 2017) through to legal processes (D'Hondt & May, 2022) and radio show conversations (Smith, 2020). Fundamentally, it involves the use of computer software to precisely explore patterns and trends in collections of texts. This pattern and trend recognition is feasible on a wider scale than manual coding alone could enable, and may include lexical and semantic grouping and numbering (McEnery & Hardie, 2012, pp.1-3). The texts within present-day corpora are varied in origin and can be as short as a SMS (text message), or span whole newspaper archives, out-of-copyright literature or online databases with billions of words in them (for example the English Corpora collections created by Davies (2004)). The focus of corpus linguistic analysis is the analysis of language which can be considered representative of the type of language under consideration, such as health communication (Brookes & Collins, 2023, p.14).

Corpus linguistic approaches may involve analysis of a chosen corpus (the target) independently or it may be compared and measured against a comparison corpus (the reference corpus). A reference corpus can consist of one text or a collection, and this comparison can enable one to find salient features that are distinct to the target text itself. Within corpus linguistics it is important to find a comparable corpus/corpora for the reference text that is relevant in terms of register and genre (Brookes & Collins, 2023, p.82). As an example, to know more about a speech from a specific politician, it is most effective to compare it to similarly themed speeches from other politicians (e.g. Amaireh, 2023).

The process of corpus linguistic analysis is typically aided and semi-automated through the use of corpus linguistic software, such as *WordSmith* (Scott, 2020); *SketchEngine* (Kilgarriff et al., 2004) or *AntConc* (Anthony, 2010). Corpus linguistic software comes in three forms, downloadable offline

software, online software and researcher-codable DIY software (Brookes & Collins, 2023, p.78). Each of these types of software has distinctive features, and it is possible to use more than one type of software on the same data set. For my own data I primarily use *AntConc* (Anthony, 2010) alongside secondary analysis from *Wmatrix* (Rayson, 2009).

Tools provided by the corpus linguistic software enable analysts to capture dominant and more peripheral linguistic properties within a given corpus. These properties may include word counts, distribution plots, keyness, collocation, Part of Speech (PoS) tagging, topic modelling and sentiment analysis (Brookes & Collins, 2023, p.13). For my study, I use keyness, collocation and PoS tagging and therefore will discuss these in further detail below. As an outline: keyness is how statistically representative something is of the rest of the corpus. Collocation is the tendency of different terms appearing close to one another with greater than random probability. PoS tagging involves the examination of categories of speech rather than simply the words.

There are some corpus linguistic tools that, whilst they have established profiles in research, I do not use in the present study. Firstly: distribution plots visibly display how individual words are apportioned across the corpus. I have chosen not to use this because the distribution of words would not help me to answer my research questions, whereas the style, syntax and theme of those words would. The small scale of my corpus and the precise focus of my research questions and analysis mean that manual qualitative analysis was more practical and insightful than using automated topic modelling and sentiment analysis.

Statistical analysis of large texts is useful, but since meaning in language is found through nuance and connection, context is essential (McEnery & Hardie, 2012, p.3). It is possible to find the context within a text by examining concordance lines. A concordance line is a section of the corpus, a few words either side, of a point of interest that provide context and may also be known as Key Word In Context (KWIC). To give an example from healthcare communication studies, Weetman et al. (2020) used concordances of the word *discharge* from patient interviews to gain insight into the experience of being discharged from a hospital:

it was affecting my kidney I was	**discharged**	after five hours of waiting with the person
one it was (.) not that I was	**discharged**	very quickly because I was out of the
I say to get the medication and	**discharge**	letter it was more than twenty four hours
3 weeks from the date I was	**discharged**	uh before the surgery received any
hanging around to wait for these	**discharge**	papers and that's what I was waiting
a whole day in hospital for that	**discharge**	letter to come from the doctors and my
little writing (.) yeah [DATE] they	**discharged**	me and I didn't have to hang
(.) to be fair when you get	**discharged**	from hospital all you are interested in is
to wait nearly 9 hours for that	**discharge**	letter so that was probably the worst part
going to be ready where is the	**discharge**	letter because we didn't know what the

Figure 2.1: Concordances for *discharge* during patient interviews (Weetman et al., 2020, Table 3, p.5)

In order to ensure this study was applied, the researchers were able suggest different ways for healthcare professionals to communicate anticipated duration of stay to patients. Healthcare professionals are then empowered to understand which expectations need to be set and how best to set those expectations. The results of this study could therefore improve recognition of patient needs and resulting satisfaction levels. Weetman et al. (2020) are an example of how it is possible to use the insights from corpus linguistic techniques to shape the practice of healthcare professionals. The insights from my study also have the capacity to shape clinical and professional practice.

A linchpin of corpus linguistic analysis is keywords: words that are characteristically more likely to be present within a target corpus when compared to a comparison corpus. 'Keyness' is used to indicate representativeness of a corpus and is often a starting point for linguists to delve further into a text (Groom, 2022). For example, Heritage (2022) examined language in a series of video games based on 'The Witcher' novels, comparing the language of games 1, 2 and 3. This involved a huge amount of data which could be efficiently processed using automated techniques and one researcher, rather than a team of hand coders. Heritage (2022) began by finding keywords, then further screened these keywords for gendered social actors. They found the keyword *King* was present for male professions and *Sorceress* was a keyword for female professions (Heritage, 2022). This indicates that within this series a male is more likely to be monarchical whilst a female has a greater chance of being a magic user. It is possible to further examine from there how each type of position is seen and from thence what the general gendered expectations are within the video game series.

The type of wider analysis performed by Heritage (2022) can be done generally through concordance lines, examining the context for each keyword, and can also be performed using collocates. Collocates highlight recurrent patterns for words that cooccur with a focus word, also known as a node, in a corpus. CL software will normally search within the 5 words either side of a node (Baker et al., 2006, p.37). For example, in the Jaworska and Ryan (2018) pain study I have previously mentioned, they searched for the node *pain* in an healthcare corpus. Their search revealed that the words *management* and *excruciating* were statistically likely to appear nearby.

There are statistical methods for calculating both keyness and collocates, and which test to use is an area of significant debate (Groom, 2022). Methods of calculating keyness range from a MI test, to chi Squared, Mann-Whitney U test , T-tests, and log-likelihood. Each have different parameters for calculation and specific situations for optimal usage (Groom, 2022). For example, log-likelihood is based on word frequency, rather than how different the texts are or how dispersed these key terms are within them (Egbert & Biber, 2019). It has been suggested in recent years that a three-pronged approach could be beneficial: looking at effect size, dispersion and association (Groom, 2022). In my corpus, I use the smaller scale chi-squared test for my *AntConc* based key word searches, alongside effect size. Chi-squared is not an option in *Wmatrix* and therefore I use log likelihood for my *Wmatrix* based key part of speech tag searches.

The tests used to check for collocates include t-score, chi-squared, MI, MI3, Z-score or log-likelihood. Each of these is useful in a different scenario, with frequency, proximity and corpus size being among the differentiating factors (Groom, 2022). For this thesis, the combined target and reference corpora are smaller than a large corpus such as the BNC (Davies, 2004) which is a one hundred million word corpus. I use log-likelihood as the likelihood measure and MI for the effect size calculation as part of my collocates searching.

It is possible to explore corpora from a different perspective, enabling a focus on properties such as tense, rather than individual lexical items, through Part of Speech (PoS) tagging (O'Keeffe & McCarthy, 2010, p.35). This can be performed in three different ways: manually by the linguist; using ready-made programs such as Constituent Likelihood Automatic Word-tagging-system (CLAWS)

which is used in *Wmatrix* (Rayson, 2009); as well as user defined tagging using programming languages such as R. For example, CLAWS is able to highlight PoS tags such as *modal verbs*, *prepositions* or *types of pronouns* with around 96% accuracy (Garside, 1987). Once a corpus is marked up according to its part of speech tags, the corpus linguist can see the higher-level patterns in the texts. Higher level pattern recognition can be instructive for our understanding of discursive patterns and linguistic conventions across genres or over time. Hunt and Carter (2012) used this technique when employing corpus led discourse analysis to evaluate the language used by Sylvia Plath in *The Bell Jar* to consider the signifiers, metaphors and context associated with her mood disorder. Deploying these corpus linguistic techniques on the utterances of people experiencing low mood in this project will enable a similar level of insight.

Lexical density and diversity are measures that corpus linguists use to examine the variety of the vocabulary of one corpus when compared to another (Kyle, 2019, p.455). One of the ways of examining lexical density is dividing the total number of distinct words (types) by the total number of words which include repetitions (tokens). This provides the type:token ratio, which is affected by corpus size, but does enable visibility of the lexical richness of data. There are also automated programmes that analyse lexical diversity, such as Measure of Textual Lexical Diversity (MTLD). This calculation analyses the lexical diversity through sampling and is therefore not affected by the size of the corpus being analysed (McCarthy & Jarvis 2010). Researchers have found that low lexical density, i.e. a relatively limited vocabulary for content words, is a potential characteristic of psychosis (Rezaii et al., 2019). Since this measure has been beneficial for understanding linguistic indicators of psychosis, I will perform a lexical density and a lexical diversity test on my own data.

The nature of corpus analytics is empirical, despite the researcher's capacity to choose only the sections of the text that are of direct relevance to their area of interest (Baker & Levon, 2015). This examination of form and constituent parts allows for a close reading of large volumes of data without researcher bias occurring until the interpretation stage. Grounding research and personal experience combine to ensure that the researcher will have preconceived ideas around mental

health. The capacity to reduce both conscious and subconscious bias in the initial stages of this analysis is crucial.

2.4 Psychology of low mood and mood disorders

This section of the literature review will include information about mood disorders, depression and low mood. The original concept behind this project was to investigate the possibility of identifying linguistic markers in the language use, or *discourse*, of people experiencing the symptoms of depression. Following extensive research into psychological literature relating to the classification of mood disorders and conditions associated with depression, I decided to amend this remit. Due to political, social and academic concerns around the potential endurance of the psychological label of *depression*, I have chosen *low mood* as an alternative term. As I provide a definition of depression, I will discuss some of the historical and current issues around this label. For this project, using the term *low mood* will ensure that my work still has relevance even if psychologists change the way they diagnose and talk about the symptoms of depression in the future. To define and understand the issues with *depression* as a linguistic label, it is important to see how it is currently understood. I will use current conventions for diagnosis, treatment and impact to set the psychological foundations to this research.

Depression is a public health concern: over 10% of adults in England were diagnosed as depressed in reporting year 2018-9 (PHE, 2020). People diagnosed with depression report it to be a condition that detrimentally affects quality of life for both the diagnosed and their loved ones (Solomon, 2001). A diagnosis of depression is often associated with comorbid diagnoses, which means two or more conditions diagnosed alongside each other. According to the American National comorbidity study, over 60% of depressed patients have at least one comorbid condition and at least 30% have two or more comorbid psychiatric diagnoses (Otte, 2008). Whilst I was unable to find statistics for comorbid diagnoses in the UK, geopolitical similarities across the nations indicate that the figures should be comparable. The proportional impact of depression and its comorbid diagnoses was one of the

motivating factors for embarking on this research. Depression has an indisputable widespread impact across society in both the UK and overseas.

The Diagnostic Statistical Manual of Mental Disorders (DSM) categorises depression as a Mood Disorder (APA, 2013). The DSM specifies that *mood* in the context of a mood disorder diagnosis includes both depression and mania (APA, 2013). *Euthymia* is the “baseline between the two” (Power, 2004, p.153). The etymology of the term is Greek: ‘eu’ is related to good health whilst ‘thymos’ encompasses drive, emotions, intellect and willpower all bound up in the word for soul (Fava & Guidi, 2020, p.42). A combination of these two concepts leaves us with a word that means wellness of emotion, thought and energy. From a linguistic point of view, this etymological breakdown indicates that it was believed that people who are not euthymic have an unhealthy soul. Psychologists clarify that all of us will fluctuate up and down along the euthymic line throughout our lives (Linden, 2020). A mood disorder diagnosis is indicative of feeling at a peak or a trough either side of the euthymic line for a sustained period of time (WHO, 1993).

Professional diagnosis of depression in the UK is commonly based on criteria set by the Diagnostic Statistical Manual of Mental Disorders (DSM) or the International Classification of Diseases, Injuries and Causes of Death (ICD) (Quevedo et al., 2019). The diagnostic tools were created by the World Health Organisation and the American Psychiatric Association respectively, in an attempt to standardise diagnostic techniques. These tools have similar criteria for a diagnosis of depression, and it will be beneficial within the context of *low mood* to clarify what these criteria are. The authors of the ICD (ICD-10) use the term *somatic* (related to the physical body) for core symptoms of depression that they consider clinically significant. Along with the authors of the current version of the DSM (DSM-5), they further specify that the configuration of “key” (essential) factors and “attributing” (secondary) factors are used to decide upon the degree of severity for a depression diagnosis. I compare these factors as described by the ICD and DSM in Table 2.1:

ICD-10 Key factors	**DMS-5 Key factors**
Majority of the time depressed mood	Depressed mood
Anhedonia	Significant reduction in interest and pleasure
Reduced energy	
ICD-10 Attributory factors	**DSM-5 Attributory factors**
underconfident	Regular fatigue or anergy
Excessive guilt	Daily excessive guilt or feelings of worthlessness
Suicidal thoughts or behaviour	Repeated suicidal ideation, planning or thoughts about death
Observed indecisiveness/vacillation	Reduction in decisiveness, ability to think or concentrate
Agitation/retardation in psychomotor (movement-based) activity	Agitation/retardation in psychomotor (movement-based) skills
Sleep disturbance	Insomnia or hypersomnia
Change in appetite alongside weight change	Significant unintended weight change or change in appetite

Table 2.1: Key and attributory factors in depression diagnosis comparison

It may be noted that the key factors for DSM-5 are differentiated from the first two key factors in ICD-10 by the linguistic terminology used rather than the properties represented. As an example of the linguistic difference but conceptual similarity, in the second line of Table 2.1 ICD-10 uses the term *anhedonia*, which has the same etymological roots as *hedonism* (*OED*, 2024). Anhedonia in ICD-10 can be equated to a significant reduction in interest and pleasure as described in DSM-5. All but one of the attributory factors for these classification schemes share the same themes worded differently. As an example, in the penultimate line of Table 2.1, *sleep disturbance* is comparable to *insomnia or hypersomnia*. To diagnose depression under both the ICD and the DSM-5, clinicians must consider the ratio of key factors to attributing factors over a course of at least 2 weeks to categorise the depression as mild, moderate, or severe (WHO, 1993, pp. 81-86), (APA, 2013, pp. 155-168). The crossover between the factors indicates some consensus in psychology at the time of compiling these manuals.

The guidelines in ICD-10 (WHO, 1993, pp. 81-86) and DSM-5 (APA, 2013, pp. 155-168) were converted by healthcare professionals and academics into questionnaires for reproducibility of

diagnostic criteria. Wright et al. (2017) advise that in order to diagnose depression, in addition to completing one of these questionnaires, the patient will receive a mental state examination. This involves the input of a psychology professional, who will check the appearance, behaviours and speech as well as mood, thoughts, perception, cognitive function and insight of the patient (Wright et al., 2017, p. 3). The questionnaires themselves may be patient or observer completed and can be idiographic (qualitative) or nomothetic (quantitative) (Power, 2004, p. 309). For classification of the informants within this project, I use a mood questionnaire that is nomothetic and patient completed. This mood questionnaire is not intended as a diagnostic tool in this context; it is used in order to standardise categorisation and enable reproducibility.

Treatments for depression as it is currently diagnosed involve managing ongoing symptoms as well as aiming to prevent a recurrence. This may involve medication that affects brain chemistry including serotonin uptake, therapy including talking therapy or Cognitive Behavioural Therapy (CBT) or in persistent and severe cases, electro-shock therapy (WHO, 1993). The long term prognosis for the individuals diagnosed with depression is unfavourable. In addition to a high likelihood of relapse, links have been found between the illness and conditions such as heart disease, as well as self-medicating behaviours that result in substance abuse (Otte, 2008).

A diagnosis of depression may also be associated with social stigma from those who do not have first-hand experience of the mood disorder (Solomon, 2001, p.6). This stigma may be exacerbated by interactions with sociological groupings that are seen as incompatible with the symptoms of depression. As an example, Galasiński (2008, p.130) writes about how some men perceive their depression diagnosis as incompatible with societal concepts of masculinity. Intersections between demographic groupings may further exacerbate this stigma, for example expectations around the intersection of gender and race. Beauboeuf-Lafontant (2007, p.2) writes that societal expectations around black women and strength are perceived as incompatible with depression. Black women exhibiting symptoms of depression may therefore be excluded as a result of this perceived dissociation from culture and femininity.

Psychologists have devised frameworks to help with the understanding of and potential treatment for depression. These frameworks can be loosely placed into the following categories: behavioural; psychodynamic; cognitive; and humanist (McLeod, 2023). Behavioural models frame depression as caused or exacerbated by behaviour reinforcement, either good or bad. These behavioural models concentrate upon a loss of positive reinforcement or reassurance for behaviours, or reduced efficacy of reinforcement. This novel lack of reassurance may be due to factors such as neglect or trauma. An example of this is a child who is no longer given praise for colouring within the lines of a picture drawn in a colouring book. The praise will formerly have positively reinforced this behaviour, but without the praise the child may not feel that this is satisfactory (this is my own illustration of this concept). Reduced external reinforcement for adults and children may result in comfort seeking behaviours, including substance abuse or crying. Associated negative environmental interactions such as angry parents may then contribute towards depression (Blakey et al., 2019, p.587).

Proponents of the humanist approach suggest that depression occurs when basic human needs are not being met, particularly the idea of self-actualisation (McLeod, 2023, p.5). Without basic needs including a feeling of self, it is possible that children and adults may feel undeserving or struggle with expectations placed upon them. Psychologists such as Şimşek (2017) posit that this lack of a sense of self results in self-oriented behaviours. That a poor sense of self results in rumination upon the concept of the self, and that this rumination is a contributor towards depression (Şimşek, 2017, p.30).

Psychologists who advocate for psychodynamic theory suggest that depression may be prompted by a sense of loss, which may be from an object, person or feeling (McLeod, 2023 p.2). It is anecdotally possible to see how losing a loved one may lead to the somatic, key and attributory factors for depression as outlined in Table 2.1.

Cognitive frameworks for the understanding of depression are based in the assumption of a pre-existing underlying vulnerability to depression. These cognitive frameworks include Beck’s theory, learned helplessness theory and attribution theory (Bernaras et al., 2019, p.7). Beck’s theory is that an event such as the death of a loved one causes the activation of an underlying pre-disposition

(Bernaras et al., 2019, p.7). From the point of this triggering event, the person negatively codes their existence and becomes depressed. Learned helplessness theory proposes that something causes an individual to lose their sense of agency, and to feel both hopeless and helpless as a result (Bernaras et al., 2019, p.7)

Attribution theory fits into a cognitive framework for the understanding of depression (Al-Sharif, 2021, p.129) and provides a structure that is compatible with the coding of linguistic patterns. I have therefore chosen to use attribution theory as a framework for some of my analysis. Stratton (1988) built upon existing research to create the Leeds Attributional Coding System (LACS) Manual, which provides specific categories for the perceived causes of events, emotions and experiences. LACS enables classification of how a person experiencing depression (and therefore, low mood) perceives situational genesis with a replicable framework. These causes may be Stable-Unstable; Specific-Global; Internal-External; Personal-Universal; Level of Controllable (Stratton, 1988, pp.29-34). To gloss these individually: stable-unstable: a present cause of pain may be due to something immutable, or a temporary situation. In terms of specific-global: the issue may be something that causes difficulty for only one person or a worldwide phenomenon. For internal-external: the cause of strife may be something the person has done, or outside factors. Furthermore personal-universal: the event may have been uniquely due to the disposition of the person, or dependent on the external situation. Lastly, controllable: the level of control and agency felt within a situation may affect how an individual feels about it. These areas of attribution can be used to analyse the way a patient orientates themselves within their present circumstances, moods and lifestyle. Having coded attribution, it is possible to understand the context for their depression, low mood or other psychological disorders. As an example, a study on primary care institutions found that patients with mental health issues were more likely to talk about their symptoms in general rather than specific terms (Greer et al., 2004, p.521). This attributional context can be used for a range of therapeutic outcomes. For example, attribution can help in perceptions of justice by providing guidance for improving agency and thereby empowering individuals to overcome learned helplessness (e.g. Al-Sharif, 2021, p.139).

Agentive power is something that Johnstone and Boyle (2018) consider important when proposing an alternative to diagnosis of mood disorders and other psychopathologies as outlined in places such as the DSM. Part of their rationale for a different method of understanding mental health is that current doctor-centred categorisation has arbitrary distinctions from one diagnosis to the next (Pilgrim, 2020). Power Threat Meaning Framework (PTMF) is one of the proposed alternatives, which places the individual's experience at the centre. Researchers have suggested looking at the structures and causes of distress and how to ameliorate maladaptive coping mechanisms rather than focusing on *symptoms*. Johnstone and Boyle (2018) examine how power impacts upon the person, defining power as "the means of obtaining security and advantage" (Smail 2005, p.28, cited in Johnstone and Boyle, 2018). Types of power include internal and external; social; economic or coercive. The negative impact of power dynamics can threaten individuals or wider social structures, with consequences including emotional distress, which is the focus of the framework.

PTMF involves foregrounding how one makes meaning of the intersection of power and threat, alongside individual responses to this interplay, as produced and displayed through physical and discourse-based output. These meanings cause general patterns which can be examined and then reconstructed into new frames of understanding, enabling research and recovery to be based on patterns of behaviour, speech and feelings rather than symptoms (Johnstone and Boyle, 2018). A neurology specialist, King (2016, p.113) writes that there is a genetic overlap in addition to symptomatic overlap between major psychological and developmental disorders. Kamens et al. (2018) indicate that psychology may be moving away from an understanding of distress and symptomology centred in guides such as the DSM. This is evidenced through frameworks such as RDoC (research domain criteria), which prioritises understanding of the nuances of mental health rather than specific disorders (Böttger et al., 2023, p.2). I will give consideration to the framework of power, threat and meaning when analysing linguistic trends for this project. Specifically, the analysis will include words, linguistic patterns and themes associated with internal feelings of power alongside the action of power and threat upon the individual.

The experiences associated with depression, as defined in guidelines and questionnaires, are commonly correlated with diminished quality of life for those experiencing them (Solomon, 2001). Despite its long tradition as a concept and the well-established frameworks and processes for understanding depression within clinical practice, researchers continue to debate the scientific validity of depression as a diagnosis (e.g. Kamens et al., 2018). Difficulty in finding biological markers alongside a potentially equivocal and diverse symptomology makes diagnosis problematic (Ballard et al., 2019). It is beyond the scope of this project to discuss the long-term viability of depression as a diagnosis within the realm of psychology. Further research in this area beyond simple amendments to current guidelines appears warranted; the results of this thesis can contribute to that discussion. The experiential embodiment of low mood that contributes to a diagnosis of *depression* as the term is presently used is an enduring concept. Consequently, I investigate the linguistic markers of *low mood* rather than a formal diagnosis of *depression*, through the comparative linguistic analysis of a corpus of elicited verbal language.

2.5 Psychology and linguistics

Qualitative analysis of speech and language to enable insights into the mental landscape is an established approach within psychology (Madill & Gough, 2008). Furthermore there has been an increase in language analysis from a medical perspective, including psychological, since the "communicative turn" (Collins & Brookes, 2023, p.8). This qualitative analysis includes discourse, thematic and phenomenological analysis (Madill & Gough, 2008, p.258). Discourse analysis may involve the examination of spoken, written or signed communication, exactly as it does for linguistics. Thematic analysis is the understanding of themes whilst phenomenological analysis is qualifying consciousness and experiential existence (*OED*, 2024).

Psychologists analysing language data may do so for many purposes, including refining diagnostic techniques, evaluating therapeutic effects or modifying treatment options (Power 2004). For example, Karmen et al. (2015) attempted to design a computer program designed to detect level of depression for individuals participating in mental health-based internet fora, based on language use.

The researchers programmed natural language processing software to seek lay synonyms of diagnostic criteria. They theorised that more synonyms referenced would equate to greater severity of depression. As psychologists, they have chosen not to include any examples of these synonyms, however the program was able to detect depression with around an 84% success rate (Karmen et al., 2015, p.27).

Psychologists may employ a range of methods for discourse analysis. These include hand coding based on a close reading of existing texts, a comparison of responses in semi-structured interviews and the use of specialist software such as Linguistic Inquiry and Word Count (LIWC) (Pennebaker et al., 2022).

LIWC (Pennebaker et al., 2022) has parallels with corpus linguistic software, but was developed to perform general discourse analysis with a psychological focus rather than collecting evidence to enable linguistic insight. The traditional disciplinary divide sees linguists primarily searching for theories to understand language (structure, variation, change), whilst psychologists focus on the creators of that language (Soames, 1984). Modern linguists have moved away from this singular level of understanding, moving towards language as the lens that shapes and filters experience of the world (Baker et al., 2006). Nevertheless, pattern recognition in corpora is interpreted differently in each methodological approach: LIWC foregrounds language as a reflection or mirror of its creator, whereas it is the creation of the language, and its properties, which are the focus for linguists. This is reflected in psychology papers, whereby the emphasis is on the summarised speech acts rather than specific examples when making important contributions.

Psychologists have sought linguistic patterns associated with mental health data, focusing on different aspects of language use including number of words, choices of words and general grammatical patterns. Molendijk et al. (2010) analyse vocabulary choice within psychotherapy sessions; Odea et al. (2018) investigated participants who use internet forums for mental health support; and Bernard et al. (2016) looked at elicited written autobiographical narratives from students. These studies indicate a range of linguistic markers that may be associated with the symptoms of depression (Şimşek & Kuzucu, 2012).

Table 2.2 is my compilation of the recurring features identified by psychologists studying linguistic characteristics of depression. An outline of the studies that reported these findings and an elaboration on what the patterns mean follows.

Feature	Pattern
High frequency	Colloquialisms
	Negative words
	Absolute terms
	Self-focused language (me, my, I)
Orientation	Personal pronouns relative in 'object' position
Temporal focus	Focused on the present when talking about memories

Table 2.2: Psychological linguistic features associated with depression

Smirnova et al. (2018) compared Russian language reports written by patients diagnosed with major depressive disorder (MDD) with similarly elicited reports written by patients who were not diagnosed with MDD. This experiment concluded that the participants with MDD used a higher frequency of colloquialisms, which is congruent with attribution theory (Stratton, 1988). Of particular relevance is that attribution of meaning, context and place for their situation is difficult for these participants. Whilst this Smirnova et al. (2018) study does give ideas of potential linguistic features to look for in my own study, I could identify no published research that confirmed or refuted these indicators. It may be that they are specific to the Russian language and culture. My own project also involves comparing similarly elicited data, but with English language informants.

Psychology researchers have discovered a positive correlation between feelings of depression and negative words across multiple mediums. Rosenbach and Renneberg (2014) took English language data and compared the language of people diagnosed with major depressive disorder (MDD), borderline personality disorder (BPD) and a control group made up of members of the public attending an event. The findings included a greater proportion of terms related to sadness in the MDD group, and of anger in the BPD group when compared with the control. This finding of increased negative words as an indicator of depressive symptoms is reinforced by Huston et al. (2019), who examined the progress of therapy over time. A reduction in depression symptoms

corresponded with a decreased use of negative words, and a concurrent increase in positive terminology (Huston et al., 2019, pp. 325-326). Their participants were from a separate study on depression and displayed indications of an overt link between a more positive outlook and linguistic choices. This linguistic change was the case for every participant considered to have a ‘good’ outcome from therapy. Based on this study, there is the potential to identify recurring linguistic properties that reflect speaker mood.

Al-Mosaiwi and Johnstone (2018) investigate the use of negative emotion words and the potential relevance to mood in a study focused on ‘absolute’ terms. Absolute words are the extremes of a semantic category; for example describing something as ‘abysmal’ rather than ‘not very nice’. On a Likert scale ‘absolutes’ would be the polarised options, such as ‘completely agree’ or ‘completely disagree’. In investigating the use of these terms, the researchers hand coded data from user’s posts on internet forums with a focus on those based around depression, anxiety and suicidal ideation, before inputting the data into LIWC. The outcome was both a higher number of negative terms and greater reliance on absolute terms than in the control (Al-Mosawi & Johnstone, 2018). This immoderate lexicon seems indicative of cognitive biases that could be associated with the symptoms of depression, as individuals seek to express their experiences of low mood. The study further supports this project’s hypothesis that language is an important facet in recognising and understanding low mood. This may be consistent with cognitive behavioural understanding of low mood, in which absolute thinking is sometimes identified as a factor contributing to maintenance of individual’s mood (e.g. Robertson, 2010). Despite this link to existing psychological literature, linguists Hunt and Brookes (2020, p28) point out that Al-Mosaiwi and Johnstone are analysing forum entries for people requesting help. These are therefore more likely to include absolute terms, as people justify their need for help.

A key finding of psychological linguistic research as summarised in Table 2.2 is self-focus, visible in the increased quantity of personal pronouns (Edwards & Holtzman, 2017). For example, Tadesse et al. (2019) found self-focused language is a prevalent theme for people experiencing depression symptoms, based on their study of the language of forum entries on Reddit. The study compares

forum entries in groups related to depression with posts from users in forums related to families as a control. This data was grouped by topic then analysed with LIWC in order to detect language features as well as topic distribution. Both methods of analysis allowed the researchers to find an increased focus on the self, particularly noticeable in the increased use of self-referential terms such as *me*, *my*, *I* from the users in the depression forums when compared to the control (Tadesse et al., 2019).

The linguistic trend towards self-focus was congruous with offline data in a study performed on data sets from six different laboratories that had been collected for unrelated projects. LIWC was used to analyse this data specifically for *I-talk*, or self-referential language. Once again, the increased use of self-referential pronouns appears to corroborate this pattern of elevated self-focus (Tackman et al., 2019). It has been posited by psychology academics (Şimşek, 2013) that the experience of depression includes rumination on the self, potentially related to a lack of clarity over personal identity. The increased concentration of self-referential language lends credence to Şimşek's (2013) theory that individuals with depression, or low mood, show a tendency towards internal reflection and a concentration on individual experiences.

Researchers have built upon this idea of self-focused language by looking at grammatical orientation. Demiray & Gençöz (2018) found that success in therapy results in an associated change of location for first person pronouns relative to verbs. Specifically, first-person pronouns move from an object position to one of agency (subject position). Whilst the study was in Turkish and there are no direct quotes available (original or translated), an English language example of passive rather than agentive speech might be *I was hugged by her* and in the active form *we hugged each other*. In this example the *hugged* is used passively in the first phrase, and actively in the latter, whereby the speaker is involved in the hugging. This indicates a possible linguistic manifestation of the earned helplessness theory (Bernaras et al., 2019), whereby a move away from low mood is associated with an associated move away from a sense of helplessness. These trends have been identified cross-linguistically; e.g. Spanish tweets from users with depression symptoms had a tendency not only towards more pronouns but also a higher volume of verbs and adverbs and a

significantly lower use of nouns when compared to users in a control data set (Leis et al., 2019). This led the researchers to conclude that the people in their depression dataset are more interested in the self and deeds rather than anything external. This may fit in with the concept of ruminative self-focus as a maladaptive mechanism (Watkins, 2004).

Himmelstein et al. (2018) took a syntactic approach to their LIWC-based analysis of autobiographical accounts from people diagnosed with depression. Participants were found to use language indicating a focus on the present when talking about memories, both positive and negative (Himmelstein et al., 2018). Unfortunately, potentially due to the psychological rather than linguistic focus of this paper, the authors did not provide examples of this present focused language. It is possible to surmise that this would involve the difference between *I look up at the fireworks that signify a new Millenium* as present tense and *I looked up at the fireworks that signified a new Millenium* as past tense, as an example. Both of these examples involve a memory of the past, but the difference in syntax represents a different orientation for those memories. The study provides an important reminder that a focus on the past does not necessarily equate to higher usage of past tense, and moreover, that context of language use is important to its accurate interpretation.

Trifu et al. (2017) took Romanian speaking participants and performed a narrative elicitation then content analysis-based study. They concentrated on semantics, syntax and morphology and compared the narratives of participants diagnosed with depression with members of the public who did not register as experiencing depression. Participants diagnosed with depression were characterised by verbosity including tautology (which is a reference to the same concept repeated and rephrased), contrasted with shorter and more abrupt sentences. The participants diagnosed with depression also displayed a greater reliance on the past tense. Trifu et al. (2017) found the use of shorter sentences alongside increased verbosity, with a greater number of words per depressed person. The Smirnova et al. (2018) Russian language study referenced earlier corroborates this finding. They found that the participants diagnosed with MDD used more words, and that their loquacity particularly focused on verbs and adverbs. The study by Leis et al. (2019) found that Spanish-speaking Twitter users with depression used a significantly lower number of characters

per-tweet than those who did not have depression. However, this does not necessarily equate to a reduced number of words- it could be a greater number of shorter words. Since the researchers in all of these studies on verbosity do not include examples of the language generated in the studies, it is not possible to see how the reduced number of characters per tweet (Leis et al., 2019) manifested in practice. Furthermore, since none of these studies were English language based, it is difficult to know if this would be a pattern for my own data.

Psychologists have also investigated the link between lexical diversity and poor mental health. Lexical diversity can be measured by counting the percentage of unique words in a text (Park & Conway, 2018, p.4). Park and Conway (2018) compared Reddit users on groups associated with depression, bipolar and schizophrenia with users on groups associated with weight loss, positivity and exercise. They found that the users on the mental health related fora exhibited significantly less lexical diversity than those in the control group. This measure of lexical diversity can be affected by the size of the texts being analysed. I therefore use a linguistic measure of lexical diversity designed to allow for different corpora sizes when analysing my own data. I discuss this in more detail as part of my analysis in section 6.4.

In the psychological studies I have outlined in this chapter, the focus is on general patterns, and researchers do not provide specific examples of these patterns they describe. Despite the potential relevance of grammar to the understanding of low mood, this does not appear to be a focus for psychological research. Based on the studies outlined, grammar-based findings are likely to be an addendum to a study with a different locus. My study therefore represents an opportunity to apply linguistic frameworks and methods to identify potential markers that distinguish the language of people with low mood from those who are not experiencing low mood.

The psychological approach to the analysis of texts, including use of LIWC, has enabled insight into the patterns outlined in Table 2.2. However, recent research suggests that the analysis of LIWC's sentiment categorisation may be problematic. Ungless et al. (2023) uncovered potential queerphobic bias by using a template based approach on automated sentiment categorisation models (Ungless et al., 2023, p.2211). They found that LIWC coded queer relationship references

less happy than heterosexual relationships (Ungless et al., 2023, p.2221). Furthermore, Hunter and Grant (2025) write that LIWC does not analyse form and function together in a reliable or accurate manner. They provide examples of LIWC misclassifying due to the lack of context, such as in the sentence “It should **be** about having a rebellious attitude” (Hunter & Grant, 2025, p.6). LIWC coded *be* as an auxiliary verb, whilst in this context it is a main verb. Adopting a corpus linguistic approach for the investigation of my datasets will ensure that the data is transparently interpreted by me rather than closed box software. This descriptive linguistic analysis can provide a fine-grained and contextualised account of the distinctive language features in corpora of those with low mood, when compared with a reference corpus.

At the beginning of this project, I began preparations to complete a systematic review of language research into mental health. This was prior to my choice to focus on low mood rather than depression. The intent was for a thorough investigation of similar research in the field to help both situate this work and inform methodological approaches. The title I chose for this review was: ‘Which language and linguistic features can be found in elicited narratives, verbal or written, from people with symptoms of depression or anxiety?’. Whilst anxiety is not a factor that is being actively considered for this project, it is often studied alongside, and comorbid with depression. This systematic review, focussing on narratives, was designed to complement a review by Vaughan (2021), which documents social media data based on similar parameters.

The intent was to publish the findings for the wider scholarly community, and to inform the data collection and analysis of the thesis itself. Due to my decision to change the parameters of this study to low mood rather than depression, I decided that this line of research was no longer relevant and so did not pursue it.

2.6 Linguistics and psychology

There are a select number of linguists who have chosen to adopt the opposite approach, and adopt a psychological focus for their linguistic work. Imahori (2019, p.22) wrote that linguistic analysis is

“an unobtrusive and more objective measure of psychological constructs”. An overview of the insight into the psychological concept of depression from a linguistic focus can be summarized using four representative examples from the field.

Galasiński (2008) performed a discourse analysis on semi-structured inpatient interviews about the experience that Polish speaking cis men and their loved ones have around depression. The men and their families revealed that they saw depression as the opposite of masculine and that those with a diagnosis felt a loss of agency in favour of an experiential existence as a result of their illness (Galasiński, 2008, p.169). These findings were based on a close reading of interviews specifically about depression. Although my data is not focused on the discussion of depression or low mood itself, in common with Galasiński, I account for gender. I examine my data in gendered groups to examine the impact of gender on the experience of low mood.

Grundlingh (2018) took a different approach, comparing authentic and false suicide notes using appraisal theory. They specifically looked for the source of affectation, how it is expressed and degree of feeling. Level of trust, self-flagellation around morality and negative appreciation were found to be significantly more prevalent in the genuine notes. Whilst low mood and suicide are not the same thing, suicidal ideation is often expressed by people experiencing sustained low mood such as that encountered as part of depression (WHO, 1993). There was the capacity for my project to find similar concentrations of terms associated with morality, negativity and a lack of trust.

Harvey (2012) took a corpus linguistic analysis of adolescent messages about depression on an online forum. This was performed primarily through a thematic analysis of keywords, alongside a search for collocates of terms related to depression. This provided an insight into the embodied experience for young people experiencing depression. The study did not provide specific linguistic indicators associated with the condition, and therefore is not a template for my own project. I use a keyword search in addition to complementary analytical methods to build from the foundations of my study.

Hunt and Brookes (2020) built upon the methodology of Harvey (2012), using corpus linguistics to analyse messages in online fora. After a similar thematic analysis of keywords, they also looked at

discursive representation of relevant aspects of the experience of depression. Attributive or suffered, depression is seen as caused by external factors and either a separate thing or something owned and experienced distinctly by an individual. The stated intent was to build upon descriptive analysis with a view to enacting revelatory research upon experiential linguistic signifiers.

Bucar et al. (2021) also analysed linguistic data from online fora using corpus linguistic methodology. Rather than searching for keywords, they used a part of speech analysis to see which grammatical features were more prevalent in forum users with a diagnosis of depression when compared to those without a diagnosis. Their data set was built from users of a subreddit called r/depression: the target corpus was users who explicitly stated they had a depression diagnosis and the reference corpus was those who did not (Bucar et al., 2021, p.3). From this data set, they found that the diagnosed users wrote using more personal pronouns and verbs than the undiagnosed users. They found that the undiagnosed users were likely to use more common and proper nouns than the users diagnosed with depression. Bucar et al. (2021) explain their findings by saying that people diagnosed with depression talk more about actions and have less interest in objects. My thought is that people diagnosed with depression on a depression forum are more likely to talk about actions in the hopes of finding help for their symptoms when compared to those who may be offering support, but this is not discussed in the paper. Unlike Bucar et al. (2021), Harvey (2012), and Hunt and Brookes (2020), for my project I have chosen to examine the language used by individuals with and without low mood rather than their descriptors of depression itself.

Many linguistic studies into depression, both from a psychological and linguistic stance, have involved data collected from the internet (Le Glaz et al., 2021, P.e15708). Le Glaz et al. (2021, P.e15708) advise that the issue with this data source is that the internet users are likely to be young, and researchers “rarely consider cultural or ethnic differences within a sample". This makes generalisable conclusions difficult, since it is not possible to be certain of how representative these samples are of the general population.

The sociolinguistic insights these researchers have found have real world applications, such as suggestions for sensitive and relevant health communication for service providers. The approach

and outcomes of these studies are not directly relevant to my research questions, but these methods are an appropriate starting point.

2.7 Social characteristics

The term *sociolect* refers to the linguistic signifiers distinctive to one's social grouping, from gender, race, sexuality, age, education level, social class and the intersection of all of these (Chambers, 1995). The study of sociolect indicates that identity including demographic groupings can have a bearing on language patterns. For example, studies have found that the gender identity of a speaker can affect their communication styles (Lakoff & Bucholtz, 2004, p.19). Further to this, the intersection of age, gender and culture can influence language patterns in a school environment (Bucholtz & Hall, 2005, p.593).

In this thesis, I examine language patterns, which are affected by belonging to or exclusion from social groupings. It is therefore important to consider how identity is constructed, since this is formative for social groupings and therefore sociolect. Bucholtz and Hall (2005, p.607) write that identity is linguistically constructed through "labels, implicatures, stances, styles, and entire languages and varieties".

Gender, within the social sciences, does not automatically match physical sexual characteristics. Zimman (2017, p.89) writes that gender is the external expression of an internal concept, and that a primary tool for constructing gender is language. To take a simple example, if someone uses pronouns such as *she* or *her* when referring to herself, the assumption of the addressee is that the speaker's gender is female. Jones (2016, p.217) advises that not all people consider themselves to fit within the binary male or female archetypes. Some individuals instead have gender identities that encapsulate elements of both, neither, or consider their gender to be in flux. This is relevant to the two individuals who took part in my study who identified as non-binary (see sections 3.5, 4.3 and 6.7 for more on gender in the context of this study). In this thesis, the term gender may be taken

throughout to be purely related to the portrayal of the internal sense of gender as defined by the individual involved, with no reference to biological sexual characteristics.

Studies referenced in section 2.5 of the literature review had a high ratio of female to male participants. This may be due to the higher incidence of females than males being diagnosed with depression (Maji, 2018). However, awareness of gender based linguistic expressions of time seems salient for my second research question: *Do people with low mood focus on the past?* Park et al. (2017) performed a study analysing how gender and personality type affect temporal focus. They found that women use more future orientated language than men and are slightly more likely to look to the past, whilst men are more likely than women to use present focused language. This difference could have relevance for our understanding of language and depression or low mood, and therefore reinforces the importance of gender as a variable within this work.

The nature of the methodologies used in psychological and linguistic research typically means that the participant demographics for the studies referenced in previous sections were unavailable to the researchers or they were homogenous. Online forum participants are often anonymous (e.g. Hunt & Brookes, 2020); participants may be from a narrow demographic (e.g. Polish men, as in Galasiński, 2008); or volunteers with minimal categorisation (e.g. Himmelstein et al., 2018), which makes it difficult to understand the full impact of sociolect. In each of these examples, participant identity was either the sole focus or not considered. Social demographics are not the sole determiner of personality and linguistic patterns. However, it seems salient to establish if social factors could make a difference to language choices in how low mood is expressed.

According to Maji (2018), there is a higher incidence of depression diagnoses in females over males, but few researchers have directly addressed potential links between depression and gender. To explore this disparity in diagnostics, Leach et al. (2008) examined an existing scale-based data set. They performed a confirmatory factor analysis and multiple group analysis and found that the Goldberg anxiety and depression scale specifically was not affected by gender bias. Other diagnostic tools were not tested as part of their study. It is therefore unclear as to the link between gender and depression diagnosis using other measures.

Maji (2018) reviewed the effect of gender upon depression and considered how societal expectations may increase incidence of depression diagnoses. For example, if standard expectations are motherhood for women and a high income for men, inability to meet these expectations may contribute towards the symptoms of a mood disorder. This was certainly the case in the male-specific linguistic data that Galasiński (2008) collected and analysed. Jaworska & Ryan (2018) found links between gender and linguistic representation of physical pain, which leads one to surmise that there could also be differences in how emotional pain is expressed. If gender can affect language as well as potentially the manifestation of low mood, then factoring this in to the analysis of linguistic patterns will have advantages. The main advantages of considering the impact of gender are greater robustness for the study as well as depth of analysis for my data.

Morgenroth and Ryan (2021) write that psychological literature about depression largely ignores the conversation around gender and sex. Parker and Brotchie (2010) are an example of this, where in a study of gender and depression, they assume that everyone has cis identity throughout. They talk about gendered roles without acknowledging any potential difference between gendered expectations and an internal sense of gender. The psychologists referenced by Morgenroth and Ryan (2021) may choose to treat their participants as cis, but this is not the approach I take. As a linguist, I use the linguistic terms chosen by the individual who takes part in my study to categorise their gender and the other aspects of their identity. Whilst this impacts category word counts, I attempt to adjust for these in my analysis.

NHS Digital statistics (2021) indicate that low mood disproportionately affects people who identify as LGBTQIA+. Argyriou et al. (2021) suggest that these higher rates of depression could be due to societal attitudes towards queer people. These attitudes may result in a sense of isolation, significantly increased stress levels and internalised homophobia. Roberts and Christens (2021) add that the negative impact of societal stressors can be worsened when combined with other marginalised identities, such as race. Within the field of psychology, a review of a decade of studies published across 3 major counselling journals found that very few were focused on LGBTQ participants (Abreu et al., 2022). Of the limited number of studies that could be found, participants

were found to be homogenous, normally comprising of individuals who were middle class, well educated, Christian, white cisgender and homosexual (Abreu et al., 2022, p.708) rather than bisexual, pansexual, transgender, marginalised races and so on. Age was not reported in this study, but it could also be a factor in the type of individual said to be representative of a population group. Since the intersection of marginalised identities can impact mental health (Roberts & Christens, 2021) studies on homogenous participants may be excluding facets of the low mood experience.

Rawlings & Bains (2020) synthesised from psychological studies that adults from minority ethnic backgrounds are disproportionately affected by depression. The authors write that due to the stigma attached to depression, help is rarely sought and adults from minority ethnic backgrounds are more likely to conceal their struggles. This study reinforced the findings of Bailey et al. (2019) who found that specifically African Americans are unlikely to seek help for depression symptoms. Furthermore, African Americans who do seek help for depression symptoms are much more likely than white Americans to be presented with an alternative diagnosis such as psychosis (Bailey et al. 2019). Including data from patients from minority ethnic backgrounds helps research to have the outreach and amelioration impact that is needed. One of my objectives was to ensure that my own research was racially inclusive for this reason, although the challenges of recruitment meant that I was not able to succeed.

Low mood such as that experienced as part of depression clearly affects a wide range of demographics, but socio-cultural background is not often marked and therefore a consideration within studies. Including this social and cultural background should enable rigorous, comprehensive insight into the embodied experience of psychological health communication based on the analysis of the linguistic preferences of individuals.

2.8 Temporal focus

Linguistic references to temporality are a focus of this project, as signified by my second research question. Temporality can be marked by tense, deixis, and lexical and phrasal structures. I examine

whether there is a difference in use by people experiencing low mood when compared to my reference group. An initial driver of this project was my anecdotal observations that the past is a focus in the discourse of individuals experiencing the symptoms of depression. In this part of the literature review, the challenges and complexities of investigating temporality from a corpus linguistic perspective will be examined. This includes an examination of one of the variables that may affect linguistic expressions of time and a potential methodological approach that can address and accommodate the complexities of identifying and analysing temporal reference points using corpus tools. Firstly, an overview of the grammatical properties of time seems appropriate.

Morphology is the structure of words when parsing for tense, and this tense-based approach is a simple way of examining the temporal focus of each word (Carter & McCarthy, 2010, p.2). English has two tenses which are marked through verbs, and these are present and past tense (Biber et al., 2010, p.150). Verb inflection can also be infinitive, the base form of a word which does not indicate a past, present or future temporal focus (Biber et al., 2010, p.115). The present tense can be used to refer to a temporal focus of the past, present or future (Biber et al., 2010, p.152). An example of the present tense being used for a past temporal focus is as follows: *last Tuesday, I **make** a cup of tea, **miss** the table, **drop** the spoon on the floor, then **have** a good laugh about it*. To construct a future temporal focus, English uses a modal or semi-modal verb in the present or past tense (Biber et al., 2010, p.153). As an example of the present tense being used for a future temporal focus: *I **will** make a cup of tea tomorrow*. The past tense can be used to refer to a point in the future from a past reference point, such as: *I **was** going to make a cup of tea but then the doorbell rang*. Tense can be modified with *aspect*, which includes a sense of dynamic perspective and can be perfect (completed) or progressive (ongoing) (Carter & McCarthy, 2010, p.892). Himmelstein et al. (2018) illustrate that a focus on the past can be created using present focused language despite participants talking about past events; this technique is conventionalised in narrative prose fiction, as detailed in Fludernik (2003, p.121). This complexity within the English language indicates that there is more to understanding temporal focus than simply looking for the past tense (Carter & McCarthy, 2010, p.598).

Temporal focus can be indicated through features beyond tense, including temporal deixis (words whose referential meaning is determined by the context of use) such as *tomorrow*; prepositional phrases which are terms related to special/temporal location such as *before*; and locative adverbials which are words that describe actions related to spatial/temporal location such as *initially* (Carter & McCarthy, 2010, p,539). Morphology and individual words indicate the time of focus for the word itself but not necessarily the temporal focus of the discourse. When examining time, there are three main areas to look at: the point in time at which the speech is happening; the point in time the participant is referring to; the point in time that the event being spoken of occurs (Svalberg, 2019). As an example, in the sentence *I used to speak with them about their school days* the utterer is speaking in their present moment, referring back to an indeterminate previous point in time (*used to speak*) at which they would converse about potentially another point in time at which the person being spoken to was at school.

A framework proposed by Svalberg (2018) encourages the mapping of tense along a line to enable a visualisation for learners unfamiliar with the nuances of English. I will outline my reworking of this tense mapping here, and then expand on in further once I have introduced my dataset. This tense tracking diagram (TTD) (Svalberg, 2018, p.424) is a tool designed for such learners but is available as a method of tracking temporal focus in personal narratives:

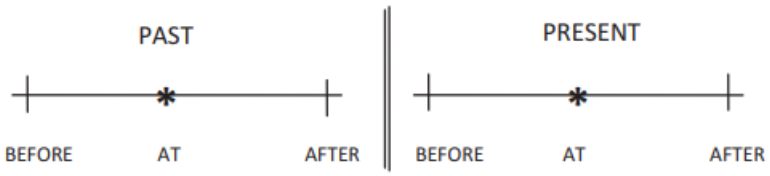

Figure 2.2: Tense tracking diagram (Svalberg, 2018, p. 424)

This diagram enables a clear visual conceptualisation of tense and includes annotations for the time a speech at takes place, the time of an event being spoken about and the metaphorical use of tense. Svalberg provides an example of this diagram in use for an extract from the novel *The Children of Men* (P.D. James, 1992):

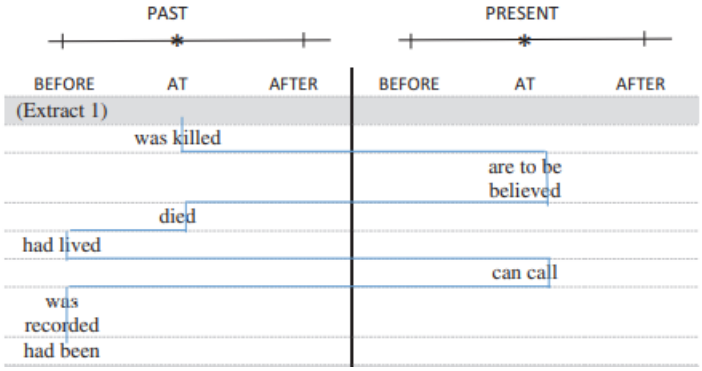

Figure 2.3: TTD in use for *The Children of Men* extract (Svalberg, 2018, p. 428)

The salient temporal words have been taken from the extract and placed into the relevant column in the TTD. I used Svalberg's TTD framework as a basis for manual tagging of my data, as follows:

PB (Past Before)	CB (Current Before)
PN (Past Now)	CN (Current Now)
PA (Past Afterwards)	CA (Current After)

Table 2.3: TTD based tags

These are based upon the TTD as outlined in Table 2.2 and will enable a clear demarcation of tense-based focus. The choice of letters rather than a numerical code is deliberate, due to the assertion of Smith et al. (2008) that numerical tags have greater capacity for user error. To map these onto *The Children of Men* extract in figure 2.2, they would appear as follows:

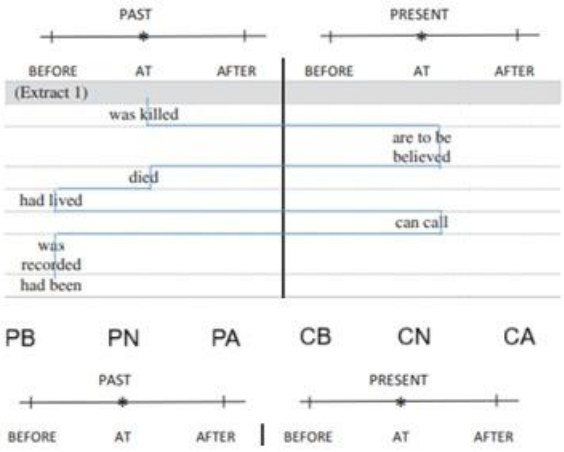

Figure 2.4 TTD amended to show my tags in *The Children of Men* extract (Svalberg, 2018, p. 428)

In the 'Past At' column terms such as *was killed* would be tagged as PN, the 'Present At' column terms such as *are to be* would be tagged as CN and terms in the 'Past Before' column such as *had*

lived would be tagged as PB. The utilisation of the visual tense tracking tool alongside associated annotations will ensure that the tense-based part of speech tagging has a reproducible, clear framework.

Mapping temporal focus is therefore not as simple as looking at the tense used for individual words. Native English speakers tend to switch between tenses without necessarily being aware of doing so (Carter & McCarthy, 2010). To examine tense use in conversational grammar, Smith (2020, p.7) used the corpus annotating tool CLAWS to perform an initial categorisation before disambiguating manually. CLAWS tags tense and aspect of verbs with one of three options*: past tense*, *past participle*, *present participle* (Garside, 1987). In this study I will therefore begin with using CLAWS to analyse my own data, to see what temporal indications can be detected through automation.

It is also possible to take a more holistic approach to temporal analysis than TTD and PoS tagging. Matthiessen and Halliday (1997, p.1), when talking about how we understand things that happen, wrote:

> processes are phenomena that unfold in time and hence have a tense system; but different process types have different temporal profiles and hence different unmarked present tense selections.

There is more to temporal focus than can be clearly marked using tense, even when including the tense tracking diagram. Since the tense tracking diagram is designed for pedagogical use, there is an element of simplicity to it. There are more options for temporal focus than past, present and future (Leech, 2004). An additional option for temporal focus is the universal focus, a concept psychology-based attribution theorists are familiar with (Stratton, 1988, pp.29-34). In this context, it is the idea that something is now as it always has been and always will be. The universal focus covers the past, present and future and therefore goes beyond the frameworks I have considered up to this point. This is an additional type of temporal focus that I code manually in order to assess its prevalence. When examining temporal focus to explore my second research question: *Do people with low mood focus on the past?* the previous literature signals the importance of considering past, present, future and universal foci.

2.9 Conclusion

This literature review provides the background and building blocks for my study. I have provided context to what applied linguistics is, what kind of data an applied linguist may work with and some of the approaches that applied linguists have taken to discourse analysis including appraisal theory. In particular, I have outlined the background and some of the techniques involved with the corpus linguistic methodology I will be using, the software that enables it and the reason behind my methodological choices. I have included an overview of the psychological grounding for this study in terms of what depression is, frameworks for understanding this diagnosis and alternative understandings behind the medicalisation including my choice to focus on low mood. I then provided a summary of published research that combines psychology, specifically low mood alongside depression, with the study of discourse. I have summarised the findings of sections 2.5 and 2.6 in table 2.4:

Linguistic feature based on comparison of depression/low mood and control	**Example reference**
Total number of words is higher overall	Trifu et al., 2017
Increased use of colloquialisms	Smirnova et al., 2018
Higher frequency of negative words	Rosenbach and Renneburg, 2014
Greater use of absolute terms	Al-Mosaiwi and Johnstone, 2018
Prevalence of self-focused language (me, my, I)	Edwards and Holtzman, 2017
Less agency, including personal pronouns in 'object' position	Demiray & Gençöz, 2018
More verbs and adverbs	Leis et al., 2019
Fewer common and proper nouns	Bucar et al., 2021
Focused on the present when talking about memories	Himmelstein et al., 2018
Rephrasing and repetition of concepts	Trifu et al., 2017
Less lexical diversity (mood disorders and psychosis)	Park and Conway, 2018

Table 2.4 Summary of linguistic features associated with depression

With an overview of potential linguistic features associated with low mood and depression,, I have also included information on the relevance of demographics for these disciplines. I continue with a section particularly relevant to my second research question- approaches to the examination of temporal focus within English language data. The chapter is completed with a thorough description of the linguistic and psychological frameworks I combine to gain a different perspective on my data. This chapter has provided a firm contextual and methodological grounding upon which to build the rest of this thesis.

Chapter 3 Data generation

3.1 Overview

In this chapter I outline the procedures used to generate data for this project. I begin with ethical considerations, followed by a discussion of my own positionality. Before generating the data for my main study, I performed a pilot study in order to test and refine data generation methods and to review the suitability of the data generated for potential analytical approaches. I detail the process for this pilot study, from informant recruitment through to interview techniques, categorisation and how this informed the approach for my main study. The data generation for my main study will be discussed in terms of the informants including recruitment and demographics. I also detail the schedule, narrative elicitation techniques and informant classification. It is difficult to create an ideal corpus within real-world contexts (Douglas, 2009, p.7) but within this chapter I will detail the effort undertaken to enable a balanced dataset.

Before discussing data generation, a note on terminology. Throughout this chapter I will be referring to the people who chose to take part in my study as *informants*. Both University of Leicester and University of Leeds use the term *participants* for their ethics application process, and this appears to be standard based on personal and anecdotal experience at other universities. However the British Association of Applied Linguists (BAAL) uses the term *informant* throughout their good practice guidelines (BAAL, 2021). After careful consideration, the people involved in my study did not simply participate, they performed a more agentive role. I have therefore selected the term *informants* to both reflect current professional terminology and the active involvement of the individuals themselves.

3.2 Ethics

3.2.1 Overview

In this section I will outline the logic behind the ethical considerations for this project. This includes my approach to consent, safeguarding the data provided to me as well as plans to minimise negative impact on both informants and researcher. Ethical research practices were important to maximise the protections for my informants and myself in terms of their data and our mental health. The main ethical issues were around consent, the personal data being collected, and potential disclosure related to feelings or experiences. Of particular concern was the mood questionnaire and how to sensitively handle disclosures of suicidal ideation.

3.2.2 Informant consent

The Participant Information Sheet (PIS) was based around a template provided by the University of Leeds. Since my intended informants were not all accustomed to the complicated wording sometimes associated with PISs, I endeavoured to simplify the wording as much as possible. I sent these PISs by email so that the informants could re-read and clearly see what was involved prior to granting their consent, as well as refer back to them after the interview. I gave the informants the opportunity to ask any further questions via email, social media or on the video call before beginning the interview formally. I allowed at least 48 hours between consent being granted and commencing the video interview itself. This was in order to provide informants with the option to withdraw prior to the interview commencing and to reduce and feelings of pressure of expectation that might arise from their involvement.

The ethical review boards advised me that they prefer informants to have an unlimited amount of time to request the withdrawal of their data. I described the methods I would be using for data analysis to these boards, including the requirement for data amalgamation. Upon this explanation it was agreed that each informant should be given at least one month from the date of their interview to be able to withdraw from the study. The precise wording I used was “you can change your mind and withdraw from the project at any time up to 1st December 2022, and you don’t need to give a reason for this” (Appendix B). Thankfully no informants took me up on this offer, and so all 52 interviews were used in the analysis.

3.2.3 Data protection

In order to ensure secure collection and storage of my main study data, I adhered to University of Leeds' ethics guidance. The data itself is saved in unsynced Office 365 cloud storage and encrypted and backed up in a password protected external hard drive. Each informant was assigned an arbitrary participant number between 1 and 200, created using a random number generator (Haahr, 1998) and out of chronological order. All data (audio-visual recordings, transcriptions and questionnaire data) is stored with reference to these arbitrary numbers, as opposed to identifying information. A key file containing salient identifying information (specifically name, email address and participant number for the purpose of feedback or withdrawal) is password encrypted and stored in a separate area of the Microsoft OneDrive. After the data was gathered and the cooling-off period had passed, I transcribed the audio recordings: pseudonymising and anonymising as I went in order to avoid data identifiable to anyone else. I used pseudonyms for all names, from pets to partners, children, friends and places. Each replacement name was chosen at random using an online name generator (Campbell, 1996). Informants were made aware of how their data would be stored using the formal University of Leeds GDPR statement (please see Appendix B for a copy of this) which was included in the information pack emailed prior to the interview.

3.2.4 Detriment mitigation

Each informant was sent a debrief email (Appendix D) after the interview, in which I thanked them for their involvement and attempted to address potential rumination on mood by providing information about mental health support. A brief guide to available support was also included in the thank you for the categorisation questionnaire (Appendix D). For the informants whose score registered them as experiencing severe low mood or who expressed suicidal ideation, I followed up with a tailored message of support and options for professional help (Appendix D).

BAAL guidance (2021, p.15) reminds us that part of the responsibility of a researcher is to guard one's own wellbeing. This is an important and sometimes overlooked aspect of ethics; for instance, Lorenzo-Dus (2021) expressed concerns around linguists examining data that could be distressing within the context of Forensic Linguistics. In the case of the project they discussed, amelioration

strategies for mental health difficulties arising included access to therapy and a supportive community who could take over the work if it became too distressing. During my PhD study there was no time or fiscal budget for these. However, a reflexive diary alongside a self-imposed limited time span for data gathering combined to keep me grounded within the data gathering stage.

3.2.5 Final points

As an incidental and reflexive point, PhD research in the discipline of English literature (the dominant discipline of the School I was initially based in at University of Leicester) rarely seems to include data gathering of this kind. The ethics boards were made up of people from outside of linguistics, and as such approval was preceded by a high level of detail and clarification from those who were working from a very different expertise. This level of scrutiny provided a useful opportunity to refine the data generation before embarking upon it. Ethics at University of Leicester was granted 15th September 2021 by Science & Engineering, Arts, Humanities and Law Research board, approval number: 29455-ks589-ss/ar:English. At University of Leeds this was granted 31st August 2022 by the Faculty of Arts, Humanities and culture, approval number: FAHC 21-127. Both boards were very kind to take the time to understand this project and ensure that all involved were protected.

3.3 Positioning the researcher

The person one is will influence how one approaches data collection, as well as how informants react (BAAL, 2021, p.2). Whilst it is not possible to control for or even classify all factors that affect this influence of positionality, acknowledging it enables active awareness which can inform the approach for data collection and analysis. This awareness will increase understanding of how informants react to me as well as why my informants chose to take part in my project. This is particularly relevant since I used my personal social media as part of the recruitment process (see section 3.5.2 for further information on the recruitment process). I am a white, middle-class, cis female, queer, MA educated, potentially neurodivergent researcher in my early 40's with a mid to lower socio-economic childhood. I also participate in the relatively unique community of Live Action

Role Play (larp) hobbyists. I entered into this project as one who has never suffered from depression, but who has actively been a support to those who have. Some of those people I have supported in the past chose to be informants within my study and as such there is a degree of circularity. The wider remit of low mood rather than focussing on those who fit the criteria for medical depression may have widened the informant pool. However, the informant pool will also have been affected by homophily, which is the tendency for people to interact with those who are similar to themselves (Block & Grund, 2014).

3.4 Pilot study

3.4.1 Overview

In the interests of straightforward logistics, I wanted to generate data from my main study informants with one interview each. It was therefore logical to trial data generation in order to ensure the efficacy of the questions and interview techniques. Data generation for the pilot study is summarised in Figure 3.1, which follows on the next page.

Outcomes from the data generation detailed in Figure 3.1 including the categorisation from the questionnaires and word count from the interviews are summarised in Table 3.1:

I.D.	Age (yrs)	Gender	Race	Interview length (words)	Depression
PI01	19	Female	Black	678	N
PI03	26	Female	White British	837	N
PI14	22	Male	White	2098	Y
PI20	22	Female	White British	962	N
PI29	18	Female	Black Caribbean	545	Y
PI39	19	Female	Black African	445	N
PI41	22	Female	White British	783	N
PI44	19	Female	White	801	Y
PI46	19	Male	Black	412	N
PI47	19	Female	Black	335	N
PI49	19	Male	White- Other - European	563	N

Table 3.1: Pilot study informant classification table

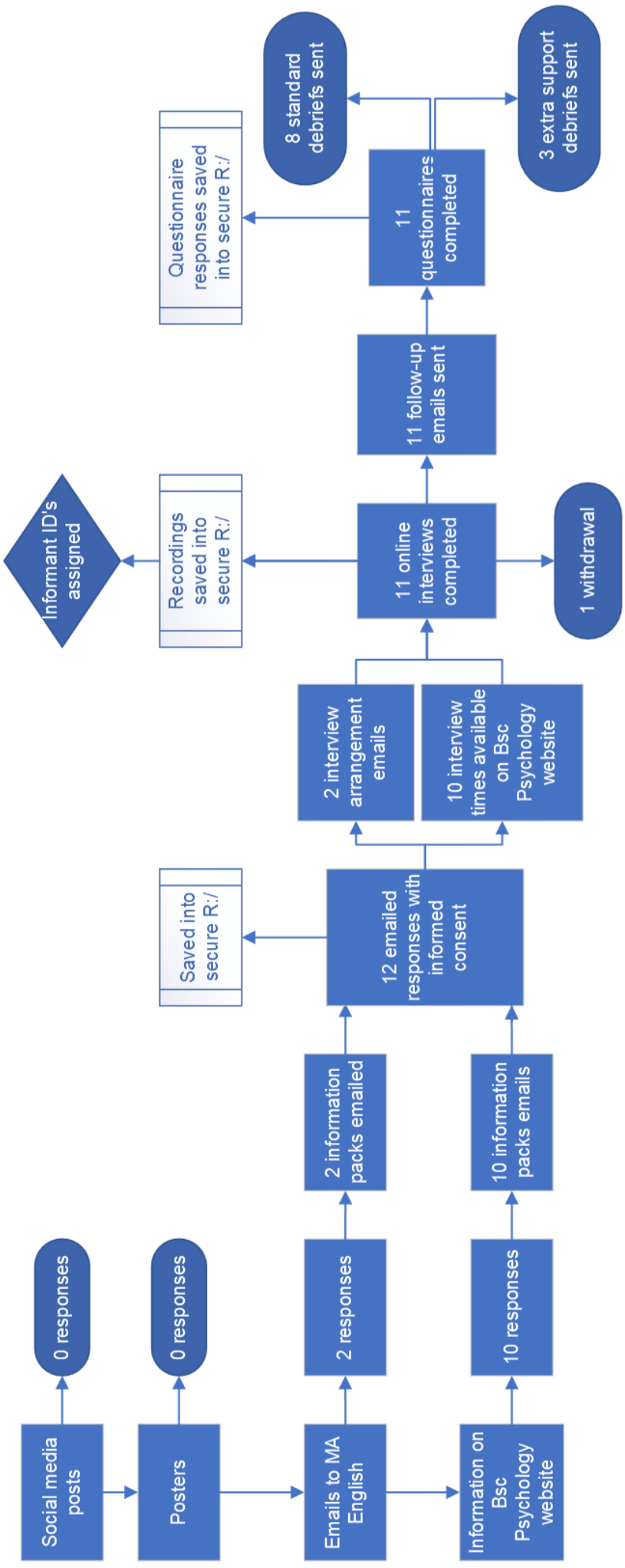

Figure 3.1 Pilot study data generation flowchart

Further information on how and why this information was obtained will be continued in sections 3.4.2 and 3.4.3.

3.4.2 Recruitment and informants

When deciding on the number of informants for piloting, I wanted to ensure that there were enough to test and refine my questions, interview and analysis techniques. This was considered against the constraints of available resources including time, so 10-15 informants enabled a balance of these factors (Bryman, 2012, p.263). I wanted to ensure that informants for the pilot study would not intersect with the informant pool for my main study. Reusing the same people would result in a reduction of goodwill and a possibly skewed data gathering, since the responses would not be spontaneous a second time. Previous studies of this nature had often used students to generate data (e.g. Himmelstein, 2018) as well as homogenous participants (Abreu et al., 2022). I wanted my informants to be diverse to enable my research to be generalisable, and this precluded the exclusive use of students in my main study. Within a university setting, students are readily available and so I decided to take advantage of this for my pilot study. I had no inclusion or exclusion criteria beyond the requirement that my pilot study informants were all students who were conversant enough with English to respond to a narrative elicitation question.

For this pilot study only, I incentivised participation with entry to a £50 prize draw, and in the absence of external or internal funding this was self-funded. There are often many advertisements encouraging students to take place in studies, and so I felt that this potential financial reward would help boost recruitment. In lieu of adequate compensation it is sometimes considered viable to enter participants into a prize draw to incentivise involvement (Oates et al., 2021, p.20). Potential downsides include attracting money motivated individuals who are uncommitted or who may withdraw their data after the prize draw (Tishler & Bartholomae, 2002). The financial incentive was only explicitly mentioned by one of the informants, who showed commitment and did not ask to withdraw their data.

Poster and university social media-led advertising were not successful in generating any interest. Professor Mary Ann Lund and Dr Alice Welham kindly advertised to their students, providing me

with 2 English postgraduates from an email and 9 psychology undergraduates who responded through an automated system. Each of these informants was given a randomly generated identification number between 1 & 50, this was not based on order of interview, and I used an online number generator (Haahr, 1998) to minimise human involvement.

3.4.3 Pilot data generation

The ethics board at Leicester felt that in the interests of data security, recordings and transcriptions should be made to and from a secure drive only accessible from the University of Leicester campus. Since this was at a point in the Covid-19 pandemic that parts of the campus were less accessible, it did affect the scheduling of interviews and transcription work. The interviews in particular were affected, since I needed to be on campus in a room with no chance of interruptions or being overheard. On a campus in partial lockdown, a room with guaranteed privacy was not readily available, and so I had to book all of my interviews in advance, and most in close succession.

Spontaneous verbal responses to a prompt carry a lower risk of self-censorship than written or pre-prepared responses. A natural vernacular style could increase the opportunity for unconscious revelations based on unguarded informality (Labov, 2013, p.3). In order to enable this vernacular style, I made the choice to deliver this prompt personally: audibly and visibly. There is some ambiguity over the effect of interviewer attributes in conducting data gathering (Bryman, 2012, p.227). I balanced the micropolitics around researcher identity (Bhavnani et al., 2020, P.256) that a personal approach would bring with the rapport-building that would be impeded by an anonymous approach. I took a reflexive approach with a view that awareness of positionality would allow me to account for some of the differences in informant responses (Jacobson & Mustafa, 2019).

I needed to encourage a sufficient response with a question or series of questions that would be open and not temporally leading. Labov (2013, p.4) advises that “death, sex and moral indignation” are tested topics that prime lengthy and interesting narratives. After careful contemplation and research, I concluded that all three of these topics are emotionally and potentially temporally charged. An emotionally charged question might lead an informant towards or away from low mood, rather than working from their baseline. A temporally charged question would skew the exploration

of my second research question: *Do people with low mood focus on the past?* Consequently, instead of taking my lead from Labov, I amended the beginning of a prompt about experiences of healthcare "When you think about things that are important to you, how do your provider and his or her staff measure up" (Grob et al., 2016, Table 2). This was tested by Grob et al. (2016) and found to deliver balanced and meaningful results. I wanted to make the question more esoteric and less about materiality in order to encourage a wider range of responses. I therefore removed the word *things*, then tidied up the ending by asking what came to mind, providing the prompt: "When you think about what is most important to you, what comes to mind?". I wrote a range of follow up open questions with the intent of only selecting the ones that were relevant based on an informant's response to the first question. These questions were:

- Important things can mean people, places, possessions, skills, health, arts, culture etc
- Please try saying whatever comes into your head
- Does importance change?
- What makes that important for you?
- What makes you think/feel that?
- How does that contribute to its importance?

Since video recording is the method I would be using for my main study, it was logical to trial it for the pilot study. Due to the constraints of the pandemic and a self-funded PhD I knew that it would not be possible to interview my informants in person for the main study. The travel that would likely be involved would be prohibitively resource heavy in terms of time and finances. The frequent use of video calling during the height of Covid-19 lockdowns taught me that conversations flow more smoothly when it is possible to read someone's facial expressions. Archibald et al. (2019) corroborate this, writing that video interviews with recorded audio can encourage more engagement than voice alone. I decided that the recorded video would also help with lip reading for the transcription stage if my informants were unclear.

I asked informants to fill in a consent form and to read a participant information sheet and privacy notice (please see Appendix A) before calling each individually rather than in a group. This choice

to call individually was in order to help with scheduling and ensure that each informant was given the same amount of attention during the interview. I used Microsoft Teams for these calls due to its existing use as part of their university studies. I called from a private room on Campus, with a standing desk and scheduled the calls over two days. The first day of interviews involved the two English postgraduates. The second day was the nine psychology undergraduates. I allowed up to 25 minutes for these interviews and therefore scheduled them at half hour intervals as well as a lunch break.

During each call I introduced myself, briefly advising the focus of my research and making sure that they were happy with the consent form, followed by rapport building. I wanted to put them at their ease and so allowed around five minutes to ask questions related to their participation in other projects, their university career and the weather. This seemed to help them to feel more comfortable and I found that conversation appeared to flow more naturally towards the end of this time. Then I began recording and advised them that it was time for data gathering before using my main prompt and any follow up questions from the list that could be considered appropriate.

Once they finished speaking, I thanked them and asked that they complete the follow up information as soon as possible before reading the debrief. I emailed a link to a questionnaire with a freeform text box for age, race and gender, followed by a self-complete depression measure called Clinical Outcomes in Routine Evaluation Outcome Measure (CORE-OM) (Evans et al. 2000). The full CORE-OM questionnaire is available in Appendix C. I chose CORE-OM because the standardised depression measures such as PH9 involved in-depth questions around potentially ethically questionable subjects for an English researcher to be asking, around such subjects as suicidal ideation. Whilst CORE-OM briefly mentioned thoughts of suicide, it is more of a wellbeing than depression measure and so I decided that it was the more ethical choice.

Once the survey response had been submitted, I sent a debriefing sheet by return of email. This included a thank you, information on where to find support if needed, reaffirmation of data protection policies and a reminder of how to withdraw from the study. I then immediately checked the results and if the informant scored particularly strongly on CORE-OM sent a secondary follow up email

(please see Appendix C) and for one particularly unwell individual involved the University's counselling service, having advised in the Participant Information Sheet (Appendix A) that this was a possibility.

Once all of the data was collected, I used the random number generator (Haahr, 1998) to decide which informant would win the prize draw. I contacted them by email in order to arrange this, and paid the incentive in a timely manner.

3.4.4 Pilot learning points

Financially incentivising participation had seemed like a good recruitment tactic. However, the personal financial cost of a monetary incentive did not translate to the data reward I had hoped for.

Scheduling 9 interviews in swift succession was achievable but resulted in a slight deterioration in quality of interview style. Towards the end of the 9 interviews I sounded more flustered and did not ask as many follow up questions. For my main study I resolved to allow at least 2 hours between interviews in order to allow myself opportunity to complete admin and reset between each one.

I adopted a conversational customer service style of interview, agreeing and expanding upon what was being said as well as using verbal nods. Verbal nods are noises like “mm hmm” and words like “yes” to show attention and encouragement. These techniques made me appear more friendly but resulted in occasional interrupting and derailing of thought processes and a less structured interview. The selection of available follow up questions also reduced the amount of control over the process, leading me to conclude that asking the same follow up questions of each informant would increase replicability.

One of the informants in this study had been taught English in a formal setting after the age of 5 rather than immersive acquisition. They expressed themselves fully and fluently and so I transcribed and analysed their data. I found distinctive linguistic patterns when comparing this data to the rest of the group that were indicative of this difference in acquisition. This was interesting to me, but beyond the remit of my research questions to investigate. This distinctive language pattern informed

my main study recruitment, whereby all of my informants learned English in an informal setting before the age of 5.

The categorisation questionnaire also did not fulfil the change in my focus from depression to low mood which happened between the pilot and main study. A questionnaire diagnosing depression could become outdated if mainstream psychology moves away from current mood disorder diagnoses. The depth of depression specific questions in CORE-OM was also more intrusive than is necessary for the purpose of this study. For example, CORE-OM asks about suicidal ideation and self harm in explicit terms. Within the context of a student, I was able to contact support on their behalf when their responses indicated severe mental health difficulties. It is much more difficult to do this for a member of the public, or a friend. I therefore deemed it necessary to find a less medicalised questionnaire and look at the idea of mood rather than diagnostic criteria. When categorising according to demographics, I had a mix of race and gender, but I also wanted to look at sexuality and socioeconomic status for the main study. These extra demographic categories will contribute towards my understanding of how representative my informants are of the general population of the UK.

3.5 Main study

3.5.1 Overview

Data generation for this main study was based on the work performed for the pilot study but with amendments. These amendments were based on both learning points from the pilot study and adjustments required for recruiting a different type of informant i.e. non-students. The flow chart to show my data generation methods is figure 3.2, and follows on the next page:

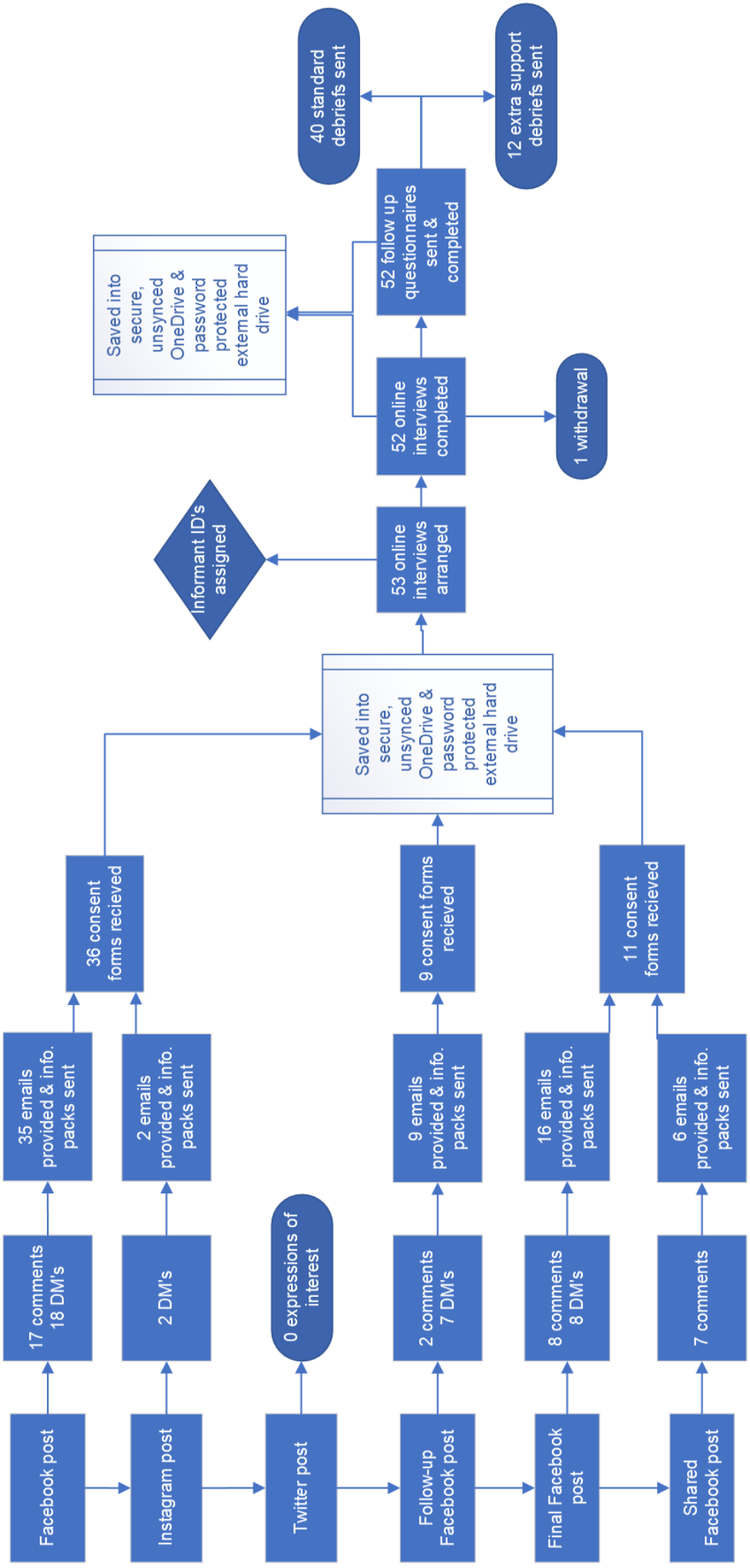

Figure 3.2 Main study data generation flowchart

3.5.2 Informant description and rationale

For the main study my aim was to interview around 50 people from as diverse a background as possible. My target was 50 people for both the size of the corpus and the anticipated proportion of those experiencing low mood to those who would not be. In terms of the size of corpus, 50 informants talking for around 15 minutes each would provide a modest number of words for a study of this kind. Rate of speech according to Adolphs and Carter (2013, p.9) was 10,000 words per hour in the CAN-CODE project, equivalent to 166.6̇ words per minute. The intent was that this would provide a combined target and comparison corpus size of around (50 x 15) x 166.6̇ = 124,950 words. This was a manageable quantity of data, to be collected and transcribed within the necessary time frame, whilst giving sufficient breadth to evaluate the potential correlation between language choices and social factors.

Estimating the proportion of low mood informants was more difficult, since I could not find official statistics on incidence of low mood within the population. I considered statistics for incidence of diagnosed mood disorders in the UK, which was 1 in 10 (PHE, 2020) and considered that the broader category of low mood would likely be higher. This would provide me with a target pool of at least 5 (1 in 10 being proportionally 5 in 50) informants with low mood, though likely much higher and a larger group of informants who were not experiencing low mood to act as a comparison.

3.5.3 Informant recruitment

My aim was to recruit a diverse selection of informants who were representative of the population of the UK. External validity would have been ideal but not possible with the constraints of this project. The resource limitations prevented any paid incentives, as well as external advertising. A combination of my varied and lengthy career alongside my hobby means that my social networks are wider than the average for a non-influencer. For example, the average number of Facebook friends in 2019 was 308 (Smith, 2019) whilst at time of recruitment I had 650. This provided a pool of potential informants with whom I already had a relationship but limited capacity for coercion based

on any existing power imbalances. I posted the following advert across my non-academic Twitter, Instagram and Facebook accounts:

“Would you be interested in helping me with my PhD research? It would take about half an hour of your time, and involve a video call followed by a questionnaire. All I want to know is what’s important to you, a little bit of information on who you are and then how you’re feeling. Pop me a PM/DM for more information if you would like to be involved!”

From this, 69 people sent a message, 68 provided email addresses and 56 sent their consent forms back. Four of those sent back were unable to interview due to scheduling issues and so in total 52 interviews from self-selected informants were conducted. A summary of interview schedules, lengths and informant demographics can be found in section 3.5.6.

3.5.4 Data categorisation

An early design element was placing a mood questionnaire after the interview, in order to prevent rumination on mental health prior to the data gathering. Including the demographics questions alongside this questionnaire was the least labour intensive way of categorising informants. Therefore the questionnaire at the end (Appendix D) included a series of free form text boxes as well as a formalised mood questionnaire. I used JISC surveys to administer this questionnaire because it enabled anonymity for my informants and I could send the categorisation questionnaire using a simple emailed hyperlink.

I wanted to enable my informants to self-identify as much as possible by using these free form text boxes. My intent was to see how representative my informants are as well as if their demographics affected their linguistic choices. I asked for age, race, gender, sexuality and highest level of education. The latter was a compromise, since I had wanted to consider socio-economic status but could not find a way to enable this through self-definition. Umney (2018) tells us of the complexities involved with class, which are more than could be easily summed up within one or two words. The final informal question asked if the informant considered themselves to be depressed. Whilst the focus of the study is no longer depression, links between self or formally diagnosed mental illness and scores on the mood questionnaire were another potential variable of interest.

I had decided to use a more feelings orientated and less diagnostic measure of mood between data collection for the pilot study and ethics application for the main study. This was due to both the difficulties involved with disclosures encountered as part of my pilot study and the change of focus from depression to low mood. I decided to use the Mood and Feelings Questionnaire (MFQ), which was developed by Angold and Costello (1987). It originated as a mood measure for children and young adults so is simple, ensuring that all informants, regardless of educational background, were able to engage with the questionnaire.

3.5.5 Interview schedule

I began advertising for informants once ethics approval was granted by University of Leeds, which was 31st August 2022. I emailed an invite, GDPR notice and participant information form as soon as potential informants provided me with an email address. Once I had a signed and dated participant information form, I allowed at least 48 hours prior to scheduling the interview itself. A consideration from the ethics board was that the timescales for withdrawal of data for informants was more truncated than they would prefer. In order to minimise opportunities for regret, I therefore allowed extra time for reflection prior to the interview itself. I allowed the informants to choose a time then sent a Microsoft Teams invite and began the meeting at the scheduled time. A learning point from my pilot study was not to have meetings in swift succession. Therefore I never did more than three interviews in a day and always allowed at least an hour between each. I also allowed 5 minutes before and afterwards to fill in a reflexive journal.

Once the meeting began, I made sure the informants were comfortable with the information sent in advance as well as reaffirming consent and signing their consent form in front of them. Then I advised each informant that they were welcome to talk for as long or as short as they would like to, and that I would avoid a conversational style including verbal nods. The latter was particularly important since I learned from the pilot study that it was important for informants to verbalise without interruption, without feeling that I was being inattentive. As Brinkmann and Kvale (2018, p.68) write, silence is an essential interview technique that allows room for full and fruitful responses. I advised my informants that there would be one main question and three follow up questions, and that the

wording was the same for everyone. Then, a final check in to ensure that recording was still permissible and the interview itself began.

The interview prompt for my pilot study was carefully researched (please see 3.4.3 for details) but the follow up questions were varied, leading and imprecise (as discussed in 3.4.4). For the main study I therefore used exactly the same one main question then three follow up questions for every interview:

- When you think about what is most important to you, what comes to mind?
    - What does importance mean to you?
    - Why is that important?
    - What else is important to you?

Once each informant was comfortable with the extent of their responses, I turned off the recording. Before ending the call, I made sure that they were happy to immediately fill in the follow up questionnaire and thanked them for their time. Then I waited for the categorisation information to register in my JISC surveys page and immediately entered the information into a spreadsheet against their participant number.

The final stage was sending back the consent form back to the informant with my signature added and a thank you. If the informant registered as low mood I also included a message of sympathy and signposting. The signposting was towards supportive telephone lines and internet forums or an urgent practical tools page created by a mental health charity (please see Appendix D for details). Whilst there was a short summary of advice services given at the end of every questionnaire, I wanted to ensure those in the most distress had the details readily available.

3.5.6 Transcription

I conducted interviews with 52 main study informants, from the 9th of September 2022 through to 17th November 2022 and these ranged in time from 2:01 up to 28:15 minutes with an average interview length of 8:15 minutes. I performed the transcription in three stages, preferring to tackle all the recordings at the same time. Having taken advantage of Microsoft Teams' automatic

transcription, the first stage involved running the data through a VTT cleaner (Azure, 2022). The VTT cleaner formatted the automatic transcription in a clearer and easier to follow way. I continued with the first stage by removing punctuation but inserting paragraphs to help break up the text. The second stage was watching and listening to the video to ensure that the wording was right and introducing pseudonymisation. This stage took a little while due to some of the accents of the informants, which the software was not able to accurately transcribe (a known issue, as investigated by Harrison (2023)). The third and final stage was adding in paralinguistic features such as pauses, filled pauses or vocalisations such as laughter.

There is a strong tradition of linguistic transcription styles upon which to draw, including those specific to corpus analysis. I chose transcription protocols from Collins & Hardie (2022), since their recommendations were based on a corpus-linguistics based ethnographic study of health communication. I found their suggestions to be efficient and logical, particularly as concerns minimal punctuation and breaking down pauses into three approximate lengths rather than precise timings (Collins & Hardie 2022, Appendix E). The only differences were adding in my own signifiers of verbalisations, specifically a gentle scream and a tongue twist noise. These verbalisations were made by my informants but had not been accounted for in the Collins and Hardie (2022) model. Grouping pause lengths, standardised spelling and incidental sounds enabled a representative corpus that could also be read and compared reliably by corpus linguistic software. For example, since all pause lengths were grouped in the same way, the software would allow me to identify whether there was any significant link between pausing, including length of pause and mood. The fully transcribed corpus combined was 52,101 words with the shortest interview word count of 74, and the longest of 4835. Koester (2010, p.67) advises that the parameters for a small corpus are up to 250,000 words. By this measure, the corpus is very small. However Hunston says “a corpus is often a compromise between what is ideal and what is possible” (Hunston, 2018, p.21). With enough attention to detail and control in the data generating process it was possible to still provide a fruitful data set to explore for my research questions. Further information on word counts including breakdowns by demographics can be found in chapter 4 and section 6.2.

3.6 Reflexivity

To preface my reflexivity on the data generation process: my original intent had been to use therapeutic transcripts from active patients as a basis to analyse the language of people experiencing depression. The ethical approvals required for this, whilst extensive, could have been achievable. However, during the course of my pre-interview research, I decided not to target active patients for my data for three reasons. The first is that I wanted to concentrate on low mood rather than depression, to enable greater scope for impact and applicability. An active patient diagnosed with depression would not necessarily be experiencing low mood on the day of their therapy session and could therefore have skewed my analysis. The second reason is that the level of anonymisation required to access data of this kind would have prevented me from a comprehensive understanding of the demographic data. Finally, this method of data generation provided me with a reference corpus that was compatibly comparable, which would not have been feasible for therapeutic transcripts.

It was useful to keep a reflexive diary throughout the main study interviewing process. My attitudes on the day as well as previous power dynamics between myself as interviewer and the informant as my interviewee have the capacity to affect the interview itself (Lazard & McAvoy, 2017). Before the interview, I took an opportunity to pause and consider my positionality in relation to each informant. I wrote a line detailing my feelings and potential power dynamics in this reflexive diary as a result of this reflection. Afterwards, I did the same, writing a line of debrief for myself. This practice gave the informant time to complete the questionnaire and gave me a clear sense of closure and separation from the interview. Since it was necessary to time the interviews around informant's work schedules, there was little opportunity for reflection or space away from the research beyond this.

Some of the interviews were lengthy, most likely due to my decision not to include conversational cues that typically encourage a sentence to come to an end. I do believe that avoiding conversational cues was the correct decision, as my data was free from interruption and the informants were empowered to answer the questions fully. Moreover, despite the clear participant information form, I am concerned that some informants may not have understood the purpose of

the study. For example, one informant after a particularly lengthy interview told me that it had been a pleasant catch up. This is a reflection of the existing relationship with my informants, and in some cases seemed to encourage more verbal data from them.

I designed the questions after reflecting on my pilot study, but the structure of one main question and three follow up questions still left room for improvement. My original aim was for the first follow up question to be *what does that mean to you?* which would refer to the response they had just given in the main question. But a detailed answer to the main question rendered this nonsensical, and so I amended it to *what does importance mean to you*. Whilst the answers were fascinating, in some way it led the responses away from the main question. This resulted in the necessity of referring back to the main question during the two subsequent follow up questions, which was temporally leading. I was attempting to avoid temporal references in order to ensure the exploration of my second research question: *Do people with low mood focus on the past?* was not led by the interview questions. By asking informants to think back, it was temporally priming them towards a view of the past. Since my data did not show a clear past focus, this change of question does not appear to have caused an impact.

From an ethical perspective, I did not realise how difficult it would be for me to discover that good friends are hiding marked low mood. Scheduling was also an issue from informants: some changed their minds; were unexpectedly too busy after initially volunteering their time; were running late or simply did not arrive. As Barratt (2021, p.231) writes, papers on interview styles do not explicitly discuss the logistics of scheduling, and to a lesser extent, recruitment. This difficulty around scheduling was therefore not anticipated, and led to me feeling rejected. Despite these challenges, all data was gathered and transcribed by the end of December 2022.

Chapter 4 Data summary

4.1 Overview

The focus of this thesis is my research questions: *What linguistic markers are prevalent in the verbal language of people experiencing low mood?* And *Do people with low mood focus on the past?* This chapter provides an understanding of the data that I use in order to address the research questions. The previous chapter was devoted to the data generation methods, while this chapter is the content and metadata that was generated. The next chapter will cover the analytical methodologies and chapters 6 to 9 are the analysis itself. In this chapter, I include a statistical breakdown of the data and an outline of the demographics of my informants. From the verbal language of the informants themselves, I discuss the topical content of their responses to my prompts. I complete this chapter with a discussion on the structure of informant responses.

I address the pilot and main study separately through this thesis, because the pilot study is a separate study in its own right and can not be grouped with the main study data. There are two main reasons for this. The first is the difference in the questions I asked each informant immediately after the main question. Specifically, I chose from a list of follow up questions during the pilot study, based on the responses provided to the main question. This meant that my pilot study informants could have anywhere from one to six follow up questions, tailored to the informant from a list of six possibilities. The lack of uniformity in this approach was useful for a pilot study but separates the data generation method from that of the main study, which involved the use of the same three follow up questions each time. The uniformity of the main study questions ensures comparable data, when I later explore the language of the low mood group when compared to the reference group.

The most important reason for the pilot study remaining separate is the mental health categorisation questionnaires, which are not measuring the same thing. When performing the pilot study, I had still intended to examine *depression* as a medical condition rather than *low mood*. I therefore used a categorisation questionnaire based on the symptoms of depression, specifically the OHMS measure. Based on my research into the long term viability of depression as a diagnosis (please see literature review 2.4), I subsequently decided to use a different questionnaire for the main study.

The questionnaire I chose was designed to benchmark mood and feelings (MFQ) rather than depression specifically. Please see Appendices E and F for the full OHMS and MFQ categorisation questions, to examine the difference between the pathologized and emotion based questions. Throughout this project I attempt to emphasise the distinction between depression as a diagnostic label and the psycholect associated with the embodied experience of low mood via its linguistic expression. To combine the depression-based pilot study and the mood-based main study would detract from this distinction.

The focus throughout this chapter as well as the rest of the thesis is on the main study. This is because the main study data was gathered in a robust and replicable way. However, because the pilot study data findings are in line with other studies, I have chosen to use this as a comparative measure. It is useful to compare the main study data to other studies to examine the differences. Researchers on the studies from my literature review have made their results available but have not yet also released the source data. Since the source data for the pilot study is available to me it can provide a counterpoint to the results of my main study.

4.2 Corpus metadata

The pilot and main study data are both transcribed verbal responses to the same question, alongside different follow up questions for the main and pilot studies. This question is “when you think about what is most important to you, what comes to mind?”. The main study involved 52 informants whose transcribed responses totalled 52,101 words with the shortest interview word count of 74, and the longest of 4,835. The pilot study involved 11 informants and transcribed responses totalling 8,459 words, with the shortest interview word count of 335 and the longest of 2,098. For a thorough guide to the informant recruitment and data generation process, please see the data generation chapter 3, specifically sections 3.4 for the pilot study and 3.5 for the main study.

I designed the pilot study primarily to test data generation techniques. A secondary purpose was to trial analytical techniques on a small scale. I also wanted to see if the pilot study results indicated any initial trends in the data. These trends did fit existing studies covered in the literature review.

However, as I will go on to show and explore, the trends in the main study data are generally less aligned with previous research. This will be explored in more detail in analytical chapters 6-9 and the discussion chapter 10.

Twelve main study informants were experiencing low mood at the time of the interview, according to the categorisation questionnaire. Their responses comprise the Target Low Mood (TLM) data set, whilst the responses of the 40 informants who were not experiencing low mood comprise Reference Low Mood (RLM). This means that TLM is 23% of the main study data whilst RLM is 77%. Three of the pilot study informants had an OHMS score indicating that they may be experiencing depression (see Appendix C). The verbal responses of these three potentially depressed informants are the Target Pilot Depression (TPD) data set. The responses from the eight pilot study informants who did not register as potentially depressed the Reference Pilot Depression (RPD) reference corpus. This means that TPD is 27% of the pilot study data whilst RPD is 73%.

In the main study post interview categorisation questionnaire I gave informants the option to self-identify as depressed. Whilst this thesis is not based on depression as a diagnostic criterion, I was interested to see if there was a correlation between self diagnosis and mood scores. I did not define depression on this questionnaire, leaving the individual informant the option to decide their own definition of depression. Informants were given the option to select *Depressed*, *Not Depressed* and *Other*, alongside a free text box to clarify if *Other* was selected.

Self-identification of depression	TLM	RLM
Depressed	9	4
Not depressed	1	30
Other	2	6
Total	12	40

Table 4.1 Main study identification of depression by corpus

In TLM, the low mood cohort, nine of the informants felt that they were depressed at the time of the interview and one felt that they were not depressed. The two TLM informants who selected the *other* option both used the free text box to advise that their feelings of depression are transient. In RLM, the reference group, four selected the option to advise they were depressed at the time of the interview whilst thirty said they were not. The six informants who identified as *other* used the free

text box to clarify that they felt as if their depression was under control at the time of the interview. Depression as it is currently defined in psychology (e.g. WHO, 2022) is a more complicated concept than that of low mood. This psychological definition of depression involves a list of symptoms that need to be experienced over a period of time in order to be diagnosable. An individual enduring serious and potentially life-ending low mood may not be diagnosable with depression but should be classified as in need of help. Worldwide figures report that 10% of people who are successful in their suicide attempts have no psychiatric diagnosis at all (Bertolote & Fleischmann, 2002, p.183). The amelioration and therapeutic need can therefore be equal in severity to a patient who fits current criteria for a depression diagnosis . As for my main study, a combination of self-awareness, willingness to accept a diagnostic term and mood at the time of the questionnaire appears to have created a mixture of responses within this self-assessment. I will discuss this further in the context of the ongoing viability of depression as a diagnosis in chapter 10.

The main study word count for individual informants ranged from 74 up to 4,835, with an average of 1,001.9 and median of 799.5. The 202.4 word difference between the mean and the median shows the variability of the length of informant speeches. I endeavour to ensure that this variability is taken into account in my analyses, to address the impact variability has on generalisability. The pilot study word count varied from 335 up to 2,098, with an average of 769, and a median of 678. As with the main study data, this lower median is as a result of the highest word count skewing the average.

When analysing data using corpus linguistic techniques, the variety of response lengths is something that needs to be taken into account. The difference in mean and median across both studies highlights this variety. It would be possible to falsely attribute a greater level of significance to linguistic patterns of the loquacious informants over those who were more concise. However, in the words of Sinclair, "underlying regularities have a better chance of showing through the superficial variations" (2004, p.189). With a corpus comprised of 52 informants, this imbalance is mitigated by volume, whilst simultaneously enabling visibility of variation.

4.3 Demographics

Data generation methodology for the main and pilot studies is detailed in chapter 3. To provide an overview of the data, here is the demographic breakdown of each study normalised using percentage-based (rounded to 0 decimal places) breakdown. I have drawn attention to the highest proportion of each group by writing the percentage in bold. For example, in the pilot study, there were more informants aged 18-20 than any other age, and so the 64% figure is in bold.

	Grouping	Pilot % of 11 informants	Main % of 52 informants
Age (yrs)	18-19	**64%**	2%
	20-29	36%	8%
	30-39		**38%**
	40-49		**38%**
	50-59		13%
Gender	Male	27%	40%
	Female	**73%**	**56%**
	Non-binary		4%
Race	White	55%	98%
	Black	45%	
	Indian		2%
Education	GCSE		6%
	A-level		15%
	Semi-degree	**82%**	10%
	Degree		**42%**
	Half masters	18%	8%
	Masters		15%
	Doctorate		4%
Mood	Low/depressed	27%	23%
	Non- low/depressed	73%	77%

Table 4.2 Comparative summary of main and pilot study informants

Further to Table 4.2, in the main study informants were aged between 18 and 56, whilst in the pilot study informants were aged from 18 to 26. The average age was 40 for the main study and 20 for the pilot. In terms of gender, 29 informants identified as female in the main study, 8 in the pilot. 2 informants identified as non-binary in the main study and 0 in the pilot whilst 21 main study informants identified as male and 3 did for the pilot study. In terms of race, for the main study 51 informants identified as white and one as Indian whilst in the pilot half the informants identified as white and half as black. Education level in my main study was varied, from those who had not progressed beyond the lowest levels of qualification available in the UK through to those who had achieved Doctorates. In the pilot study informants were all students studying towards an undergraduate or masters level qualification. I did not ask about sexuality in my pilot study, whereas

within the main study seventeen (33%) of my informants identified as “fluid”, “Queer”, “Bisexual” or “Pansexual”, though no informants overtly identified as exclusively attracted to their own gender.

Since the entries for the classification questions on gender, sexuality, race and education level were free form text boxes, there are a range of responses, particularly to the sexuality and race questions. For the data analysis I have grouped each of the classification types. For example, in the *education* section, *part of a masters* includes informants who are still studying towards a master’s degree or who have completed a qualification that is between the level of undergraduate and masters. Whilst these groupings enabled analysis of the language of the demographic collectives, the terminology chosen by the informants could be a source of linguistic analysis in its own right.

I was surprised that the majority of main study informants were educated to degree level or above. Based on the contacts on my social media from mandatory schooling, I had expected a higher volume of individuals who were educated to GCSE or A-level only. My theory as to the reason for this is a combination of disinterest and homophily (Crawford et al., 2018, p.153). In my social media advertisements I wrote that the interview was for a PhD project. My informants were mostly those who also had undergraduate degrees: homophily at work.

In providing free form text boxes for demographic categorisation, I enabled the informants to identify themselves in their own words. The responses for gender were 2 nonbinary, 28 female, 21 male and 1 cisfemale (sic). Specifying gender rather than sex meant that any informants who are trans and identify as the opposite gender to the one they were raised as were able to do so without referencing their birth sex. This is affirming for trans men, trans women and those outside of the gender binary. However this does not take into account the sex an informant was assigned at birth and therefore the influence of early socialisation as a different gender. To clarify, if one of my non binary informants was assigned female at birth, their experience of society up until the point that they socially presented as non binary would be different to that of a non binary individual who was assigned male at birth (Stockard, 2006, p.215). I have deliberately chosen gender over gender and sex, and investigating the trans experience specifically is outside of the remit of my study, and capacity of my data. This could be an avenue of exploration for future research.

Table 4.3: Main study informant demographic summary

	Criteria	Informants
Age range	18-19	1
	20-29	4
	30-39	20
	40-49	20
	50-59	7
	Average	40 yrs old
Gender	Male	21
	Female	29
	Non-Binary	2
Sexuality	Straight	10
	Heterosexual	24
	Mostly hetero	1
	Fluid	1
	Bisexual	11
	Pansexual	4
	Queer	1
Race	White	9
	White British	23
	White English	1
	White Irish	1
	White Welsh	1
	White European	1
	Caucasian	6
	Human	3
	British	5
	English	1
	Indian	1
Education	GCSE	3
	A-Level	8
	Part of degree	5
	Degree	22
	Part of masters	4
	Masters	8
	Doctorate	2
Community	LARP	41
	Work	6
	Family	1
	Other hobbies	4

This table features the precise demographics for the main study informants only.

The free form text box approach to self-identification provided a range of responses for sexuality, which have been outlined in Figure 4.1.

Fig 4.1: Sexuality wordcloud

The capacity to self-define gave these informants the power to be nuanced in their descriptors for sexuality. When examining this data, there are no informants who identified as purely attracted to their own gender. I therefore separated the informants into those who identify as straight and those who identify as queer. Even under the banner of heteronormativity, there are 5 different terms for those who identified as definitely straight, with another term that was borderline. The informant who chose to identify as “mostly hetro” (sic) was difficult to categorise, since if one is mostly heterosexual, it leaves the possibility that one is partially not heterosexual. Whilst corpus linguistics is useful for examining large swathes of data, some of the nuances can be lost. The kind of fine grained linguistic analysis that would enable illumination on the subtlety of these sexuality based identities is not easily accommodated by corpus linguistics (Brookes & Collins, 2023, p.42). In this data set, I chose to simply group the informants into mostly straight and mostly queer, thereby including ‘mostly hetro’ (sic) in the former category. This is in recognition of the primary aim of the study, which is to identify the dominant patterns across the dataset.

Whilst this is outside of the parameters of this study, I would like to draw attention to the three people who wrote *human* when asked to describe their racial identity. In my pilot study each informant clearly stated their ethnic origin, including the person who chose to write *white- other*, rather than specifying their nationality. This led to my belief that asking for race was a clear way to categorise that would be simple and unlikely to be misunderstood. Each of the people who wrote *human* in this

main study were white British, and based on their communication afterwards this was an attempt at humour. This could be part of the downside of working with informants who have an existing relationship with the researcher, and/or it could be a lack of clarity over the purpose of the question. In any case, the informant demographics do not reflect the racial diversity that I had hoped for. Once again, this homogeneity is likely due to homophily whereby my racially diverse friends self-excluded. My aim was for my informants to be representative of the population of the UK and to therefore constitute a range of ethnicities, but I was not able to achieve this. This is an area where an expansive future study could enable targeted recruitment and therefore the opportunity for greater representation and inclusivity.

To further explore the intersectional demographic breakdown of my main study informants, I have tabulated the intersection of demographic responses grouped according to low mood (TLM) and reference (RLM) informants. Tables 4.4 and 4.5 follow on the next two pages, and indicate that the intersectional spread of demographics are well spaced out across both TLM and RLM. To clarify, there is no general skew towards a particular gender, sexuality or education level in either sub-corpus. The only potential intersectional bias is that the two informants educated to doctoral level are both straight women and in the reference group.

Table 4.4 TLM intersectional informant demographics

		Age range					Gender			Sexuality		Education						
TLM	Criteria	18-19	20-29	30-39	40-49	50-59	Male	Female	Non-binary	Straight	Queer	GCSE	A-Level	Part of degree	Degree	Part of Masters	Masters	Doctorate
Age range	18-19																	
	20-29		2					2			2				1		1	
	30-39			7			3	3	1	5	2	1		1	5			
	40-49				2		2			1	1				2			
	50-59					1		1			1			1				
Gender	Male			3	2		5			4	1	1			4			
	Female		2	3		1		6		2	4			1	4		1	
	Non-Binary			1					1		1			1				
Sex-uality	Straight			5	1		4	2		6		1			5			
	Queer		2	2	1	1	1	4	1		6							
Education	GCSE			1			1			1		1						
	A-Level																	
	Part of degree			1		1		1	1		2			2				
	Degree		1	5	2		4	4		5	3				8			
	Part of masters																	
	Masters		1					1			1						1	
	Doctorate																	

Table 4.5 RLM intersectional informant demographics

		Age range					Gender			Sexuality		Education						
RLM	Criteria	18-19	20-29	30-39	40-49	50-59	Male	Female	Non-binary	Straight	Queer	GCSE	A-Level	Part of degree	Degree	Part of Masters	Masters	Doctorate
Age range	18-19	1					1			1			1					
	20-29		2					1	1		2		1				1	
	30-39			13			6	7		9	4		3	2	5	1	1	1
	40-49				18		6	12		15	3	1	3	1	7	2	3	1
	50-59					6	3	3		4	2	1			2	1	2	
Gender	Male	1		6	6	3	16			15	1		3	2	6	3	2	
	Female		1	7	12	3		23		14	9	2	4	1	8	1	5	2
	Non-Binary		1						1		1		1					
Sex-uality	Straight	1		9	15	4	15	14		29			5	2	12	4	4	2
	Queer		2	4	3	2	1	9	1		11	2	3		2	1	3	
Education	GCSE				1	1		2			2	2						
	A-Level	1	1	3	3		3	4	1	5	3		8					
	Part of degree			2	1		2	1		2				3				
	Degree			5	7	2	6	8		12	2				14			
	Part of masters			1	2	1	3	1		4	1					4		
	Masters		1	1	3	2	2	5		4	3						7	
	Doctorate			1	1			2		2								2

4.4 Topics

The main question that I asked my informants was: *When you think about what is most important to you, what comes to mind*? This section is an overview of the topics that the main study informants stated that they find the most important. In order to catalogue topics, I read though each transcription and wrote down each topic that was described as important. For some informants, this was a simple task because they listed things in a simple way: for some more ambiguous responses I needed to carefully evaluate their language. For example, from Informant126:

> Quote 4.1: obviously I do a lot of larp and I'm gradually coming to the conclusion that I prefer to be. an NPC or a crew member because then I'm one of the people who's. telling the story or providing the story.

In quote 4.1, I chose to disambiguate storytelling as the thing that the informant considered most important. The informant had spoken about the importance of storytelling in the previous section, and it was clear that in this section their focus is not on the hobby itself but in the creative storytelling process. In order to count the topics of important, I loosely combined those with similar wording or meaning, for example, I included a husband, wife, girlfriend or partner under the heading of *Partner*.

I have grouped the TLM (low mood) and RLM (reference) corpora together for this section, and will perform a comparative topic analysis in 6.3. The lowest number of topics that my main study informants spoke about was two, the most was thirty and the average was eight.

The TLM and RLM informants classified the first, second and third most important thing to them in different ways. In Table 4.6, I include a list of each topic in order of popularity alongside a percentage figure for the percentage of TLM and RLM informants who chose this. For example, 32.69% of TLM and RLM informants said family as their first response, when asked what was most important to them.

TLM & RLM combined most important		TLM & RLM combined second most important		TLM & RLM Combined third most important	
Family	32.69%	Friends	17.31%	Friends	10.20%
People	7.69%	Family	15.38%	Family	6.12%
Child/ren	7.69%	Happiness	5.77%	Health	6.12%
Friends	5.77%	Partner	5.77%	Loved one's happiness	6.12%
Contentment	3.85%	Pet	3.85%	Connection	4.08%
Happiness	3.85%	Child/ren	3.85%	Happiness	4.08%
Partner	3.85%	Connection	3.85%	Larp	4.08%
Relationships	3.85%	Loved ones	3.85%	Money	4.08%
Being liked	1.92%	A better world	1.92%	Loved ones	4.08%
God	1.92%	Being a good person	1.92%	Self-actualisation	4.08%
Green space	1.92%	Being a role model	1.92%	Support	2.04%
Career	1.92%	Being supportive	1.92%	Calm	2.04%
Health	1.92%	Change	1.92%	Child/ren	2.04%
Home	1.92%	Choose your own destiny	1.92%	Chosen family	2.04%
Caring for parent	1.92%	Community	1.92%	Creativity	2.04%
My stuff	1.92%	Right no. of family/friends	1.92%	Silliness	2.04%
Myself	1.92%	Decarbonization	1.92%	Environmental concerns	2.04%
My rehab	1.92%	Equality	1.92%	Feeling wanted	2.04%
Safety	1.92%	Family health	1.92%	Helping people	2.04%
Self-improvement	1.92%	Feeling valued	1.92%	Integrity	2.04%
Self-actualisation	1.92%	Health	1.92%	Work	2.04%
Purpose	1.92%	Existence enablers	1.92%	Justice	2.04%
Self-worth	1.92%	Politics	1.92%	Kindness	2.04%
Stability	1.92%	Relationships	1.92%	Love	2.04%
		Self-honesty	1.92%	Make world better	2.04%
		Sense of control	1.92%	Memories	2.04%
		Stability	1.92%	Music	2.04%
		Stories	1.92%	Others' happiness	2.04%
		Time alone	1.92%	Self-improvement	2.04%
				Partner	2.04%
				Taking care of people	2.04%
				Time	2.04%
				Time alone	2.04%
Total responses	52	Total responses	52	Total responses	49

Table 4.6 TLM and RLM Topics first, second and third response

Family, friends, children, happiness and partners were all particularly popular as the most important thing to my informants. One informant said that the second most important thing to them was "the things that allow me to exist so. uhm for example my. headphones", which I have classified as *existence enablers* for the purpose of this chart (the only use of this category). I found two of second responses surprising during the interviews, one was *decarbonisation*, the other was *the right number of friends and family*. Most informants felt that friends and family were important, but the latter response struck me as distinctive, and the informant did not clarify on what they meant by this. It is

noticeable from this chart that the topics of importance become more idiosyncratic as the informants progress beyond their initial thought.

I have grouped all the rest of the informant responses into a word cloud, since many of them were unique and therefore not suitable for display in the same way as the top three responses.

Figure 4.2: TLM and RLM grouped Topics 4th-30th responses

These detailed responses provide an opportunity to see the wide range of things that informants described as being important, after their top three responses. These feature the most idiosyncrasies, and show the range of interests and vocabulary that my informants prioritise. Friends continue to be popular, but more specifics begin to be mentioned such as academic interests, general aspirations and wider global concerns. I further group the topics that informants talk about as part of a

comparison of TLM and RLM responses in section 6.3. However this section demonstrates the nuances of the topics that individual informants found important. This speaks to the importance of an individualistic as well as grouped approach to my linguistic analysis.

4.5 Narrative

Narrative is a term used in the experiment design of some of the studies outlined in the literature review (e.g. Jaworska & Ryan, 2018) and since these studies were successful my original aim was to generate narrative data for this project. From a linguistics perspective, Toolan describes a narrative as "a perceived sequence of non-randomly connected events" (1998, p.7). From psychology, Dennett (1992) argues that if one were to translate the physics-based concept of gravity to narrative, the self would be the epicentre. To clarify, the self is not an easy thing to define but narrative can enable us to orientate around what an individual is or stands for. Examining a narrative data set for language patterns therefore appeared to be a logical way of getting closer to a sense of the individual and their low mood. This became an established practice in sociolinguistics, since the constructionist “narrative turn” (Tirvassen, 2019, p.101).

Labov (2013) wrote extensively on personal narrative and their work led me to believe that this was a good way to get rich verbal data. In the forms to obtain ethics approval (Appendices C and D), the term *narrative* is heavily featured, since I expected my data to resemble a narrative form. Whilst performing exploratory analysis, it became apparent that *narrative* is not an appropriate term for my data. As discussed in 3.4.3, I based my main data generation question on one posed by Grob et al. (2016, Table 2). The foundational question was designed to encourage their participants to talk about their experience with a healthcare provider and provided in depth and meaningful responses. There was no linguistic data shared from this original experiment that provided any insight as to the narrative nature of the responses. Based around my reading, particularly Labov (2013), the question appeared to lend itself towards a narrative or storytelling-based response. I aimed for a narrative response because Gross et al. (2016, p.1252) wrote that narratives engender complete, in depth and balanced data. They clarify that narrative responses can be more meaningful than alternative

methods of data generation such as quantitative studies. This is why I chose the main interview question: *when you think about what is most important to you, what comes to mind?*

During the pilot testing phase, narratorial response responses were frequent within the data, as shown in pilot quotes 4.2 to 4.4.

> Quote 4.2: We had a really big friend group at first, but then it got smaller and smaller as we all stopped hanging out. But the two of us have just stayed
>
> Quote 4.3: I was sort of a bit of a loner really for quite a long time. But then, I found a community of people that, I feel safe within
>
> Quote 4.4: get away and being like at peace and home and going for drives and watching the sunset with my friends and, going for picnics at the beach

I categorised these examples as narrative due to the unfolding events each informant outlines. In quote 4.2, the event is being in a big group of friends. The events that follow are the reduction of the friendship group ending with two individuals, which completes this small story. In quote 4.3, the narrator sets the scene with a loner who is used to their own company, building in a new event of adding community to the picture and then finishing the story with the image of safety and belonging. The last example, quote 4.4, is a list of events, with more of a sense of narrative elements linked through the sense of home that the narrator is longing for.

Based on close reading of the pilot study transcripts, the follow-up questions that were temporally leading were those most likely to lead to a narrative-based response. For example, “does importance change over time” led to many stories in response. This is because narrative is “discourse that communicates information keyed to a temporal framework” (Hühn et al., 2014, Diegesis – Mimesis section 1.Definition). Time, specifically temporal focus, is therefore an integral part of narration. I chose to omit the question about how importance changes over time from the main study because it had too much chance of priming informants towards thinking and therefore talking about the past. This was to prevent a bias towards the hypothesis underlining my second research question: *Do people with low mood focus on the past?* In order to consider if something had changed, informants would need to think about the past, the present and then think forward to the future. This would have enabled data generation that was primed towards temporal thinking, which would have affected the temporal focus of informants’ speech.

In order to avoid temporal priming, the questions I posed of my informants may have inadvertently deterred them from narrative responses. In order to tell the sequential events of a story it is necessary to place oneself into a lived past experience or hypothetical present/future event. As Fludernik (2009, p.44) writes, “The narrative discourse then turns back to events which happened previously (analepsis) or jumps forward to discuss what will happen later (prolepsis)”. This time-hopping was actively discouraged in the question design and therefore affected the narratorial element of informant responses. However, whilst I chose a narrative eliciting question specifically in order to provide sufficiently lengthy and rich data, the resulting dataset suggests that my informants were able to provide a wealth of data for analysis. This wealth of data was irrespective of the narrative component; in this sense, a lack of narrative response is not innately problematic.

Some of the main study responses were narratorial. However, many were more likely to involve lists or in one case just an individual word. An example of each follows with a narration (quote 4.5), a list (quote 4.6) and an individual word list (quote 4.7) respectively.

Informant 128 narrates the reason friends are important; as this is a long speech I have included an excerpt:

> Quote 4.5: I suppose from being Junior School onwards uhm I had. a couple of very close friends but I didn't have a wide social circle. uhm as a. a teen and and into my early 20s late teen into my early 20s uhm I discovered that uhm. I had a wider social circle because of the hobbies that I used to do so I used to go ice skating. uhm but it was speed skating with all the boys. so it wasn't uhm wasn't figure skating it wasn't pretty. it was dashing round the ice at top speed uh trying not to fall over uhm but that created uh a wider social circle outside of people that I knew from school… uhm and I enjoyed it I enjoyed having. a wider social circle… then when I met my ex-husband now uhm at that time he had uhm he's from quite a big family so there's. eight children uhm so our social circle. narrowed to being. his family.

Informant 76 lists and expands on things that are important without providing any events to confirm this:

> Quote 4.6: People friends and family mostly… uhm… cats… uhm. are kind of people uhm… yeah mostly it's it's people. who I can trust and a community or a family.

Informant 120 simply listed words that represent the things that are important to them:

> Quote 4.7: Family. health. happiness… peace. equality. fairness. justice. Friendship

It is clear from this that the majority of my informants did not respond to this prompt (Grob et al., 2016, Table 2) with a narratorial response. I have discussed that the lack of temporal direction in the follow up question may have been the reason for this lack of narrative, though this may also be due to my inclusion of the word *most*. This superlative encourages the informant to place an order of priority on things that they consider important, but does not encourage a narrative structure.

My data may not have been as narrative in structure as I had anticipated, based on my research for the interview question. If I was to perform a similar or follow up study, I would keep the same main interview question, despite the lack of narrational data generated. This is because I have a valid dataset that has provided fruitful analysis and potential for future similar studies.

4.6 Vocabulary

The vocabulary of the informants within my study is markedly different to the terminology of participants quoted in the studies cited in the literature review section 2.6. Whilst many of the studies from the literature review do not include linguistic examples, those that do involve high frequency lexis, associated with low register language. As an example, the disclosures of adolescents as quoted from their internet forum posts “i am forever hungry i may have a habit of eating i dont know wot is wrong with me and i am not fat i am skinny kind of just rite really” (Harvey, 2012, 363). If one discards the spelling and grammar of digital discourse, since this is not relevant to spoken data, the simplicity of this language is clear. As a comparison, one of my informants spoke on a similar topic: *they are always. sort of. nebulous… want to be fitter I want to eat healthier*. The difference in vocabulary here is marked. Whilst the adolescent is showing self-awareness, they do not appear to have the lexicon to express this as fluently as the informant in my data, who uses polysyllabic Latinate lexis *nebulous* alongside superlative structures.

The difference in communication style for the adolescent posting on the internet and my informant may be due to their relative age and education differences. A different example as quoted from a depression forum for adults “I was terribly shy and self-conscious and was tall and fairly advanced for my age” (Hunt & Brookes, 2020, p.128). The spelling and grammar for this quote are significantly

closer to written standard English, and the polysyllabic expressions such as *self*-conscious and *advanced* are more complex. Whilst the term *terribly* could be considered archaic, the concordances around collocates still portray a greater linguistic sophistication. The vocabulary of my informants is closer to this example, as exemplified in the word cloud of topics in Figure 4.2, section 4.4. I theorise that the nature of the academic context of the interview may have affected the vocabulary style. My informants were aware that they were taking part in a PhD study and could have chosen to use an extended lexicon as a result. Whilst the elevated register of my main study informants is distinctive, it is similar across the low mood and reference sub-corpora and therefore was comparable using corpus linguistic techniques.

4.7 Data summary review

This data summary chapter has provided an overview of the form and contents of the data generated using the methods discussed in the previous chapter. The size of the corpus, both overall and between informant contributions, is modest by modern standards (Hunston, 2018, p.20). However I believe that the differences are valuable for contrasting with previous research demographics. Due to the limitations in size and composition of my corpus, I aim to be careful in handling quantitative as well as qualitative data analysis approaches. I outline these methodological approaches in chapter 5, and use them in chapters 6 to 9.

Chapter 5 Analytical methodology

5.1 Analytical methodology introduction

This is an overview of the methodologies I use as part of the analysis of the data discussed in chapter 4. Each analysis chapter (chapters 6, 7, 8 and 9) includes an expansion on the specifics of how I implement the methodology for that section.

I have chosen to use mixed methods for this study, combining qualitative and quantitative perspectives to examine the data thoroughly. I make use of, in the words of Madill and Gough (2008, p.254) the “permeability of the quantitative-qualitative boundary”. Methodological plurality is particularly well suited to a corpus linguistic approach to health communication (Brookes & Collins, 2023, p.2). Since corpus linguistics and the theoretical frameworks I use are nuanced, a review of how they work generally will provide a foundation for the analysis chapters.

5.2 Corpus linguistics

The interview and transcription process outlined in chapter 3.5 provided a data set of 52,101 words. I grouped the informants who spoke these words using the categorisation questionnaire at the end of the interview. This enabled me to separate the informants according to mood and demographics. The main separation is that the speech acts of the 12 informants experiencing low mood at the time of the interview were grouped into a corpus labelled TLM (target low mood) which consists of 10,442 words. The language of the 40 informants not experiencing low mood at the time of the interview were grouped into a comparison corpus called RLM (reference low mood) and consists of 41,659 words.

The size of data set produced during this study lends itself well to corpus linguistics. Corpus linguistic analysis enables one to examine large bodies of qualitative data from a quantitative, statistical standpoint. This data driven analysis enables pattern recognition whilst minimising personal bias in the initial stages. I cover the development and background of corpus linguistics in the literature

review, section 2.3. Whilst the data set is large enough to enable corpus linguistic analysis, it is relatively small when compared to data sets such as the British National Corpus (BNC Consortium, 2007) which is 100 million words. Throughout this thesis I am aware that the small size of this corpus affects the outcome of the analysis. With few words and informants, small dialectical changes affect each comparison, therefore preventing these results from being generalisable. However, as the title of this thesis indicates, I am still able to search for tendencies that are indicative of a low mood psycholect.

5.2.1 Keywords

A linchpin of corpus linguistic analysis is keywords; these are words that are characteristically more likely to be present within a target corpus when compared to a reference corpus. In this study, the target corpus is verbal data informants experiencing low mood at the time of interview (TLM), whilst the reference corpus is the verbal data of informants who did not have low mood at the time of interview (RLM). The keywords that are more likely to appear in the target corpus than the reference corpus are often used to indicate the 'aboutness' of a corpus (Groom, 2022). Keywords are therefore often used as a starting point for linguists to provide data driven insight into a text.

To find keywords in my data, I use *AntConc* (Anthony, 2010) software to compare the TLM and the RLM corpora. My primary reason for choosing *AntConc* (Anthony, 2010) is the functionality, particularly the capacity for customisation. This customisation includes the ability to input your own data and then choose how keywords are calculated. To calculate keywords, the software performs significance testing to establish which words are more likely to appear in the target when compared to the reference data. There are a range of formulae used to calculate keywords: log-likelihood looks purely at the frequency of a word, whilst t-test and Whitney & Mann consider distribution of words as well as frequency (Egbert & Biber, 2019). Based on the range of interview lengths, I used chi Squared which is currently favoured by many linguists as a preference to log likelihood (Groom, 2022) and this measure is also available in *AntConc*.

5.2.2 Part of Speech tagging

Another element of linguistics that can be automated with corpus linguistics software is Part of Speech (PoS) tagging. We have established in section 5.2.1 that it is possible to statistically analyse text to find the significance of individual words. PoS tagging involves grouping individual words into grammatical categories, also known as parts of speech. These grouped categories can then be compared across corpora, to establish the significance level of a category in a target corpus when compared to a reference corpus.

I use Constituent Likelihood Automatic Word-tagging System (CLAWS) which is accessible through *Wmatrix* (Rayson, 2009) to perform automated PoS tag analysis on my data. *AntConc* does not have an automated PoS tag analysis tool, which is why I have chosen to use *Wmatrix* for the PoS tag analysis. This software gives each PoS tag a code, for example *JJT* refers to a general superlative adverb such as *oldest* or *best*. These tagged codes are then automatically compared across corpora to provide a ranked list of categories. The ranking is in order of keyness, and this enabled a swift comparison of grammatical categories across data sets.

5.2.3 Prototypical texts

In order accommodate the fine-grained analyses in chapters 8 and 9, I needed a sample of my data rather than the full corpus which is used in chapters 6, 7 and part of 8. Whilst I was researching for this thesis, Anthony & Baker (2023) developed a programme (*ProtoAnt*) that enables the researcher to rank individual texts in order of representativeness. *ProtoAnt* (Anthony & Baker, 2023) takes a list of keywords as a primary input then analyses them against a group of individual texts to work out which best represent that list of keywords. In my case, the group of individual texts to be ranked were the transcribed informant interviews, separated out by informant. I used *Wmatrix* (Rayson, 2009) corpus linguistic software to find the initial keyword list to input into *ProtoAnt*. Using the software I was able to rank the informant interviews in order of prototypicality. That is, it showed me the informant utterances that were most representative of the entire cohort, in order of representativeness.

I began by taking the first 74 words of every main study interview- both TLM (low mood) and RLM (reference) corpora. In order to find the most prototypical texts of all the data, I needed the keywords that represented all the data, which is why I combined TLM and RLM. I chose to sample the first 74 words only from each informant, to prevent the varying interview lengths from skewing the analysis towards the more loquacious informants. The choice to sample texts to enable equal representation for each informant is accepted practice in corpus linguistics (Brookes & Collins, 2023, p.47) The shortest interview was 74 words long and therefore sampling the first 74 words from all ensured that all the texts were exactly the same length. I chose the first 74 words in the understanding that the gut feeling response is the least intellectually driven and therefore viscerally meaningful. The first instinct fallacy is the psychological understanding that we are most likely to stick with our first response to a question (Kruger et al., 2005). I could not find any studies related to how meaningful the first response is, but it is the most lasting response.

The first 74 words of each informant created a total corpus that was which was 3,848 words long, which is modest, but this was the balance I chose to ensure a fair representation of all informants. I compared the sampled TLM and RLM combined corpus to a sample of the spoken word section of the British National Corpus (BNC Consortium, 2007) in *Wmatrix* (Rayson, 2009). This is an existing corpus of 10 million words transcribed from a combination of informal and formal conversations in a range of situations. I chose to compare the TLM and RLM corpus to a sample of the spoken component of the BNC because this would provide a list of the words characteristically more likely to appear in all of my spoken main study data when compared to regular speech. This provided a keyword list of 100 words to input into the *ProtoAnt* software as the parameters for the prototypical search. These 100 keywords also provided insight into the linguistic preoccupations of my informants when compared to general speech. Table 5.1 features the positive keywords in order of significance with the number of times each word appears in the relevant corpus, normalised per 100 words.

Keyword	Sampled TLM & RLM	Spoken BNC
uhm	119.00	0.00
uh	1.00	0.00
family	1.40	0.02
pause	0.94	0.00
important	1.00	0.02
my	1.72	0.24
most	0.86	0.04
life	0.67	0.02
friends	0.51	0.01
happy	0.46	0.01
things	0.86	0.11
me	1.35	0.29
being	0.59	0.06
i_think	0.43	0.03
people	0.97	0.20
priority	0.16	0.00
kind_of	0.19	0.00
importance	0.16	0.00
health	0.24	0.01
question	0.32	0.03
guess	0.19	0.01
sense	0.19	0.01
thing	0.54	0.11
probably	0.38	0.06
because	0.70	0.20
without	0.24	0.02
that	3.48	2.15
so	1.29	0.57
i_suppose	0.16	0.01
and	3.93	2.57
myself	0.19	0.02
safety	0.13	0.01
comes	0.13	0.01
ooh	0.27	0.05
lots	0.16	0.02
mind	0.16	0.02
really	0.49	0.18
is	1.67	1.02
sort_of	0.27	0.07
other	0.35	0.11
of	1.99	1.30
kind	0.13	0.02
its	0.13	0.02
around	0.13	0.02
word	0.13	0.03
everything	0.13	0.03
also	0.19	0.06
to	2.29	1.69

Table 5.1 Keywords (norm. per 100 words) from sampled TLM and RLM against spoken BNC

Each of these keywords has a log likelihood figure indicating at least 99% chance of significance and appears more than 5 times in the sampled combined TLM and RLM corpus. I have used log likelihood in this case, because this is the measure that *Wmatrix* used and is the figure in the output for the keyword analysis.

The keyword list in Table 5.1, much like the topical data summary in section 4.4, is an insight into the topics that my informants of any mood prioritise when asked what is important to them. Some of these keywords, including the most significant: *Uhm* and *uh*, alongside *pause* are due to the differences in my transcription protocols compared to those used for the BNC spoken component. *Wmatrix* (Rayson, 2009) has combined some of the words despite there being no underscore connecting them in the transcript, for example *i_think*. No reason was provided for these word combinations within *Wmatrix* itself or the external help functions.

The keywords created from this sample are representative of the entire data set and therefore useful for finding texts that are representative of the whole corpus. This insight into my data when compared to the sample of the spoken section of the BNC shows the difference in data generation techniques. This can be seen by the synonyms and lemmatized version of *important*, which are prevalent in Table 5.1. This use of derivations of *important* is part of the informants endeavouring to answer the data generation question about what is important to them. Since they are based on the sampled combined TLM (low mood) and RLM (non low mood) corpora, they are not actively contributing to the main aims of this study, which are around low mood specifically. These keywords have proven useful in terms of examining the effect of my data generation, in terms of the topics people concentrate on once they are asked about what is important to them. The keywords are also essential for finding the most representative texts for manual coding.

I entered the full positive and negative keyword list based on the comparison of this sampling with the sample of the spoken section of the BNC into *ProtoAnt*. *ProtoAnt* used this data to create a list of all transcribed TLM and RLM interviews in order of representativeness. The calculations that were used to sort the interviews appear to favour shorter texts. My theory as to why this is, is that the shorter texts are more likely to feature a greater keyness for the keywords based on the first 74

words of all the texts when compared to the BNC sample. A list of the most representative interviews enabled me to perform a close reading of texts in order of salience. This provided me with an efficient way of testing the usefulness of the TTD (Svalberg, 2019) and the combination of attribution and appraisal theory.

5.3 Theoretical frameworks

I covered the background of attribution theory as part of section 2.4, however I will be using Leeds Attributional Coding System (LACS) in chapters 8 and 9. In chapter 9 I combine attribution with appraisal theory. As such, it is logical to lay out these frameworks in greater detail as part of this foundational methodology review. Attribution theory enables systematic analysis of the psychological features that are perceived to have caused emotions (Stratton 2003). Appraisal theory is a linguistic analytical framework whilst attribution theory is a psychological analytical framework. Appraisal theory enables one to systematically examine how situational emotion is linguistically constructed (Martin & White 2007).

5.3.1 Attribution theory

Attributional coding comes from the discipline of psychology, and involves considering dimensions that are perceived to characterise the causes of outcomes such as actions or emotions. Psychologists have used attributional coding for understanding therapeutic conversations, including in a family context (Stratton, 2003). Psychology academics have also combined attributional coding with corpus linguistics, including finding units of speech to use attribution theory on (Rae, 1993). Rae used early corpus linguistic techniques to find the words *because* and *cos* in spoken transcripts, which gave them a focus for their attribution coding. As part of my first research question, I use attribution theory to understand how informants linguistically characterise the causal mechanisms of importance and their situations or moods.

The dimensions of attributional coding are reliability, scope of outcome, area of origin, distinctiveness and level of controllability (Stratton, 2003, p.142). Each dimension has specific parameters that are coded on a relative basis. To clarify, the basis of coding is the dimensions

relative to other factors. These factors may be based on the informant, the situation and other persons or items that are involved. Every attribute is coded with valency taken into account, there is the option to consider causes as emotionally positive, negative or neutral. To understand more about how I coded the dimensions, I provide clarification on what is meant by each.

Reliability refers to how stable or unstable the cause of the emotion is. For example, finding happiness in a solitary specific walk in the countryside- the stable elements of this are the part of the countryside and the capacity for the individual to walk. If someone's favourite area of natural beauty was threatened or their mobility impaired, this would affect their perception of the stability of this source of happiness.

Scope of outcome refers to who the attributable cause of the emotion affects, this could be either a global phenomenon or something specific to the individual experience. To continue with the example from the previous paragraph of an individual person enjoying a specific walk: the cause of this emotion is both walking and the area of countryside, the happy person is the only individual who is affected by their capacity to walk in this case. The specific area of beauty is more of a global phenomenon, with local wildlife as well as other people affected. If it is a particularly solitary spot, the individual may consider the area itself to also be specific to their individual experience. Since each of these dimensions are relative, the belief of the person whose output is being coded is important.

The area of origin refers to the source of the emotion, this could be internal or external. In the case of our walker, the source of the happiness is both the exercise and the sights. The exercise causes endorphins and is therefore an internal source of happiness. The sights of the area of natural beauty are an external source of happiness because they come from without.

A distinctiveness attribution involves the consideration of whether this situation is personal: that it could only happen to the person experiencing it, or if it is more universal. In the case of the happiness found in a walk through nature, this is distinctively universal. There are many people who find joy in walking through areas of natural beauty.

Lastly the level of controllability, is the situation around the emotion controllable or not. The capacity for control in the situation I have exemplified throughout this situation is layered. The individual is not always able to control their ability to walk, the weather or any changes to accessibility of the route. An individual who finds happiness in natural walks would tend to have the control to go for one of these walks to help increase their happiness.

Stratton (2003) writes that part of attributional coding is considering each of the dimensions alongside one another, since looking at each in isolation does not provide the intersection points that can cause distress. Combining these dimensions gives a clear picture of who is speaking in the context of how they consider emotions or situations to have arisen. This is briefly discussed in chapter 8 as part of exploring temporal focus and used alongside appraisal theory in chapter 9.

5.3.2 Appraisal theory

Appraisal theory is a linguistic framework that involves considering the components that combine to discursively construct our emotions. The way this framework is set up to achieve this is by coding individual words, phrases or units of meaning. This coding is based on categories, which are outlined in Table 5.2. The framework originated with systemic functional linguists, but has since branched out to include varieties of discourse analysts (Bednarek, 2008, p.12). The theoretical framework provided by appraisal theory has evolved over time, as well as opening up the possibility of combining with other analytical techniques. In particular, Bednarek (2008) combined CLAWS PoS tagging of different sections of the British National Corpus (Davies, 2004) with appraisal theory. This was in order to compare how different registers such as newspaper articles and recorded speech construe emotion.

The full components that constitute appraisal theory are laid out in Table 5.2, followed by an outline of the meaning of each category:

Engagement		Mono-glossic	Hetero-glossic		
Attitude	Affect	-ve/ +ve feelings	personal capabilities	personal normality	
	Judgement	moral truth	moral ethics	+ve/-ve	Direct/implied
	Appreciation	Positive	Negative		
Graduation		force of emotion	focus of emotion		
Feelings	Happiness	Cheer	Misery	Antipathy	
	Security	Confidence	Trust	Disquiet	Surprise
	Satisfaction	Interest	Admiration	Ennui	Displeasure
	Affect	Fear	Desire		

Table 5.2: Appraisal categories

The components that constitute appraisal theory can be loosely grouped into: *engagement, attitude, graduation and feelings* (Martin & Rose, 2003, p.54). Briefly, these components are the source, impact and power of the emotion, followed by the fragments of the emotion respectively. Engagement refers to the source of the emotion, which may be based on the informants own opinion or with reference to outside sources. Attitude is the impact of the emotion, divided into the smaller branches of affect, judgement and appreciation. Each of these sub-groupings of attitude can be positive or negative.

I will be using two of the components of appraisal theory as part of my analysis of chapter 9: from affect, the positivity or negativity of feelings, and from graduation, the focus of emotion. It is established practice within discourse analysis to use the parts of appraisal that are relevant to the research (e.g. Grundlingh, 2018). In order to provide a methodological background, I summarise each of the sub-groupings of appraisal theory but concentrate on valence and focus.

The sub-group affect encompasses how emotions as a whole can be expressed through discourse. This expression can be either direct or implied, as well as with positive or negative valence (Martin & Rose, 2003, p.25).

The sub-group of judgement takes into account morality, ethics and personal capacity and conventionality. Graduation takes into account both the intensity of emotion and how directed it is (Martin & Rose, 2003, p.54). Appreciation is an evaluative measure, referring to *reactions*, *compositions* and *values* (Martin & White, 2007, p.69) and each of these can be positive or negative.

Reactive appreciations are how interesting or pleasing something is; compositive appreciations are how something maintains its equilibrium alongside complexity; value appreciations refer to how appropriately timed they are, or inventive, or authentic (Martin & White, 2007, p.69).

There is also a specific feelings based category, which provides a method of classifying the components that make up the linguistic expression of a feeling. These components include affection, misery, trust, disquiet and desire (Martin & Rose, 2003, p.58).

The two parts of appraisal theory that I concentrate on in chapter 9 are affect: positive and negative feelings; and graduation: focus of emotion. This is because the analysis using corpus linguistic techniques in chapters 6 to 8 had indicated that valence and specificity in language patterns could be related. I therefore used appraisal theory to explore this connection and found links that I detail further in chapter 9.

Martin and White (2007, p.46) explicitly state that feelings should be classified as positive or negative based on popular culture, or “vibes” rather than psychological constructs. When classifying emotions as positive or negative, it is important to examine the context of a word to ensure accuracy (Martin & White, 2007, p.47). When classifying the focus of emotion, the prototypicality of the emotion conveyed is the true measure under consideration (Martin & White, 2007, p.137). Terms that increase prototypicality such as *real* or those that detract such as *kind of* are instrumental in this classification.

Appraisal theory has been used in a range of discourse analysis based research where emotion and its linguistic construal is particularly relevant. This includes deeper understanding of language acquisition or translation, such as comparisons of translations of the novel *Mansfield Park* (Alsina et al., 2017). Appraisal theory has also been used to consider different aspects of works of literature such as *Oliver Twist* (Lian, 2018). In keeping with discourse analysis of health related communication, appraisal theory has been used in an analysis of news articles about Covid (Ye, 2021). In their analysis of news discourses about Covid, Ye (2021, p.160) compared articles from Wuhan and London to understand the levels of perceived objectivity. They concentrated on affect, and found that CNN reported approximately similarly across appreciation and affect, though used

slightly more judgemental terminology when referring to Wuhan when compared to London (Ye, 2021, p.168).

There is precedent for the combination of appraisal theory and health communication, in particular the language of mental health. Grundlingh (2018) found this framework helpful for screening the difference between a genuine and a falsified suicide note. Since appraisal theory enables analysis of the discursive construction of emotion, it was possible to see the difference between the emotions conveyed. Specifically, Grundlingh (2018) found that the categories *trust*, *moral ethics* and *negative appreciation* were more prevalent in the genuine suicide notes. I build on this work by comparing the use of appraisal theory categories for people experiencing low mood at the time of interview (TLM) when compared with my reference group of people who were not (RLM).

5.4 Analytical methodology summary

In this section, I have covered the practicalities of corpus linguistics in general and then the three main applications of this method of analysis in my work. Specifically, I covered how I performed searches for key words and key part of speech tags in my own data. Further to this, I described how I used corpus linguistics to search for prototypical texts to enable manual coding. Lastly, I discussed the theoretical frameworks of attribution and appraisal theory and how they can be applied to my data.

Chapter 6 Word count, content and diversity analysis

6.1 Introduction

Through this chapter, I address my first research question. This is: *What linguistic markers are prevalent in the verbal language of people experiencing low mood?* To do so, I explore word counts, topics, lexical variety and keywords of the language of people experiencing low mood when compared to my reference group. For shorthand, I will refer to the corpus of the language of low mood informants from my main study as TLM (target low mood). The corpus of language from the informants who were not experiencing low mood I will refer to as RLM (reference low mood).

6.2 Word counts

My first research question involves identifying linguistic markers that are prevalent in the language of people experiencing low mood. These linguistic markers include the number of words used, for which I examine overall word counts. Existing research indicates that people diagnosed with depression are likely to use more words when responding to a prompt about the self, when compared to a reference group (e.g. Smirnova et al., 2018). Table 6.1 follows with an overall summary of the transcribed word counts across my pilot and main studies.

	Word counts					No. of Informants
	Total	Mean	Median	Lowest	Highest	
TLM	10,442	870	752	74	2841	12
RLM	41,659	1041	877	87	4835	40
Main overall	52,101	1002	800	74	4835	52
TPD	3,444	1057	792	545	2098	3
RPD	5,015	627	621	335	962	8
Pilot overall	8,459	769	678	335	2098	11

Table 6.1: Word count summary for pilot and main study

My analysis that follows shows that my main study indicates no clear correlation between mood level and word count, although the pilot study corroborates the existing research trends.

I grouped the data for the main and pilot studies separately and then compiled each into a visual representation. In figures 6.1 and 6.2 I have grouped the word counts into 100 word increments, then normalised the informant numbers by examining the percentage of informants who use that number of words. As an example, of the 11 informants, if one used 236 words and one used 297 words, both would be using between 200-300 words. If they were the only informants to use 200-300 words, they would make up 2/11 of the total pool, which is 18% so 18% would be marked in the 200-300 column. The figures provide a visual representation of how the informants' word counts are clustered including where any outliers are. The informants with depression (TPD) and those with low mood (TLM) are shown in dark grey in figures 6.1 and 6.2, respectively. This is to contrast with light grey for the RPD and RLM groups.

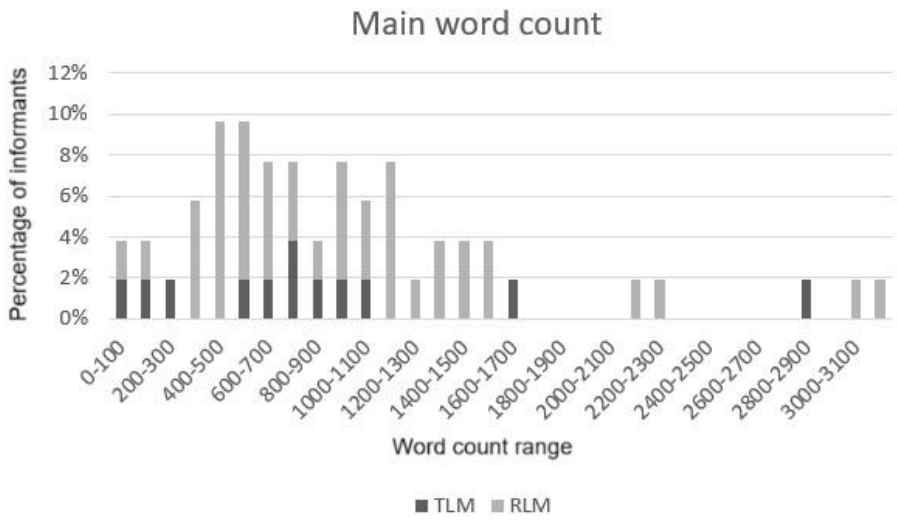

Figure 6.1 Main study grouped word count by percentage of informants

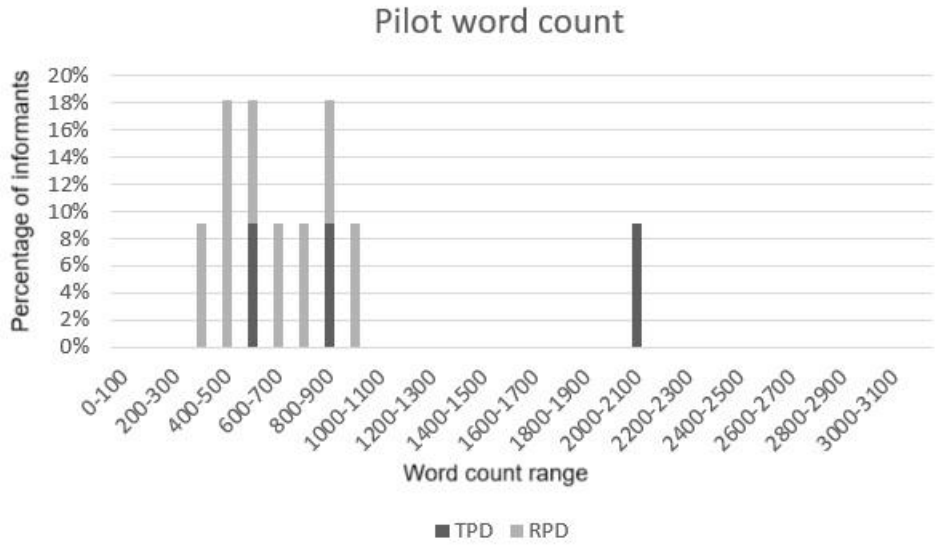

Figure 6.2 Pilot study grouped word count by percentage of informants

The distribution of TLM and RLM informants is mostly centred around the 400-1200 words. The word counts of the TLM and RLM informants does not appear to have a clear pattern based on this representation. The clustered data (400-1200 words) alongside gaps (1800-2100 and 2400-2700 words) does provide visual representation of the reason for the difference in median and average word counts discussed in section 4.2. The skew caused by these loquacious informants cannot be taken to be a feature associated with low mood because it is based on the speech of one informant.

Within the pilot word count, most of the TPD informants are in the same area as the RPD informants. Specifically, 200-1000 is the most popular word count. One low mood informant has a much higher word count than the rest of the informants. Once again, this helps explain the difference between mean and median as described in data generation summary section 4.2.

Examining these figures next to one another, there are outliers identifiable in each study. The loquacious depressed informant from the pilot study set the expectation that my English language data may have similar trends to those identified in the Russian language work of Smirnova et al. (2018). However, within the main study there are two RLM informants who are the most loquacious, although one of the TLM informants is very close with the number of words they used as part of their interview. Most of the clustering for the main study (both TLM and RLM) appears around 300-1200 words, which is similar to the pilot study (for both TPD and RPD). There is no clear difference in dispersion across the TLM and RLM groups. This initial exploration indicates that number of words used is not a linguistic marker that can be used to differentiate the language of people experiencing low mood.

There are three potential reasons for the difference in results for the main study low mood data compared to the pilot depression data. The diagnostic parameter for Smirnova et al. (2018) as well as my pilot study was depression rather than low mood. Since there is much more to depression than purely low mood (please see chapter 2.4 for diagnostic details), it may be that the word count is affected by more of the symptoms of depression than low mood. Congruent with this, depression is associated with under confidence and a reduced ability to think or concentrate (WHO, 2022). However, it may be that the existing personal relationship between myself as interviewer and the

informants of my main study encouraged involvement for those who were feeling most low, when the lowest may not have participated in a study with a stranger. To clarify, the lowest mood informants would potentially talk the least, and self-exclude from the type of studies that involve researchers who have no previous relationship with this informant. The culture and community around the researcher-interviewee relationship is known to impact involvement and verbal contributions (Copland et al, 2015, p.33). Finally, Smirnova et al. (2018) was a Russian-language study, and it may be that some linguistic patterns associated with low mood arise from language or cultural differences. These are potential reasons for the lack of correlation between low mood and word counts based on all of TLM and all of RLM. To interrogate the word counts of my main study data further, I chose to perform data visualisation based hypothesis testing. This enables an exploration of the link between word count and some of the key demographic factors of my dataset.

6.2.1 Hypothesis testing for word counts

I took a hypothesis-based approach in order to interrogate the number of words in TLM when compared to RLM. These hypotheses were based upon the literature review and anecdotal insight then refined based upon initial impressions from transcribing the data. I used graphs to visualise these hypotheses as part of this exploration.

My initial hypothesis was based on loquaciousness, since this was something that stood out in the literature review. Interrogating correlations between the mood score (MFQ, appendix D) and word count provides a nuanced approach when compared to the binary TLM to RLM word counts in figure 6.1. Both my pilot study and the depression-based findings of Smirnova et al. (2018) found that informants with depression are more verbose than informants who have not been diagnosed with depression. However, when researching socialisation, psychologists Brown et al. (2011) found that low mood experienced as part of depression is associated with a withdrawal from company and feeling a lack of connection to fellow humans. Contrary to the pilot study, Brown et al. (2011) indicate in their work that there is a reduced likelihood of a low mood informant being talkative. This is in contrast with one informant in RLM corpus who spoke at length about all the things they liked under the guise of importance. Low mood can engender a ‘survival mode’ at which point the informant will

only talk about what is deemed necessary to complete their task (Pintea & Gatea, 2021, p.333). Within the ICD, under-confidence and reduced energy are both considered symptomatic of depression, and both of which could contribute to this potential reduced word count (WHO 1993). A combination of these factors led me to amend the initial hypothesis to consider low mood to be indicative of a lower word count. In figure 6.3, the higher the mood score (as determined by MFQ, see Appendix D), the lower the mood, with 27 as the cut off point (represented as a bold vertical line). Figure 6.3 shows no correlation between TLM and RLM informants.

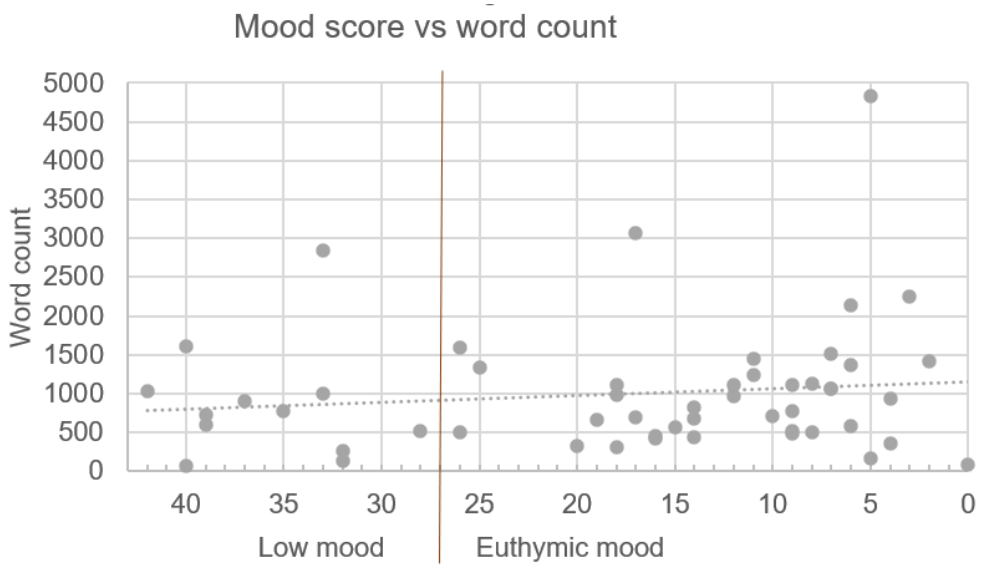

Figure 6.3 Main study mood score vs word count

Within this graph, the exact word counts are plotted on the y axis whilst the mood scores are plotted along the x axis. I chose to plot the scores along a scale rather than a strict binary of TLM or RLM for this part of the analysis in case level of mood was a factor in word count. There is a modest trend towards higher word count for a higher mood score. However, the informant with the lowest mood score has a word count similar to the informant with the fourth highest mood score, and the informant with the highest mood score has a similar word count to one of the informants with the second lowest mood scores. There are no significant links between mood score and word count based on this visualisation, giving us a null hypothesis. In terms of my first research question, this means that word count is not a linguistic marker of low mood in this data set.

My next hypothesis was that men with low mood talk less than women with low mood. Patriarchal western gendered expectations of women is that they are emotional, whilst men are expected to be

strong and only show emotions like anger rather than sadness (Lakoff & Bucholtz, 2004 p.22). In a society whereby women are expected to talk feelings (Charteris-Black & Seale, 2010 p.1), they seemed more likely to have the vocabulary to articulate as well as to feel supported and listened to. Within this context, men would more likely to hide their 'inappropriate' feelings by talking less. However, as can be seen from figure 6.4, this is not the case:

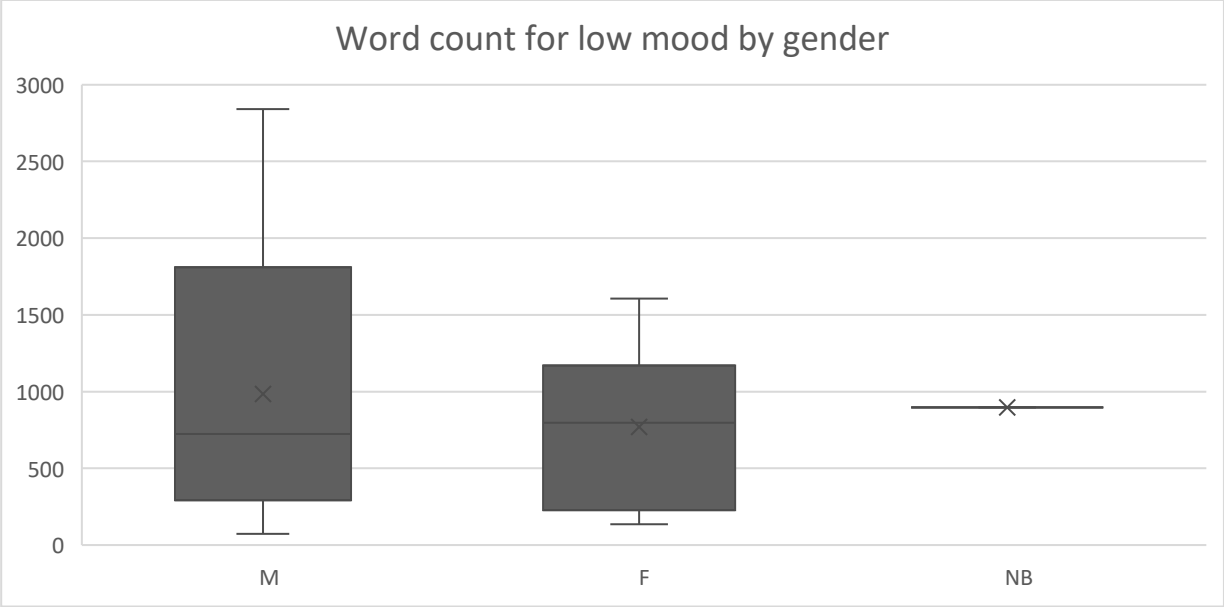

Figure 6.4: Word count TLM by gender- male, female & non-binary

The average word count for the TLM informants is similar for all genders, though females have the lowest, then non-binary, and males have the highest average word count. Median word count is around the same as the average for the female and non-binary low mood informants, whilst the median is lower than the average for the male low mood informant group. This means that more word count values fall within the lower end of the scale than the upper range for males with low mood, something that is confirmed by the placement of the upper and lower quartiles in this graph. There is less range within the word counts for the female low mood informants. There is only one non-binary low mood informant therefore it is not possible to know if the word count statistics are generalisable without further research.

Based on these results, none of the testing for my initial hypotheses was conclusive. The pictorial representations of word counts by mood and by gender showed no incontestable trends. I therefore turned to the Statistical Package for the Social Sciences (SPSS) (IBM corp., 2020) to perform

significance testing on the TLM and RLM data. Within SPSS, a chi-squared test can be used to establish if there is a likelihood that any difference in results is due to chance or an underlying cause. A chi-squared value of below 0.05, or 5% is normally the benchmark for significance (Dörnyei, 2007, p210). This test revealed no significant link between the results of the Mood Feelings Questionnaire (the metric used for mood measurement) and word count, as can be seen from Table 6.2:

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	49.183[a]	50	.506
Likelihood Ratio	53.409	50	.345
N of Valid Cases	52		

a. 102 cells (100.0%) have expected count less than 5. The minimum expected count is .23.

Table 6.2 Chi-squared test for significance of mood against word count

This table shows that the likelihood of a significant link between the mood score and word count is 0.345 which is higher than the 0.05 threshold. Only scores below this threshold are significant, and therefore I can eliminate the presence of a significant link between mood and loquaciousness from this main study data.

I also used SPSS to test the link between low mood and age, gender, sexuality, education or community (e.g. larp or work). Once again, none of the chi squared results were below 0.05, meaning that word count for low mood informants was not significantly affected by demographic within my main study data. The age, gender, sexuality, education level and community membership has no significant impact on interview length. This may be due to sample sizes being relatively small in comparison to larger corpora such as the BNC (BNC Consortium, 2007). In terms of my first research question, this means that number of words is not a linguistic marker associated with low mood, either collectively or based on demographic.

6.3 Topics

This is a comparative study of responses to the main question that I asked my informants, which was: *When you think about what is most important to you, what comes to mind*? I compare the topics spoken about by TLM informants when compared to RLM informants (see section 4.4 for an overview of the topics covered in the corpora). The average number of topics spoken about by TLM was 7.2, compared to the 8.6 of the RLM informants.

To enable a view of the types of topics spoken about by informants, I have placed them into groups. These groups were loosely based on Maslow's Hierarchy of Needs (Maslow, 1943) and Bourdieu's social capital (Bourdieu, 2021). The groups are: self; connections; society; nature; spiritual; physical; hobbies. *Self* is anything related to the psyche or the ego of the informant, including their personal happiness, the importance of exercise, or the desire for self-actualisation. *Connections* are around existing relationships or the hope of future interpersonal relationships. *Society* is on a grander scale than connections, including a desire to help people, politics and social justice. *Nature* refers to the environment, including causes campaigning for climate change awareness and protecting the natural world. *Spiritual* is a category for the informants who cited their faith and their concept of God as important. *Physical* refers to places and objects such as noise cancelling headphones for noise-sensitive informants, through to a car, and a house. Finally, *hobbies* includes reading, LARP, board games and storytelling.

The grouped topics that the TLM informants found more important when compared with the RLM informants are in figure 6.5. The percentage of topics is based on the number of informants who referenced that topic rather than the number of mentions per informant or how long they spoke about the topic. To clarify, if an informant were to talk about *nature* five times in their interview, and they spoke about *nature* for 20 minutes, that informant would be counted as one for the *nature* topic.

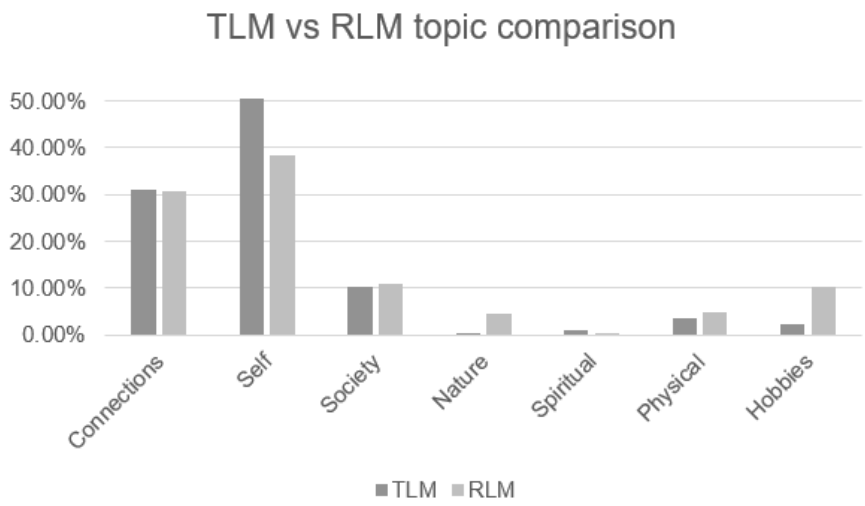

Figure 6.5: TLM vs RLM topics of importance

The TLM (low mood) and RLM (reference) informants spoke about things related to themselves the most, when asked about what is important to them. The importance of connections was proportionally similar in RLM when compared to TLM. The TLM informants covered each of the topics, although only one TLM informant spoke about nature and one spoke about their faith. The verbal data of the RLM informants makes it clear that nature and hobbies were more important to them than physical things.

The increased focus on the self for TLM (51%) compared to RLM (38%) is in keeping with psychological understandings of contributory factors to depression. Specifically, that self-focus is a contributor in “perpetuating depression” (Watkins, 2004, p.1037). Whilst both the TLM and RLM informants talk about themselves, there is a greater proportional focus on this topic by the TLM group. There is a greater interest in connection, society, hobbies and nature from RLM when compared to TLM. Each of the topic groupings that are more prevalent in RLM over TLM are those that enable informants to look outside of themselves. To provide two examples from the RLM data:

Quote 6.1, Nature: “being mindful of the planet. is important to me”

Quote 6.2, Society: “justice. and about making things better for people”

RLM quotes 6.1 and 6.2 each show a different way of looking outside of the self. The physical is of greater importance to TLM than RLM, and physical things are not the same as connecting with loved ones, society, hobbies and nature. Quotes 6.3 and 6.4 from TLM exemplify their interest in the physical:

> Quote 6.3, Nature: “nature. uhm and how especially as I get older I need to be outdoors. more and more… uhm I need to be alone outdoors”
>
> Quote 6.4, Society : “other important things are being able to do- put- uhm put events on that I can put money into. and food into a food bank”

The contrast between how the TLM and RLM informants talk about nature in 6.3 and 6.1 respectively is typical of the difference between TLM and RLM informant verbal data. The RLM informant talks about the planet as a cause, whilst the TLM informant talks about nature generally as it directly affects them and their life. Even in terms of society, the RLM informant in quote 6.2 talks about the cause of social justice whilst the TLM informant in quote 6.4 aims for the specific act of running small scale charity events. As a new finding, based on this data, this difference in how generalities and specifics are discussed may be indicative of a low mood psycholect. This is something that will require further analysis in subsequent chapters to explore in more detail.

6.4 Lexical diversity

The first research question is: *What linguistic markers are prevalent in the verbal language of people experiencing low mood?* As part of this exploratory analysis, I have found indications that lexical diversity may be a feature of the low mood data. In this context, I am using *lexical diversity* to describe the “range of different words used in a text” (McCarthy & Jarvis, 2010, p.381). There are a range of near synonyms for the term lexical diversity, including lexical richness, lexical variety, and lexical density (Kyle, 2019, p.454). I discuss the nuances that differentiate the terms *lexical diversity* and *lexical density* in section 2.5.

Mental health researchers Park and Conway (2018) examined lexical diversity in online forum users. They found that members of mental health related groups used significantly less lexical diversity when compared to members of control groups. Rezaii et al. (2019) performed a study examining

the relevance of lexical richness to psychotic episodes. They found that a low number of content words compared to a higher number of function words was an indicator of a psychotic episode. Content words are lexical words that convey meaning, such as nouns, verbs or adjectives (Baker et al., 2006, p.47). Function words are words that grammatically construct a sentence, such as conjunctions or auxiliary verbs (Baker et al., 2006, p.76). I could find no studies into the effect of specifically low mood or depression on lexical density or diversity, only grouped data. The main study data is suitable for an examination of lexical diversity because the two corpora, TLM and RLM were elicited in the same way, and are therefore comparable. The data is also specifically low mood rather than depression grouped in with other mental health concerns and therefore can enable a more focused picture of the effect of low mood on lexical diversity.

There are a range of ways to examine lexical diversity or density, some of which are simple calculations and some of which require complex computer algorithms, such as the study cited above (Rezaii et al., 2019). I analysed a sample of my data using the same type of lexical density analysis as that employed by Rezaii et al. (2019). I used the sampled version of my corpora that I created in order to find prototypical texts, as discussed in chapter 4.2.3. I chose this sampled version to ensure that individual relative loquaciousness would not affect lexical density measures. Both the sampled TLM and sampled RLM corpora had around 60% content words to 40% function words, with only a 1% difference between the two. Because I could not be sure that the lack of difference was due to the sampling, I chose to concentrate instead on the lexical diversity methods that included the entire corpus.

I employ two methods in this section, one simple calculation and one computer assisted analysis. The first was to discover that this form of analysis is relevant to my material; the second was to examine this relevance in further detail. The two types of analysis are type to token ratio and automated lexical diversity analysis.

6.4.1 Type to token ratio

The type to token ratio is a simple calculation, sometimes used to provide an indication of the potential lexical diversity of data. This is merely an initial indication, because the difference in corpus

size between TLM and RLM affects the type to token ratio. A *type* is a specific word form in a corpus, whilst a *token* refers to all word forms that appear in the corpus, including recurring types. To provide an example from a TLM informant:

> Quote 6.5: in my personal life in my life. in my own family unit and my friends within my community

In quote 6.5, the specific word form *in* is used 3 times, *my* is used 5 times, *personal* is used 1 time, *life* is used 2 times, *own* is used 1 time, *family* is used 1 time, *unit* is used 1 time, *and* is used 1 time, *friends* is used 1 time, *within* is used 1 time and *community* is used 1 time. This makes *in* + *my* + *personal* + *life* + *own* +, *family* + *unit* + *and* + *friends* + *within* + *community* = 10 types of words. There are 18 total words (tokens) in the excerpt, because some of the word types are repeated; e.g. *my* occurs 5 times.

A type to token ratio involves dividing the total number of distinct words (types) in each corpus with the total number of words (tokens) in the same corpus. The type to token ratio for the excerpt earlier in this section is 10 types divided by 18 tokens, providing a figure of 56%.

The basic measure of type-token ratio is affected by the size of a corpus, where a longer corpus will normally involve a lower type-token ratio than a shorter one (Kyle, 2019, p.457). This is because a longer corpus has more potential for repetition of words than a shorter one. Type-token ratio can provide an indicator of whether lexical diversity may be a meaningful measure for my analysis. The TLM (low mood) type to token ratio is 1275:10919, which is 12% whilst the RLM (reference) type to token ratio is 3192:43712 which is 7%. A higher figure can be indicative of more lexical diversity, or simply a smaller corpus. I chose to investigate further using a calculation based on type token ratio, but with controls in place to allow for different corpus sizes.

6.4.2 Automated Lexical Diversity analysis

McCarthy and Jarvis (2010) developed a method of calculating lexical diversity, known as: Measure of textual lexical diversity (MTLD). Due to the sampling based method of calculating lexical diversity, this is not affected by corpus length. The MTLD calculation is “the mean length of sequential word strings in a text that maintain a given TTR value (McCarthy & Jarvis, 2010, p.384) where TTR stands for type to token ratio. Each word in a text is individually analysed based on the word immediately

before and the word immediately afterwards. This analysis is performed both forwards and backwards, and is done for every word in the corpus. This initial calculation produces a “factor count” for the corpus (McCarthy & Jarvis, 2010, p.385). The total number of words in the corpus is then divided by the factor counts for the forward analysis and then from the backward analysis. The MLTD figure is therefore the mean of the values from both analyses. Due to the small word strings analysed as well as the inclusion of all data, this calculation is applicable for any data set over 100 words (McCarthy & Jarvis, 2010, p.384).

I calculated the MTLD figures using a website by Reuneker (2017), analysing the full TLM and RLM corpora separately. The MTLD figure for TLM is 37.19 and for RLM is 47.28. This indicates that the RLM informants use more lexically diverse verbal utterances than the TLM informants. In real terms, it means that the TLM informants use a less varied vocabulary than the RLM informants.

In this section on lexical diversity, both measures have indicated that the TLM informants have a less varied lexicon in their verbal data. The type-token ratio can be affected by the corpus length, though MTLD is not. This section builds on the work of Park and Conway (2018) and their exploration of written internet forum entries with users with a range of mental health issues. They found that users of mental health fora used less lexically diverse language than users on non-mental health fora. In corroboration, this section, particularly with the use of MTLD, I have confirmed that my TLM informants use less lexically diverse language than the RLM informants.

6.5 Keywords direct comparison

In this section I build more upon my previous analysis of the volume, content and complexity of grouped words used by looking more closely at locution. Keywords are terms that can indicate potential linguistic topics, stylistic features or patterns based on a comparison of one corpus against the other (Groom, 2022). When comparing corpora, it is possible to search for both positive and negative keywords. These are words characteristically more and less likely to appear in the TLM (low mood) corpus than the RLM (non low mood) corpus respectively. The modest size of this corpus means that an individual who talks a lot and repeats themselves can affect parts of corpus linguistic

analysis including keywords. During this keyword search, there were some keywords that were in the language of only one informant. I have deliberately excluded these from this table since they were part of an idiolect, and not representative of the cohort. The remaining keywords therefore provide an insight into the type of language used by the low mood informants when compared to the reference. The positive and negative keywords based on a comparison of TLM and RLM are ordered by keyness, based on the chi squared calculation of likelihood. The frequency figures for TLM and RLM are normalised per thousand words and I have also included effect size to show how meaningful the keywords are (Groom, 2022). Whilst likelihood shows the probability of a word being key, effect size takes into account the size of the corpora being compared. This data follows in Table 6.3:

	Type	TLM norm. Frequency	RLM norm. frequency	Keyness (Likelihood)	Keyness (Effect)
Positive	would	4.12	1.62	29.118	0.009
	easy	1.10	0.11	27.227	0.002
	to	32.51	26.83	26.826	0.063
	change	1.01	0.14	24.534	0.002
	relationships	0.92	0.14	21.265	0.002
	interactions	0.46	0.00	20.018	0.001
	useful	0.82	0.09	19.717	0.002
	come	1.65	0.43	19.019	0.003
	already	0.46	0.02	18.912	0.001
	vital	0.55	0.05	18.799	0.001
	me	14.65	9.88	18.549	0.028
	these	2.01	0.64	18.045	0.004
	be	9.52	5.95	16.887	0.018
-ve	yeah	2.66	6.73	20.141	0.006
	just	2.84	6.43	17.738	0.006

Table 6.3: TLM vs RLM keywords

In this section, I particularly concentrate on positive keywords *would*, *to*, *change*, *me* and *these*. That is because these keywords are the most relevant to language patterns associated with low mood, though I will briefly cover the main themes of the rest of the keywords. These are a sense of negativity, absolute thinking and abstraction of language. In terms of absolute thinking, *Easy* is used adjectivally throughout TLM to indicate polite indignation at the difficulties of life. The clearest example of this is: *I don't think. the world uh should it should be easy but I don't think the world should always be uh quite as hard as it is*. To provide two further illustrative examples: *There are*

days when it is very not easy, and *if it's important it's not always easy*. The keyword *easy* when used in contexts such as this is consistent with an increased amount of linguistic negativity found in previous linguistic studies of depression (Newell et al., 2018, p.451). The keyword *vital* is used as an emphasiser for how important a concept is. Framing something as vital indicates extreme or absolutist thinking. To clarify, absolutist thinking is looking at the polarised side of a semantic category (Al-Mosaiwi & Johnstone, 2018). An absolute thinker within this context would describe something as vital or a necessity rather than purely important or engaging.

Many of the TLM keywords focus on abstract concepts, in the form of nominalisations (*relationships*, *interactions*) and attributes (*useful*). This is in contrast with how the RLM group expresses similar concepts, where they focus on specific and concrete examples. For instance, the keyword *relationships* arises from the TLM informants listing this conceptual noun as important, whereas RLM informants prefer to name individuals such as a spouse, children, friends and so on. In a similar way, the keyword *interactions* reflects a focus on conceptual behaviour important to the TLM informants, rather than conversations with specific individuals in RLM. An example of this from RLM is: *I know talking to my my siblings*: the informant references the importance of this specific type of interaction rather than interactions broadly. The keyword *useful* is something that the TLM informants talk about as something they aspire to be, rather than someone in particular that they want to be useful to. As an example: *when I can be useful to people, that gives me great joy*. The final keyword I examine in less detail is *be*, which is covered as part of the PoS tag analysis in section 7.4.

The keyword which is most key is the modal verb *would*, defined in the Oxford English dictionary as "feeling or expression of conditional or hypothetical desire or intention; a (mere) wish or intention" (*OED*, 2023). Carter and McCarthy (2010, p.650) advise that *would* is commonly associated with the past tense but is also used conditionally or "to soften statements and requests". The association with the past tense is of particular interest for my second research question: *Do people with low mood focus on the past?*, so I have explored the use of this keyword across the TLM and RLM corpora.

Examining some examples in context in addition to an examination of collocations offers insights into how the language patterns for low mood informants may be different to those for non-low mood informants. I will include a selection of 5 examples from each corpus, beginning with quotes 6.6 to 6.10 from the main study low mood corpus (TLM) followed by quotes 6.11 to 6.15 from the main study comparison (RLM).

Quote 6.6: so that would be a significant way to enhance. your feeling
Quote 6.7: people that were old when I was young… would be who we would all become
Quote 6.8: important thing that and… I would say as well and this falls under
Quote 6.9: at the moment… I think those would be. on top… because they
Quote 6.10: those. those things would. would breathe more life into the happiness

Within TLM, the collocates (or words most likely to be found in the lexical company) of *would* are: *be*, *say*, *would* and *those*, whilst in RLM they are *be* and *say*. I have therefore included an example of each of the TLM collocates in quotes 6.6 to 6.10. *Would* is used in the data from this corpus in both conditional and hypothetical senses of the word. Most of the collocates indicate these dual uses, and *would* as a collocate is used anaphorically as well as reflecting repetition in spontaneous speech.

Quote 6.11: nurture me or my friends and family would be the other category
Quote 6.12: worth pursuing uhm that in its absence it would be felt.
Quote 6.13: … well kind of I would say a level of priority
Quote 6.14: fun for which I would include live role play
Quote 6.15: then I would probably… I would probably say if I if I had to

The collocates for *would* in RLM are *say* and *be*, I have therefore included an example of each of these within quotes 6.11 to 6.15. As with the low mood corpus, conditional forms of the word *would* are frequently used. Two of the collocates are shared across the corpora, indicating similarity in some of the linguistic patterns. However, unlike the low mood narratives, *would* and *those* are not collocates of *would* in the speech patterns of those who were not experiencing low mood. The lack of *would* collocating with *would* suggests a reduced level of rhetorical emphasis and/or a more purposeful manner of communicating. The lack of *those* collocating with *would* indicates that the low mood informants in TLM use grouping terms alongside this expression of desire or intent whilst the

informants in RLM use specifics. This echoes the observations made of the content based keywords, discussed above, regarding abstract concepts versus concrete (lived) examples.

The keyword *to* for the TLM corpus is used in the context of alternative possibilities each time it is used. The node *to* is exclusively an immediate collocate of *me*, *you* or *be* within the target corpus, though *you* is used generically and therefore is a self-reference term similar to *me*. These collocates indicate self-orientation whereby the main parts of the clause are linked to the speaker and their experiences and views, as in quotes 6.16 to 6.19:

Quote 6.16: but my rose bush is important to me because it was my marma's

Quote 6.17: that's very important to me and safety to me might mean different things

Quote 6.18: I think are pretty and look nice and I want to be able to look around and see

Quote 6.19: If ever you find something is not important to you uh k- kind of no matter

In quote 6.16, the informant discussed how a plant is not just important, but constructs that importance specifically in relation to themself, showing an orientation around the self in addition to a lack of confidence. In quote 6.17, the informant relates a conceptual importance to their personal understanding of the idea, showing the desire to clarify without a sense of directness. Quote 6.18 is an epistemic desire for more beauty, personal and external, whilst showing an under confidence. Quote 6.19 is the last example and also involves an orientation for importance, this time around the generic *you*. This self-orientation is represented within the linguistic studies covered within the literature review, such as Şimşek (2013) who asserts that self-orientation and depression frequently coexist and contribute to one another.

The next most significant keyword to undergo scrutiny is *change*, defined as either a noun or a verb, referring to something that could be swapped or transition to something different respectively (*OED*, 2024). The word *change* occurs 12 times in TLM, and 6 in RLM, some representative examples of change from TLM are shown in quotes 6.20 to 6.23:

Quote 6.20: God but as I perceive God to be. which is… a mixture of change and human connection

Quote 6.21: how to leave the world a better place. climate change etcetera you know

Quote 6.22: I guess... uhm…I can't change… other people I can only change myself

Quote 6.23: everyone has the chance and should have the chance to. to change you know the world

The keyword is used as a noun once in quote 6.21, within the context of climate change, and as a nominalisation in quote 6.20. Beyond this, *change* as a verb appears in contexts where the informants are unsure, talking about the transient nature of things, their interests, opinions, communications, other people and the world as a whole. There is a self-referential focus in only 2 of the 12 instances the word is used in total throughout the TLM corpus. The rest of the time the informants refer to their surroundings- be this personal (e.g. *other people*) or global (e.g. *the world*). Based on the linguistic studies covered in the literature review section 2.6, I expected low mood informants to be more self-referential in their speech than the RLM group. Whilst keywords such as *to* evidence references to the self, *change* does not. The use of this keyword is more concurrent with the psychological framework of attribution theory, particularly the attribution of external blame towards present circumstances (see 9.2 for further attributional analysis).

The keyword *me* is a pronoun that has been identified as a salient feature in previous research into depression, as covered in section 2.5. Previous studies have found that self-referential language, particularly the use of terms such as *me*, are linked to depression (Tackman et al., 2019). Its presence as a keyword in the TLM corpus therefore aligns with this finding, as does its collocation with *to*, discussed above, to foreground the speakers' orientation of their ideas to their personal experience and perspectives.

These is a keyword that further foregrounds the greater use of abstract language in the TLM corpus, compared with RLM, as in quotes 6.24 to 6.26:

Quote 6.24: my secondary answer talks about kindness and grace these are. linked

Quote 6.25: we keep evolving I guess and it's really huge like these ideas I didn't really think about

Quote 6.26: Self honesty is important to me integrity these these concepts these ideas

Contrastingly, compare these three uses of the keyword *these* in RLM, in quotes 6.27 to 6.29

Quote 6.27: It's incredibly important to have these people around me and in my life

Quote 6.28: if you saw it posted something about he was on one of these exercises

Quote 6.29: you you hear these stories from so many different cultures where. you know. Uhm. is absolutely love the idea of the Mexican idea of the day of the dead

TLM informants are more likely to talk conceptually (including, literally, about ‘concepts’ as in quote 6.26) , whereas the RLM informants use the demonstrative to point to specific and concrete entities and things. This fits with the research of Greer et al. (2004, p.524), who found that patients with mental health issues were more likely to talk about their symptoms generally rather than specifically. This pattern of specifics over generalities is a common theme, from the topics section through to keywords such as *these*.

The negative keywords are those characteristically less likely to occur in the language of informants experiencing low mood when compared to the reference group. This short list of items: *yeah* and *just*, provides further support for the trends observed in the positive keywords data that may be linguistic markers associated with low mood. Negative keywords complement the positive keyword analysis to show the differences between corpora, though it has been suggested that this is rather an essentialist view (Motschenbacher, 2018, p.158). The context of *yeah* does not provide a lot of insight into the mindset of the low mood informant, beyond the idea of using more negative words when experiencing low mood or depression as reported by Rosenbach and Renneberg (2014). Examining the term *just*, it is used as a modifier, gently reducing polarised responses for example the informant who says: *just something with enormous heft*. Al-Mosaiwi and Johnstone (2018) found that people experiencing depression use a greater proportion of absolute terms. The use of *just* as a hedge, a term that reduces absolutism in the reference group would therefore fit with the idea that informants not experiencing low mood are less likely to use extreme language.

6.6 Keywords BNC comparison

In this section I examine the keywords that arise based on a comparison of TLM with the sample of the spoken section of the BNC (BNC Consortium, 2007) and RLM with the sample of the spoken section of the BNC (BNC Consortium, 2007). Comparing my specialist corpora with a large representative general corpus of English provides a different perspective on the language, and addresses some of the restrictions encountered in the keyword analysis of the TLM/RLM comparison resulting from the small corpus size.

By contrast, the large size of the comparison corpus means there are a larger number of keywords than in the direct comparison of TLM and RLM. I have therefore introduced a cut-off at the 30th most significant keyword for each, in order to enable a manageable amount of data to examine, following established practice in corpus linguistics (Groom, 2022). I used *Wmatrix* (Rayson, 2009) for this analysis, which includes a sample of the spoken component of the BNC.This list of the 30 most significant keywords based on a TLM compared with the BNC sample against RLM compared with the BNC sample follows in Table 6.4:

Keyword TLM vs BNC	Log likelihood	Keyword RLM vs BNC	Log likelihood
important	974.46	**important**	2144.44
uh	585.79	**uh**	2071.85
[pause]	479.73	**things**	1024.17
i**mportance**	263.02	**[pause]**	889.15
things	254.84	**my**	561.47
me	223.87	f**amily**	530.32
my	168.05	**life**	416.86
life	138.11	**importance**	408.78
to	137.28	**i_think**	357.83
i_think	130.64	**me**	345.91
that	117.94	**friends**	328.00
family	104.69	kind_of	323.93
people	103.16	so	316.21
hmm	78.18	**because**	271.58
world	67.33	**that**	263.78
and	64.06	sort_of	241.90
friends	57.56	**people**	230.76
most	56.12	**being**	188.97
being	54.51	happy	175.43
safe	53.94	**and**	168.21
yourself	52.26	**sense**	154.26
like	49.51	**world**	153.52
value	49.50	guess	144.44
sense	49.32	i_suppose	142.73
myself	48.95	**like**	136.12
something	47.70	feel	128.40
content	47.56	kind	128.22
because	44.68	love	122.73
is	43.80	**im**	122.57
i**m**	39.84	priority	120.73

Table 6.4 TLM vs spoken BNC and RLM vs spoken BNC keywords

The bold words signify a crossover between the keywords identified in the TLM comparison with the BNC sample and RLM comparison with the BNC sample. This crossover is most likely to be because

the sample of the spoken word component of the BNC is based on conversations in a range of settings, whilst my data was all elicited using the same questions and the same method. This is particularly evident with terms like *most*, *important* and *importance*, which represent my informants repeating the interview question: *When you think about what is most important to you, what comes to mind?* Keywords such as *uh* and *pause* are due to my transcription protocol (Appendix E) which is not the same as that used for the BNC spoken component (Love et al., 2017).

I compared the full keyword lists based on TLM vs spoken BNC sample and RLM vs spoken BNC sample. Eleven TLM keywords within the 99.99th percentile likelihood of significance were not also in the RLM list. These keywords are in Table 6.5 in order of significance:

Keyword TLM vs BNC	Log likelihood
vital	35.42
useful	34.61
grace*	31.99
core	30.71
easy	29.61
safety*	28.44
be	23.75
alone*	18.57
respect	18.57
weird*	17.10
again	15.43

Table 6.5: TLM vs BNC keywords not also in RLM vs BNC (*appear in the contributions of a single informant)

These keywords are the terms that are distinctive to the TLM data over the RLM data when considered against a general reference of spoken language. Of these: *grace*, *safety*, *alone*, and *weird* are the words of one informant only and therefore part of their idiolect. It is notable that the keywords *vital*, *useful*, *easy* and *be* were also identified in the RLM comparison, signalling their markedness in the language of the low mood informants (see section 6.5). Therefore a closer examination of the keywords *core*, *respect* and *again* have the capacity to provide more evidence of potential language patterns associated with low mood.

There are six instances of the keyword *core*, as follows in quotes 6.30 to 6.35:

Quote 6.30: what's important to me is that uh the very core concepts and the important things

Quote 6.31: value …uhm a sort of… a c-… it's a core element needed. for something

Quote 6.32: If I did not deal with them I would betray the core concepts of who I want to be

Quote 6.33: it's really these core concepts that are absolutely integral to what we want and who we are

Quote 6.34: we need them to enjoy and to live but. the core principles you can't take away

Quote 6.35: not to the detriment of all others and my core self in fact they are a compliment

This keyword is used to talk about the heart of an idea, and as referenced in section 4.6, the vocabulary used is distinctively elevated, literary in style. Five of these six examples are from the language of one informant, indicating that these keywords have been slightly skewed by their loquaciousness and repetition. The elevated style of both of these informants is particularly noticeable in comparison to studies performed on internet fora (e.g. Hunt & Brookes, 2020). This elevated lexicon of my informants when compared to those involved in other studies was discussed further in section 4.6.

The keyword *respect* is used five times in the TLM corpus, three of these times are used as part of a definition of what importance means to the informants. For example,: *importance is giving something uhm enough respect and status*. The other two times are used close to the word *we*, used as a generic pronoun, by the same informant, for example: *we have to respect that, you know*. This impersonal use of *we* is similar to the *you* as one of the collocates of *to*, as discussed in the previous section. In both cases, this is a way of using ostensibly personal pronouns to reflect a broader audience.

The final keyword specific to this BNC spoken sample comparison is *again*, which is used 28 times in the TLM (see Table 6.5). It has emphatic, temporal and interactive functions as can be shown in quote 6.36 to 6.39:

Quote 6.36: creative expression uhm… and again I think that that links back to. interconnection

Quote 6.37: it's gone from your life and will never be again.

Quote 6.38: it's not misplaced and then again it's important

Quote 6.39: sorry could I get the question again sorry

Quote 6.36 is an example of the most frequent use of the keyword *again*, whereby an informant acknowledges repetition across the themes or topic of their response to the question. This metalinguistic speech pattern is found in the data for three informants on multiple occasions, although it is not characteristic of the whole TLM dataset. Other instances see it being used to request assistance from the interviewer, acknowledging the repetition in their interaction. Finally, the keyword reflects repetition in the experiences being discussed in response to the question of importance

There is some crossover between the keywords based on the TLM comparison with BNC sample and the RLM comparison with BNC sample. An example of this is the keywords *friends* and *family*, which were the two most frequent topics for the TLM and RLM informants. This inter-corpus comparison has provided more evidence of a psycholect that features self-referential focus for the TLM informants when compared to the RLM informants. The comparison also supplied evidence of abstract rather than concrete language to describe topics of importance in the psycholect of the low mood informants.

6.7 Keywords and social background

In order to address my first research question on the language patterns associated with low mood, I need to understand factors which may affect those language patterns. Social background affects language patterns (Chambers, 1995) (see literature review 2.7 for further information). In this section, I use keyword analysis to explore the lexical make-up of the TLM corpus through a comparison of intra-informant demographic groupings. It is important to address this because of the difference in proportions of low mood to reference informants between the straight and queer sexuality groups. Within the straight group, 6 in 35 informants could be categorised as low mood. Within the queer group, 6 in 17 could be categorised as low mood. Within this data, queer informants were almost twice as likely to be experiencing low mood when taking this study than their straight counterparts. This would appear to fit in with the statistics discussed in the literature review chapter 2.7. Namely, that members of the LGBTQIA+ community have an increased likelihood of experiencing low mood (NHS Digital statistics 2021). It is essential to ensure that intersectional data

is actively included in any study into low mood. It is popular for health communication studies to use data from internet fora, which often contains little to no demographic data (Brookes & Collins 2023, p.44). This study is therefore an opportunity to see how that often absent demographic data affects language patterns. The smaller size of my pilot study corpus means there is insufficient data for this kind of analysis, and so the pilot corpus is not considered here.

These demographic groups are gender, sexuality and education level. For details on the number of informants in each demographic, see section 4.3. The corpus size for each demographic follows:

	Gender			Sexuality		Education level				
	Fem.	Male	Non-binary	Straight	Queer	GCSE	A-level	Dgre/part	Mstrs/part	PhD
TLM	4835	5146	938	5712	5207	74	0	10196	651	0
RLM	21945	20593	1174	31767	11945	2116	5224	21581	12575	1815

Table 6.6: Word counts according to demographics

In the education setting, I have combined the undergraduate degree educated informants with those who have the equivalent of part of a degree, as well as the masters level educated informants with those who have part of a masters. This was in order to provide sufficient data for analysis of these groups. Categories in which there was only one TLM informant, for example the non binary, GCSE and Masters demographics are not compared individually. Since each informant fits into other categories, all collected data is still used as part of this analysis.

The corpora I examine are modest in size when broken down into these demographic groups, and therefore the analysis in this section is exploratory only. Throughout this thesis I am looking for language patterns that indicate a low mood psycholect, rather than language patterns that indicate social grouping. I have therefore chosen to compare social group against social group, for example, to compare the language of female TLM informants with the language of female RLM informants. This is to ensure that the only variable, the only thing that is different, is low mood. If I was to compare Female TLM informants with female, male and non-binary RLM informants, this would bring more variables into consideration. I therefore compare the female section of TLM with the female section of RLM; the queer section of TLM with the queer section of RLM etc. In order to enable an intersectional analysis of the positive and negative keywords by demographic, I have

included all of them together in Table 6.7, ranked by keyness (chi-squared) score. Keywords that are also general keywords based on the full comparison of TLM and RLM are highlighted in grey. Since these have already been covered in section 6.5, I will not go into detail on their use in this section.

	Gender				Sexuality				Education level	
	Female		Male		Straight		Queer		Degree/pt dgre	
	Keyword	Key-ness	Keyword	Key-ness	Keyword	Key-ness	Keyword	Key-ness	Keyword	Key-ness
Positive	relationships	21.87	important	39.52	these	33.74	within	19.08	to	21.65
	alone	17.12	would	26.03	useful	26.34			change	21.02
	old	17.12	easy	22.93	change	24.58			me	18.14
			content	22.55	important	20.63			would	17.59
			useful	22.55	selfish	20.20				
			selfish	22.22	would	19.69				
			is	21.47	easy	16.15				
			are	20.32	know	15.77				
			these	18.32	you	15.75				
			not	17.45						
			importance	16.82						
			kindness	16.10						
			vital	16.10						
Negative	them	14.42	just	15.68	yeah	17.12	Just	15.24	he	26.28
	yeah	13.64	laughs	14.80	where	14.61			yeah	23.42
			do	14.00	laughs	14.57			been	22.39
									just	19.31

Table 6.7: Keywords by demographic; general keywords from full comparison of TLM/RLM (section 6.5) are highlighted in grey)

In this table, there are a greater number of keywords for the straight and male demographic groupings. Both of these corpora are the larger of the gender and sexuality groups respectively. As shown in Table 6.6, the word counts for the demographics across corpora were approximately comparable, though the type to token ratios were lower for the TLM male and the TLM straight corpora when compared to their TLM female and TLM queer counterparts. By this measure, the TLM informants who identified as male and/or straight expressed less lexical density than the TLM informants who identified as female and/or queer. The larger number of keywords for the male and straight groups may therefore be due to this reduced lexical variety. If the men and/or straight TLM informants repeat themselves more than the queer and/or female TLM informants, this repetition will result in a greater chance of each repeated word being classed as significant. There are six shared keywords across the straight and male keywords list, which is also due to the presence of

the same informants (male and straight) in these demographic groups. The difference in the repeated words that the low mood males and straight people exhibit is sufficient to also affect the keywords for the overall data. This is visible in the high crossover between the male or straight keywords and the general keywords. This means that my general low mood keywords as discussed in section 6.5 are in part more representative of male straight informants than female, non-binary and/or queer informants. This is one of the benefits of an intra-informant analysis of this kind: it provides more detail on the sources of the language data.

Alone and *old* as well as *relationships* are the keywords that are specific to the female TLM informants when compared with the female RLM informants. Examining the former keywords in context, *alone* is something that some of the female TLM informants aspire to, in terms of time to themselves. The keyword *old* is found in the speech of two low mood female informants, and is used in the context of reflective thoughts on generic identity. Specifically, around how societal expectations around age have changed, and in one case, how expectations will continue to change for future generations:

> Quote 6.42: when I was younger I thought… people that were old when I was young… would be who we would all become but I became. I don't know how but I suddenly became very aware that actually we won't because we'll be… products of that time and the. old people we will be. at at the time that you know like in our 70s or 80s… it will be will be completely different and… you know. uhm people born in the 1400s… as old people then were very different to the people born the next generation etcetera

Alone is not in the speech of the male or non-binary TLM informants, and *old* is used twice by the same male TLM informant who talks about “old trauma” but not at all by the non-binary TLM informant. This could be evidence of cultural expectations around gendered behaviour, that low mood women are more likely to seek time alone or to reflect on age than people of other genders. An examination of the uses of the negative keyword *them* indicates that the female TLM informants use *them* to talk about indistinct groupings of people. This is in contrast to the RLM female informants, whose increased use of the word *them* is congruent with caring responsibilities, or generally orientating themselves around others. As an example, see quote 6.43 from a female TLM informant and quote 6.44 from a female RLM informant:

> TLM Quote 6.43: being furious at someone because I feel like they took advantage of me is is is something I can… oh I can't change them. I can only change my response

RLM Quote 6.44: a sort of nice feeling of being able to like. give something to them in the way that people helped me early in my career

This lack of introspection and willingness to look for importance outside of the self could fit with the work of Tadesse et al. (2019) and their findings that depression correlates with increased use of self-focused language. It therefore follows that informants without low mood would have an increased use of externally focused language, as is the case for the RLM female identifying informants. It also fits with the general items of importance for the collected TLM informants when compared to the specific items of importance for the collected RLM informants.

For the male TLM informants, the words that are not also found in the general keywords list are *important, content*, *selfish*, *is*, *are*, *not* and *kindness*. *Important*, *are* and *importance* are used to repeat the interview question and refer back to it as part of speech. *Content*, and *kindness* are both listed as terms that the male TLM informants hope to embody and portray: the former is referring to a sense of contentment. The keyword *selfish* is used by male TLM informants as a characteristic that they do not wish to have, or to be perceived as having. As an example, a male TLM informant talks about the importance of *giving something meaning and worth without becoming selfish*. The keywords *is* and *not* are used as part of a list, or to show a lack of surety, on different occasions as in: *standing up and supporting for them is not easy*. The negative keyword *do* does not appear to have any connotations distinctive to the words of the low mood male informants.

When examining the keywords from the male identifying TLM informants when compared with the female TLM informants, men are more likely than women to aspire to, or try to avoid characteristics. The women were more likely than the men to be reflective or desire a specific situation, such as time alone. Both of these patterns are indicative of introspection, despite the difference in how this introspection manifests. A sense of self-orientation has been a theme through the data, and this section has shown that it can be expressed in different ways by men and women.

Positive keywords exclusive to the straight identifying low mood informants were *important*, *selfish*, *know* and *you*. The demographic crossover with the male informants means that *important* and *selfish* are used in the same way for the straight TLM informants as for the male TLM informants.

Most of the occurrences of *you* and *know* are found as lexical bundles, as the informants seek reassurance with the discourse marker *you know* (Buysse, 2017). This could be partially due to my decision to not use verbal cues to encourage speech as part of the interview in the data generation process. The few occasions that these words are not collocates of each other, *you* collocates with *have*, indicating a sense of obligation. Based on the work of Galasiński (2008) a sense of obligation is something that his straight male participants spoke about in terms of their mental health; this can be seen in the following example, given by a straight male TLM informant: *people don't- might not want to hear what you have to say*. In context of the rest of his speech, this informant feels an obligation to say things that some people might find unpalatable. The collocation of *you have* is an indication that the intersection of male and straight fits with this idea of obligation. The negative keyword *where* is used to signify a situation rather than a specific place in both the TLM and RLM straight corpora. This example from RLM exemplifies this: *I think that's quite heavily affected by where I'm at in my life*. The use of the term where does not change with low mood, purely the frequency. The pattern of my data prior to this point has involved indications that TLM informants are more likely to talk about generalities and RLM informants about concrete examples. This negative keyword is a rare example where both reference general situations rather than specific places.

The only queer keyword in the speech of more than one individual is *within*, which occurs six times in this subsection of the corpus, shown in quotes 6.45 to 6.50:

Quote 6.45: Actually you… you are important as well within that and you deserve that
Quote 6.46: it might be a high priority within that as in I I the way
Quote 6.47: I thought that I was a bit of a oddball. within my own family
Quote 6.48: to have. a feeling of accomplishment within oneself
Quote 6.49: who I am in the world and who I am. within the context of my family
Quote 6.50: giving something uhm enough respect and status within yourself

In quotes 6.45 and 6.46, the informant is qualifying what may be important as part of a grouping, whilst in quotes 6.47 and 6.48, the informant is referring to their own sense of being strange when compared to their family. In quotes 6.49 and 6.50, the informants are talking about their internal sense of wellbeing. The word choice for each of these is poignant, since the context shows the

sense of exclusion that many within the queer community have felt for much of their lives. *Within* is a linguistic reflection of the stories shared and typified in books such as *Desire, Love, Identity* (Robinson & Villasenor, 2019). This book is a collection of stories written by Nottinghamshire residents about their experiences around their sexuality. *Desire, Love, Identity* (Robinson & Villasenor, 2019) includes stories of exclusion from mainstream society as well as finding belonging in the queer community.

The group based on informants who have an undergraduate degree or the equivalent of part of an undergraduate degree had one keyword specific to this demographic. The negative keyword specific to this educational demographic is *been*. Within the RLM group of the undergraduate degree grouping, the other keyword *been* is mostly used to talk about past experience. Within the TLM group from this demographic there is one example of the word being used: *I hadn't been able to be me*, which is more personal and less past focused to the multiple uses of the term in the comparison group's recollections. Seeing the use of this keyword in the comparison group provided an indication of how to address my second research question. The use of *been* specifically could suggest that the reference informants were looking to the past more than those experiencing low mood, although tense is not a consistent marker of temporal focus. This is an area that I explore further in chapter 8.

6.8 Word count, content and diversity analysis discussion

Through this chapter I have performed an initial exploration into my first research question: *What linguistic markers are prevalent in the verbal language of people experiencing low mood?* This exploration was specifically based on examining the language of the low mood informants when compared to the reference group.

In examining word counts, I found that there was no statistical link between low mood and number of words used, either generally or when examined according to social grouping. This is despite potential greater loquaciousness for informants experiencing depression in the pilot group, which fits with an existing Russian language study into depression (Smirnova et al., 2018).

In terms of what my informants discussed, I found that TLM informants were likely to talk about slightly fewer topics than RLM on average, but that there were many shared topics between the two corpora, with connections, the self and society as the most common. The main distinction was a greater likelihood for TLM informants to attribute importance to physical rather than emotional topics.

The next part of the chapter considered how my informants spoke about these topics. To start, I analysed the lexical diversity of the informants' language, and found that the TLM informants used language that was less lexically diverse than the RLM informants. To clarify, the TLM informants were more likely to repeat the same words rather than utilise a diverse vocabulary when discussing their topics of importance. This is potentially a significant finding, because it builds on a study by Park and Conway (2018) in which they found users on mental health internet fora used less lexically diverse language than users on recreational physical health fora. Their study highlights that people experiencing mental illness generally have a less lexically diverse vocabulary than a reference group. My study shows that low mood as a distinct experience also results in a reduction in the lexical diversity of their own speech when compared to a reference group.

The TLM compared with RLM keyword searches, alongside some wider contextual analysis of BNC spoken sample, have identified two potential linguistic features characteristic of low mood. The first of these is an increased likelihood of self-focus in TLM when compared to RLM. The second is a concentration on abstract concepts rather than concrete things or people of importance for the TLM informants when compared to the RLM informants. An increased focus on the self is also a finding of Bucur et al. (2021) whose analysis of social media posts concluded that an increased use of personal pronouns were an indicator of depression. A focus on abstract rather than specific things is a new finding for this study. A similar finding from psychology is that patients with mental health issues were more likely to talk about their symptoms generally rather than specifically (Greer et al. 2004, p.524). A wider tendency towards abstraction over concrete findings are distinctive to this corpus and requires further investigation. The analysis in this chapter suggests that there are linguistic characteristics identifiable as a psycholect in the language of informants with low mood, and that these may have diagnostic value in future research and applied settings.

Chapter 7 Part of Speech tagging analysis

7.1 Introduction

This chapter builds upon the exploratory analysis of the previous chapter to further answer my first research question: *What linguistic markers are prevalent in the verbal language of people experiencing low mood?* In this chapter I take an inductive approach, using part of speech tagging analysis to perform a corpus linguistic examination of the texts based on observable patterns. In order to look at grammatical patterns of the speech of informants experiencing low mood (TLM), I will be looking at the part of speech tagging categories that these informants were more likely to use, when compared to the reference group (RLM). This will contribute towards the outline of a psycholect associated with the verbal utterances of people experiencing low mood.

Part of Speech (PoS) tagging has been used on health communication data in studies related to Dementia diagnosis (Wen et al., 2023); schizophrenia (Bayer et al. 2023); and online suicide notes (Adil Jafar & Abdul-Salam Jasim, 2022). When analysing data from individuals with dementia, Wen et al. (2023, p.e36590) found that these individuals used more adverbs than their control data set. Further to this, the use of prepositions, personal pronouns, particles, present participle verbs, predeterminers, singular proper noun, adjectives, common nouns, past tense verbs, base form verbs, and wh- pronouns (such as *who* or *whose*) were distinguishably different for patients with dementia when compared to the control group (Wen et al. 2023, p.e36590). Bayer et al. (2023) are analysing the PoS tags of individuals experiencing schizophrenia to test if these individuals used less syntactically complicated sentences than their control set. The results of this study are not yet available, but their intent is to use a combination of PoS tagging with speech graphs to analyse how complicated the syntax used by the individuals with schizophrenia is (Bayer et al., 2023, p.85). Adil Jaafar and Abdul- Salam Jasim (2022) examined suicide notes from online fora, finding an increased use of personal pronouns in the suicide notes against a comparison corpus. In a part of speech tag comparison, they found that the suicidal forum users wrote using more verbs, nouns, adjectives and adverbs than the reference users (Adil Jaafar & Abdul-Salam Jasim, 2022, p.15).

Of greatest relevance to my data, Bucar et al. (2021) used a comparative PoS tagging approach on a depression based-data set. They compared the language of online depression forum users who explicitly stated they had a depression diagnosis with users who did not state this (Bucar et al., 2021, p.3). From this data set, they found that the depression dataset had a higher incidence of personal pronouns and verbs, compared to fewer nouns when compared to the posts from undiagnosed users. The researchers suggest that the reason for this is “a defective linguistic structure with less interest in the environment” (Bucar et al. 2021, p.4). To clarify, this indicates a combination of less complicated writing patterns and a greater focus on the self rather than external factors for the depression users when compared to the reference group. In my PoS tag analysis, discussed below, I also find a difference in usage of some personal pronouns and verbs, when comparing TLM with RLM. I explore the most key elements of my PoS tag analysis concentrating on prepositions, general adjectives, the verb *be*, personal pronouns and modal verbs.

7.2 Parts of Speech tagging

I used the *Wmatrix* interface to tag my corpora, and calculate which PoS tags had a higher or lower incidence in the verbal language of my TLM (low mood) informants when compared to RLM (the reference group). I normalised the data output per 1,000 words to enable a direct comparison of the TLM incidence of each PoS tag with the RLM tagged corpus. I used the filter function built into *Wmatrix* to show PoS tags that appear more than 6 times and with a log likelihood score above 6.63. This log likelihood score indicates a significance of at least 95%, and ensures that I am considering only the most relevant PoS tags (Groom, 2023).

The full results of the part of speech tag comparison of TLM against RLM are outlined in Table 7.1. I list the part of speech tag immediately followed by the sub-type of that part of speech in round brackets, followed by the word form of the part of speech tag in curly brackets. For example: the second most key part of speech tag is a type of personal pronoun; that type is objective first person singular, and this part of speech is the term *me*.

	Key Part of Speech tag	Norm. TLM	Norm. RLM	LL
Positive	General preposition	51.47	40.45	21.95
	Personal pronoun (objective 1st person singular) {me}	13.28	8.74	16.47
	General adjective	61.73	52.07	13.04
	Verb (base form) {be}	8.79	5.65	12.09
	Personal pronoun (singular, reflexive) {e.g. yourself, itself}	3.21	1.83	6.82
	Auxiliary modal {eg can, will would}	13.19	10.71	6.36
Negative	General adverb	33.06	45.18	34
	Personal pronoun (3rd person singular objective) {him, her}	0.55	3.13	31.05
	Verb (lexical, past tense) {e.g. gave, worked}	4.03	6.91	13.22
	Verb (present participle) {doing}	0.46	1.65	11.4
	Interjection {e.g. oh, yes, um}	12.82	16.91	10.26
	Verb (do, infinitive) {I may do... To do…}	1.28	2.61	7.77
	Verb (lexical -ing participle) {e.g. giving, working}	12.27	15.67	7.64
	Verb (past tense) {had}	0.55	1.44	6.79
	Verb (finite base form) {have}	3.75	5.65	6.78

Table 7.1 Key parts of speech TLM compared to RLM

The existing PoS tag studies mentioned in the introduction, such as Bucur et al. (2021) purely concentrated on positive PoS tags, but I chose to also consider the negative PoS tags, for a more complete picture. Since the negative PoS tags are more representative of the reference corpus, this provides a baseline to compare the TLM corpus with throughout the rest of this chapter.

Starting with an overview of main trends from this comparison of TLM and RLM, there are more negative than positive key part of speech tags. This means that there are more PoS tags that are statistically less likely to be in TLM compared to RLM than there are PoS tags that are statistically more likely to be in TLM compared to RLM. A collective finding from the key PoS tags is reduced lexical variety; a focus on the self, and varying levels of power and agency. *General prepositions* are statistically the most characteristic of the data, and these are short linking words. There are two types of personal pronoun in this positive key PoS tags list: the objective first person singular *me* and the singular reflexive which includes terms such as *yourself*. Another key PoS tag is *general adjectives*, which are describing words and includes terms such as *content*, *important* and *physical*. Verbs are words that signify “actions, events, processes and states” (Carter & McCarthy, 2010, p.158) and the base form of the verb *be* is a key PoS tag. Finally, *auxiliary modal verbs* which are a type of verb that indicate levels of confidence and importance (Carter & McCarthy, 2010, p.420). I

will explore the themes that run through these positive key PoS tags later in this chapter, in sections 7.3-7.6, but first, I discuss the negative PoS tags.

The most key negative PoS tag is *general adverbs* which are modifiers such as *specifically*, *inextricably* and *emotionally*. A reduced use of adverbial descriptors may be indicative of a reduced lexical complexity in the TLM informants when compared to the RLM informants. Lexical complexity, or lexical diversity, refers to how many different rather than repeated words someone uses (McCarthy & Jarvis, 2010, p.381). Lexical density and diversity are described in more detail in section 6.4. Examining the words that fit into the category *general adverbs*, there are 93 different types in TLM compared to 211 different types in RLM. More informants in the RLM corpus means that they are more likely to have different speech patterns and therefore exhibit a greater range of vocabulary. compared the proportion of *general adverbs* in TLM and RLM to the BNC spoken sample and found that TLM has 4% fewer and RLM has 4% greater proportion against the BNC sample. Therefore general adverbs are used differently across my TLM and RLM corpora when compared to spoken British English generally, as well as when compared against each other. To explore how these *general adverbs* are used, here are some examples of *general adverbs* in use from TLM and RLM, the TLM examples are quotes 7.1, 7.2 and 7.3 and the RLM examples are quotes 7.4, 7.5 and 7.6:

TLM quote 7.1: and when I say safe I mean emotionally safe rather than physically safe
TLM quote 7.2: important thing to me is is probably . being relatively happy in myself
TLM quote 7.3: which uhm I freely admit is usually quite a struggle to achieve
RLM quote 7.4: that 's why you 're so emotionally sort of aware of yourself
RLM quote 7.5: uhm. but having a relatively comfortable life where we can make the decisions
RLM quote 7.6: well-being is ... is usually at the top

In the examples for *emotionally*, in quote 7.1 the TLM informant speaks of a desire to feel emotionally safe, when compared to the RLM informant in quote 7.4 who speaks about emotional self-awareness. Feeling emotionally self-aware is hierarchically different to emotional safety. Safety is something important to survival, and something one can build upon in order to obtain awareness. The difference in use between these informants may indicate a contributory factor for the TLM informant to feel low, compared with the security of the RLM informant. In quote 7.2, which is the

TLM example for *relatively*, the informant uses this adverb as a qualifier, they do not believe they can be completely happy in themselves, and therefore can only hope for being relatively happy. The RLM informant in quote 7.5 instead talks about a desire for relative comfort. Their happiness is assumed, as it was with the term *emotionally*. The term *usually* is used by the TLM informant in quote 7.3 to refer to the regularity of their own struggle, but by the RLM informant in quote 7.6 to refer to how other people view things.

The negative key PoS tag that is second in significance is the third person subjective singular personal pronoun, specifically *she* and *he*. Because of the significant role of pronouns in the PoS tagging analysis, I investigate personal pronoun usage throughout the corpora in section 7.5, including *she* and *he*. My second research question is *Do people with low mood focus on the past?* and I devote chapter 8 to answering this question. Therefore the third and penultimate most significant negative PoS tag: *past tense of the lexical verb* and *had*, will be included as part of my exploration of temporal focus. I will divide the rest of the negative PoS tags into *verbs* and *interjections*, to examine them efficiently. In terms of *interjections*, these include terms such as *yeah*, *hmm* and *gosh* and these are “words that help us talk and think, scaffolding the complexity of language” (Dingemanse, 2024, p.257). Both sets of informants use the interjection *uh* the most, and the TLM informants use 9 different interjections compared to the 18 different interjections that the RLM informants use. Dingemanse (2024, p.266) writes that interjectors used to push a conversation or story along are indicative of greater complexity of speech. The reduced number of *interjections* in TLM when compared to RLM is largely due to the relative corpus sizes. However, they may also be a further indicator that the TLM informants are less lexically complex in their verbal responses than the RLM informants. Based on the work of Taavitsainen (1998), *interjections* are also used as signifiers of emotion (Taavitsainen, 1998, p.197). A reduced usage of interjections in TLM when compared to RLM may therefore also be indicative of reduced emotional expression. This is not certain, since the most popular interjection is *uh*, which is a disfluency marker rather than an affective response. The multifunctional uses of interjections such as *uh* are something that a wider study could enable further insight into. There is mostly positive valence to the answers to the main

interview question: *when you think about what is most important to you, what comes to mind?* A lack of emotional lexis around positivity may imply anhedonia, which is a reduced interest in enjoyment or gratification (discussed in 2.4). Anhedonia is one of the symptoms of depression (WHO, 1993, pp. 81-86) and the reduced use of the *interjections* PoS tag in TLM compared to RLM indicates that anhedonia may also be associated with low mood.

The negative key PoS tagged verbs from a comparison of TLM and RLM that are present, infinitive, lexical and base in tense are: the present participle of do: *doing*; the infinitive of *do*; the lexical form of the -ing participle such as *giving* and *working*; the finite base form *have*. Verbs as a group are terms that “denote actions, events, processes and states” (Carter & McCarthy, 2010, p.301). *Doing*, *do* and the lexical form of the *-ing* participle are all indicative of agentive action. This would indicate that the RLM informants are more likely to articulate a sense of power and dynamism in their language. The finite base form of the verb *have* can be defined as active possession (*OED* 2025), once again indicating dynamism rather than passivity. The sense of agency and power indicated by these verb forms used less by the TLM informants when compared to the RLM informants is something that I will explore more in section 7.6.

7.3 Prepositions

The part of speech tag most likely to appear in the low mood corpus (TLM) compared with the reference corpus (RLM) was “general prepositions”. Linguistically, prepositions are short linking words which can be grouped into spatial and temporal categories, then further categorised into those with an intrinsic (e.g. on), relative (e.g. near) or absolute (e.g. at) frame of reference (Coventry & Garrod, 2004, p.136). Prepositions are function words, and the ratio of function to content words is one of the methods of examining lexical density (McCarthy & Jarvis, 2010). A higher proportion of this type of function words may help to explain the observed differences in lexical density between TLM and RLM.

General prepositions made up 10.34% of the words in the low mood verbal utterances, and 9.97% of the words in the reference corpus, which *Wmatrix* software registered as the most characteristic difference between TLM and RLM. Table 7.2 shows all prepositions used by TLM informants only, across both corpora, and by RLM informants only. Asterisks signify words that are in the speech of one informant only.

General preposition

TLM only	Norm freq	Both	TLM (norm)	RLM (norm)	RLM only	Norm freq
alongside*	0.3	about	3.5	3.4	above*	0.0
versus*	0.2	along*	0.2	0.2	across	0.1
		around	1.1	1.5	after	0.5
		as	4.7	5.0	against*	0.0
		at	2.8	3.7	amongst*	0.0
		before	0.2	0.5	beside*	0.0
		behind	0.1	0.1	down	0.9
		between	0.7	0.2	including*	0.0
		beyond*	0.1	0.1	inside	0.0
		by	1.0	1.1	less	0.5
		during*	0.3	0.1	off	0.5
		from	2.9	1.6	onto	0.1
		in	13.3	10.8	out	2.4
		into	1.8	1.7	outside	0.2
		like	8.3	8.3	pre*	0.0
		on	3.6	4.4	round	0.2
		over	0.2	0.9	since	0.1
		through	1.0	0.8	throughout*	0.0
		to	35.9	26.8	underneath*	0.0
		towards*	0.2	0.1	up	2.2
		under	0.5	0.2		
		upon	0.4	0.0		
		within	0.6	0.4		
		worth	0.4	0.1		

Table 7.2 Prepositions from TLM and RLM (normalised per 1,000 words)

Note that the TLM informants use only two words that the RLM group do not, whereas the RLM informants use 20 words that the TLM group do not. This indicates a potential difference in lexical diversity within this single PoS tag between these two groups. RLM is a larger corpus, and with more words and more informants there are more idiolects being analysed from RLM compared to TLM. In section 6.4, we found indications of increased lexical diversity for RLM informants when compared to TLM informants across the corpus, not related to the size of the corpora. In this section we explore

if the increased diversity for the RLM informants is due to the greater wordcount or reduced lexical diversity from the TLM informants for the PoS tag *prepositions* specifically.

I used *AntConc* to perform a search of the collocates of the *prepositions* PoS tag. Collocates are words that are statistically likely to be found in the company of a target word (Baker et al., 2006, p.36), see section 2.3 for information on collocates and how they calculated. *AntConc* does not have the function to search for a specific PoS tag, and so I prepared a version of the main study corpora in which I replaced every instance of a preposition with a code that did not exist elsewhere in the texts (specifically the code *XX*) then used this code as the search term for *Antconc* to calculate collocates. The collocates for *prepositions* PoS tag in TLM and those for RLM are in Table 7.3, the word likely to be found within 5 words of prepositions is in the collocates column, the normalised frequency is the next column, the number of times the collocate and *be* are found beside one another are in the Freq LR column, the number of times the collocate is found to the left and right of *be* are in the next two columns respectively, then the statistical likelihood of them being found together, measured with log likelihood, is the column on the end.

Collocate		Freq(norm.)	FreqLR	FreqL	FreqR	Likelihood
TLM	able	210	46	27	19	36.753
	s	2860	163	93	70	28.857
	be	1040	137	61	76	28.347
	me	1600	193	65	128	26.955
RLM	xx	34850	2442	1221	1221	49.838
	able	440	81	60	21	49.208
	s	10610	676	407	269	40.53
	be	2600	299	124	175	39.469
	think	3590	194	106	88	36.362
	trying	430	68	53	15	28.702
	want	720	100	82	18	28.677
	terms	190	38	8	30	27.337
	going	810	108	83	25	26.874
	looking	250	44	32	12	24.178
	me	4320	432	154	278	22.778
	t	3330	197	124	73	21.087
	go	1010	122	71	51	20.355

Table 7.3: Collocates of PoS *prepositions* in TLM and RLM

The collocates of *prepositions* from TLM are also in RLM, and in the same order of likelihood. The increased likelihood of prepositions collocating with one another in the language of RLM informants when compared to TLM is something to explore. Whilst the TLM informants are statistically

significantly more likely to use prepositions than RLM informants, the RLM informants are statistically likely to use prepositions in close proximity with each other. An example of using these close together in RLM with prepositions (in bold) is: *we're able* ***to****.* ***as*** *much* ***as*** *we can help* ***out*** *other people*. An example of prepositions (in bold) close together in TLM: ***to*** *uhm…* ***to*** *things* ***like*** *what* ***like*** *we've talked* ***about***. It is clear in the difference between these representative examples that there is slightly more repetition and hesitation from the TLM informant, compared to the more lexically dense language of the RLM informant.

The repetition of lexical terms from TLM informants compares to the greater lexical richness from the RLM informants across the *prepositions* PoS tag. This is a small-scale example of the greater lexical diversity from RLM compared with TLM discussed in section 6.4. The use of prepositions across my corpora further builds on the indications that lexical diversity is part of the language patterns that distinguish TLM from RLM. Specifically, that informants experiencing low mood are more likely to repeat words rather than using new terms, when compared to the reference informants.

7.4 General adjectives alongside *be*

General adjectives are the third most key PoS tag based on the comparison of TLM and RLM, whilst the base form of the verb *be* is the fourth. Similar to the prepositions of the previous section, the *general adjectives* PoS tag is indicative of reduced lexical diversity in the verbal language of TLM when compared to RLM informants. I group *general adjectives* and *be* together because they are both indicative of informants talking about generalities rather than specifics, as introduced in section 6.3.

Adjectives are describing words, and CLAWS divides adjectives into PoS category tags: general, catenative, comparative and superlative. Within TLM, *general adjectives* include *silly*, *corny*, *comfortable*, *good* and *peaceful*. In total, the number of distinct *general adjectives* in TLM is 190 compared to 540 in RLM. As with the *prepositions* PoS tag, this is fewer than half of the specific words that fit into this PoS tag in the TLM corpus when compared to the RLM corpus. Therefore, whilst there is a greater relative volume of general adjectives in TLM, there is greater repetition of

these general adjectives when compared to RLM. This once again indicates reduced lexical diversity for the TLM informants when compared to RLM.

Some of the general adjectives from TLM are also keywords, these being *easy*, *useful* and *vital*. *Easy* is used as an example of negativity, whereby the informants express their dissatisfaction at the difficulties of life. *Vital* is used to emphasise the importance of a concept and is an example of TLM informants using extreme modifiers. Finally, *useful* is an abstract (rather than tangible and specific) concept that the informant would like to exemplify. I explored the usage of these words across the corpora in section 6.5.

In sections 6.5 and 6.7 we established that the TLM informants were more likely to talk about concepts rather than physical things. A greater volume of describing words is a linguistic reflection of this; the data suggests it takes more descriptors to talk about concepts than it does to talk about a physical thing. For example, from TLM:

> Quote 7.7: so it's **human** connection. **creative** expression learning **new** things… feeling both **connected**. but also having time **alone** and uhm… joy

The *general adjectives* PoS tag as categorised by *Wmatrix* are in bold, and this speech is typical of the TLM informants. To clarify, rather than identifying a connection with a specific human, this informant speaks about connections with humans generally. Every concept in quote 7.7 is modified with an adjective, except for joy. The informant chose to use descriptors to outline an idea rather than a thing.

To contrast the example of a TLM informant using modifiers when talking about concepts is an example of a typical RLM informant talking about connection, and the general adjectives they use:

> Quote 7.8: I suppose you could say people **human** connection but that's that's **important**… uhm [pause] yeah I I think all in all regardless of. how **rich** you are or this or that I think the the people and it may not be you know you you don't need… hundreds of people you just need a few people who you know and know you and after that I think life is far easier

Once again, the *general adjectives* PoS tag as categorised by *Wmatrix* are in bold, and there are less instances of this PoS tag in this RLM excerpt than in the shorter TLM quote 7.7. RLM quote 7.8 is exemplary of both the use of *general adjectives* and also the attitude towards importance. Unlike TLM quote 7.7 which was before this, the informant talks about human connection but also qualifies

that this is a limited number of mutually known people. The RLM informant in quote 7.8 talks about mutual relationships with known individuals rather than the wider concept of human connection. In the use of general adjectives, unlike the TLM informant, the RLM informant does not modify the language they use for emphasis or clarification. The nebulousness of the topics of importance listed by the TLM informant in quote 7.7 requires greater adjectival input as part of the explanation of what they are, when compared to the distinct connections foregrounded by the RLM informant in quote 7.8.

The base form of the verb *be* is defined in the Oxford English Dictionary (2025) as “Without required complement: to have or take place in the world of fact, to exist, occur, happen”. Theoretically, this appears in direct contrast to the previous example, whereby the TLM informants talk about generalities rather than the specifics of the RLM informants. I examined collocates to see if there was a difference in how the TLM and RLM informants used the word *be*. The results are in Table 7.4. In this table, the word found within five words of the word *be* and the corpus are under the collocates column, the normalised frequency is the next column, the number of times the collocate and *be* are found beside one another are in the Freq LR column, the number of times the collocate is found to the left and right of *be* are in the next two columns respectively, then the statistical likelihood of them being found together, measured with log likelihood, is the column on the end. Collocates of *be* that are in both the TLM and RLM corpora are in bold.

	Collocates	Freq(norm.)	FreqLR	FreqL	FreqR	Likelihood
TLM	**to**	3940	91	60	31	57.94
	would	470	26	20	6	49.90
	able	210	17	5	12	44.09
	order	30	5	5	0	20.01
	content	80	7	2	5	19.12
	oneself	40	5	2	3	17.09
RLM	**to**	11730	186	136	50	138.92
	might	320	22	19	3	68.95
	would	710	24	23	1	44.56
	able	440	18	1	17	39.29
	should	290	14	13	1	34.67
	honest	70	6	0	6	21.32
	thoughtful	30	4	1	3	17.75
	could	540	13	11	2	16.99
	can	1630	25	23	2	16.99
	tends	40	4	4	0	15.43

Table 7.4 Collocates of *be* from both TLM and RLM

Three of the collocates are the same across both corpora (*to*, *would* and *able*). The other three TLM collocates for *be*, specifically *order*, *content* and *oneself*, are not collocates for the RLM informants. Since the three TLM specific collocates are infrequent items in the corpus, I chose to compare one word that spanned both collocate lists. This is in order to offer indicative examples of collates in use that span both corpora. *Able* collocates with *be* 17 times in TLM (210 normalised frequency) compared with 18 times in RLM (440 times normalised frequency) and so this was my choice for a direct comparison. I compared each use of *able* and *be* when collocated, and found more positivity, agency and certainty in the RLM group when compared with TLM. For example, quotes 7.9, 7.11 and 7.13 from TLM when contrasted with quotes 7.10, 7.12 and 7.14 from RLM:

TLM	RLM
Q. 7.9 to be able to look in the mirror and. be happy	Quote 7.10 would be able to find some joy
Quote 7.11 to be able to be. on this planet	Quote 7.12 I like to be able to do things
Quote 7.13 to be able to help and support people	Q. 7.14 position to be able to help somebody else

Table 7.5: Collocate in context of *be* and *able* across TLM and RLM

Quotes 7.13 and 7.14 particularly illustrate the recurring pattern of general grand scheme ideas rather than specifics in the language of the TLM informants when compared with the RLM informants. The TLM informant in quote 7.13 wants to be able to help and support people generally (as a mass noun), whilst the RLM informant in quote 7.14 talks about *some body* else, one body, one individual, rather than the grouping of *people*. Whilst this individual is still abstracted rather than specific for the RLM informant, it is easier to help one body than a generalised group. The RLM informant therefore sets more achievable expectations for their capacity to support than the TLM informant who wishes to support people generally.

7.5 Personal pronouns

Two of the most key PoS tags from the comparison of TLM with RLM are personal pronouns, specifically the objective first person singular *me* and singular reflexive personal pronouns. Wales describes personal pronouns as “useful ‘shorthand’ referring expressions, used as alternatives either to ostensibly more explicit or descriptive forms, or to those more cumbersome” (Wales, 1996, p.4). According to a meta- analysis performed by Edwards and Holtzman (2017), increased personal

pronoun use, specifically first person, is a common theme for people experiencing depression. I examine the use of all personal pronouns across both corpora to enable oversight of how these patterns of personal pronoun use occur. I begin by examining the use of personal pronouns across both corpora:

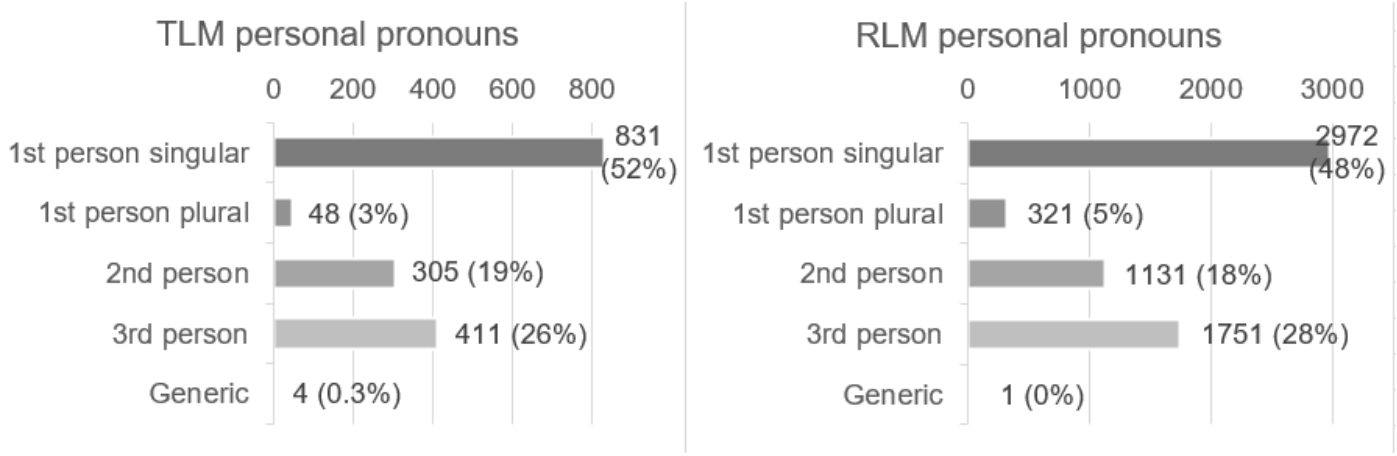

Figure 7.1: Proportion of personal pronoun use across TLM or RLM

In these charts, the proportions of each category are approximately the same for both corpora, though the first person plural and second person are both slightly proportionally higher in RLM. The usage of generic personal pronouns when compared to the rest of the categories is too small to display- 4 times for TLM and 1 time for RLM. I used Carter and McCarthy (2010, pp.376-385) to decide which pronouns to put in each category, and a breakdown of each follows in figures 7.2, 7.3 and 7.4. Carter and McCarthy (2010, p.382) define terms such as *our, your* and *their* as “possessive determiner personal pronouns” and these determiners have therefore been included in the list. I chose to separate each type of pronoun by person and number since the breakdown of the overall volumes of each type of personal pronoun as displayed in chart 7.1 is not significantly distinct between the two corpora.

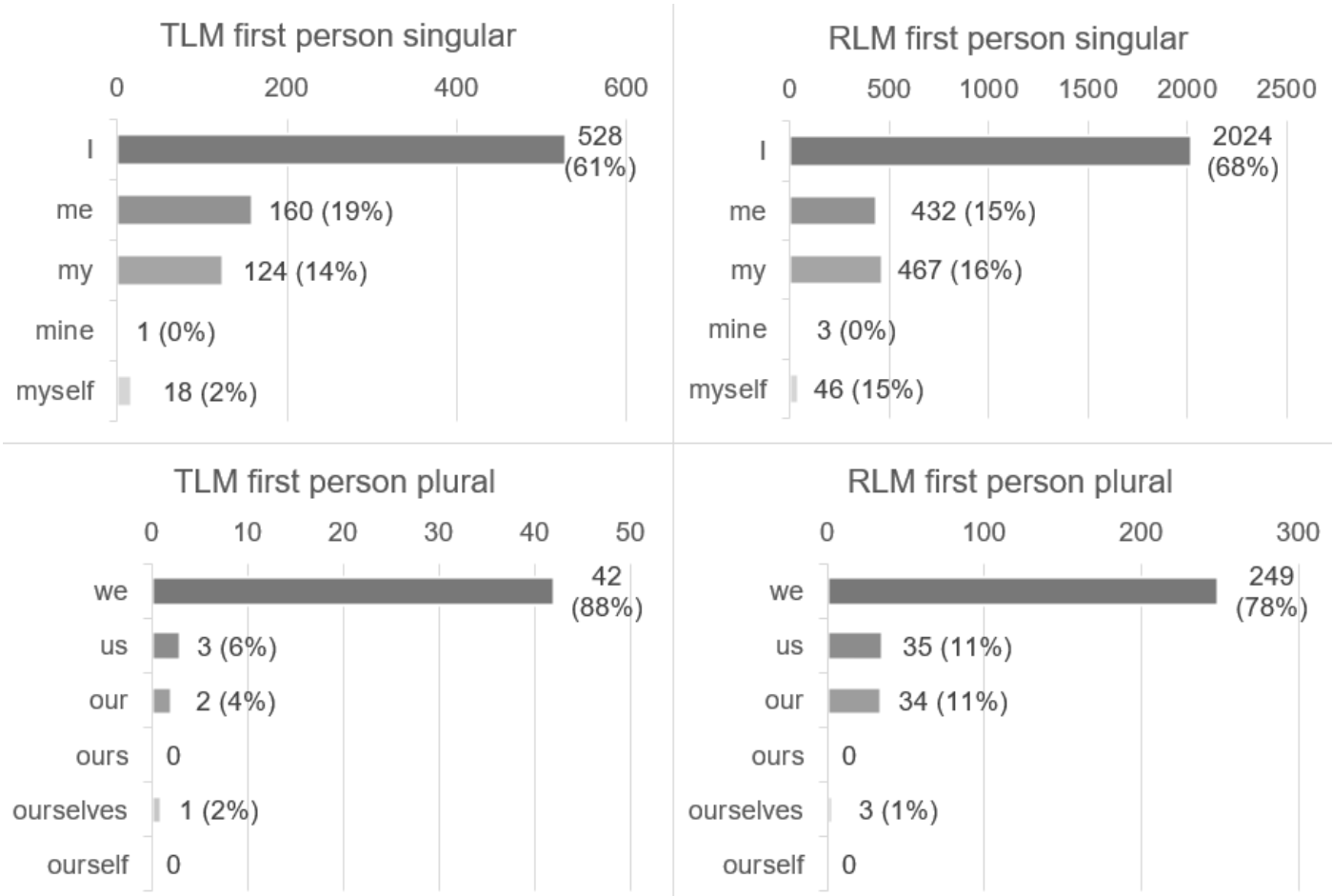

Figure 7.2: First person single and plural pronoun use in TLM and RLM

In figures 7.2, 7.3 and 7.4, the percentage figures are the total use of that specific pronoun compared to others of this type. For example, *I* is used in 61% of the instances of first person singular pronoun use in the TLM corpus. The use of the first person singular forms is similar across both corpora, and *mine* does not register since it is used once in the TLM corpus and three times in the RLM corpus. The charts show that both TLM and RLM informants favour the first person plural *we*, whilst RLM informants are proportionally more likely to use the first person plural *us* and *our* when compared to TLM. Since this difference is not part of the keyword or key PoS tag search, they are not statistically significant but the informants do show mild preferences.

The use of *me* is particularly relevant due to its prominence as a key PoS tag and because it is one of the specific personal pronouns mentioned in the many existing studies on the language associated with depression, as meta analysed by Edwards & Holtzman (2017). I give quotes 7.15 up to 7.18 for examples of *me* in context from TLM followed by quotes 7.19 up to 7.22 for examples of *me* in context from RLM.

TLM quote 7.15: important to me because they allow me to exist. In the world uhm other things
TLM quote 7.16: until then. I hadn't been able to be me and when I went/ when I left my marriage
TLM quote 7.17: everybody else uhm. yeah looking after me probably needs to matter more than it does
TLM quote 7.18: going for a. walk in the morning to me is quite important just to. balance

RLM quote 7.19: it's the things that. surround her and me and my husband and sort of… protect
RLM quote 7.20: impact on… both the world close to me and the world around me… uhm… but
RLM quote 7.21: happened without the people around me so you know credit where it's due
RLM quote 7.22: she would literally destroy me. uhm obviously my chums… chums are important

Quotes 7.15 to 7.18 from TLM display a trend towards the theme of introspection that is congruent with the work of Şimşek (2013). They claimed that self-rumination rather than self-reflection is symptomatic of depressive behaviours which also maps onto this low mood data. In quote 7.15, the informant speaks about their important things as permitting them to exist, showing a lack of agency and positivity. They are existing rather than living, and they do not mention anything that makes their existence positive, only things that allow that existence to continue. Quote 7.16 is an informant talking about how their sense of self had been repressed, until they left their marriage. Within the wider context of this informant's speech, they claim that their previous spouse tried to force them into roles that they were not suited to. The informant further elaborates that this is something they only came to terms with once they had distance from the previous spouse and opportunity for introspection. In quote 7.17, the informant claims that self-care should be more important to them. They are aware that it is important as a concept but they choose not to prioritise their own, purely to think about how important the concept is. In 7.18, the informant clarifies that a walk in the morning is not generally important, to me is a qualifier. They indicate that their personal morning walk or a morning walk in general may not matter to others, but it does to them. This theme of introspection is in keeping with Şimşek (2013) but in contrast to the RLM informants.

The RLM informants also use the term *me*, as exemplified from quotes 7.19 to 7.22. However their usage is as part of a compound noun phrase, making the referent less likely to be the speaker than in the language of the TLM informants. Despite *me* being a personal pronoun, RLM informants use it to group themselves in with an external person or entity. These include the informant's husband in quote 7.19, *the world* in quote 7.20, surrounding people in quote 7.21 and the wife and friends in

the quote 7.22. This indicated a reduced level of introspection and a sense of belonging that is not linguistically present in the TLM group.

In terms of first person plural pronouns, *we* is the most frequently used for both TLM and RLM, though *we* is proportionally used more in TLM (88%TLM:78%RLM). This does not appear congruent with the self-focus involved with a greater than average usage of first person singular pronoun use reported by Edwards and Holtzman (2017, p.64) in their meta-analysis. I therefore examine the use of *we* across TLM and RLM, with quotes 7.23 to 7.25 from TLM first followed by quotes 7.26 to 7.28 from RLM:

TLM quote 7.23: I think I feel like. uh… we're we're getting more. aware of… uh how to be…
TLM quote 7.24: how. we as people and are connected to everything around us
TLM quote 7.25: things like what like we've talked about the the those,those those concepts
RLM quote 7.26: all tied up in all of those things that we've been talking about about nurturing
RLM quote 7.27: she is good at and the other way round yet we're not complete opposites
RLM Quote 7.28: I know people say we're destroying the world. uhm.

In the majority of the TLM uses of *we*, the informants use the term generically, as in quotes 7.23 and 7.24. Quote 7.25 from TLM and the quote 7.26 from RLM involve the informant talking about the interview, and grouping themselves in with the interviewer. Using *we* as a grouping term with the interviewer is rare in both corpora. Quote 7.27 from RLM is a frequent occurrence in RLM but does not occur at all in TLM, whereby the informant groups themselves together with a specific person who is not the interviewer. Quote 7.28 from RLM involves more of a generic *we*, as the informant groups themselves in with the rest of humanity. The generic use of *we* is a smaller proportion of the uses of *we* in RLM than it is in TLM. Every use of *we* across both corpora is inclusive, whereby the informant groups themselves in with others rather than separating themselves out. The various uses of *we* across the corpora indicate a greater sense of belonging in the RLM informants and an increased likelihood of speaking generically in the TLM informants, when compared against each other.

The usage of second person pronouns was proportionally similar across TLM and RLM, as evidenced in figure 7.1. Existing research does not show any correlation between depression and

the use of second person pronouns (Edwards & Holtzman, 2017). In the interest of being thorough in the exploration of my data, I explore the usage of second person pronouns within these groupings across the corpora.

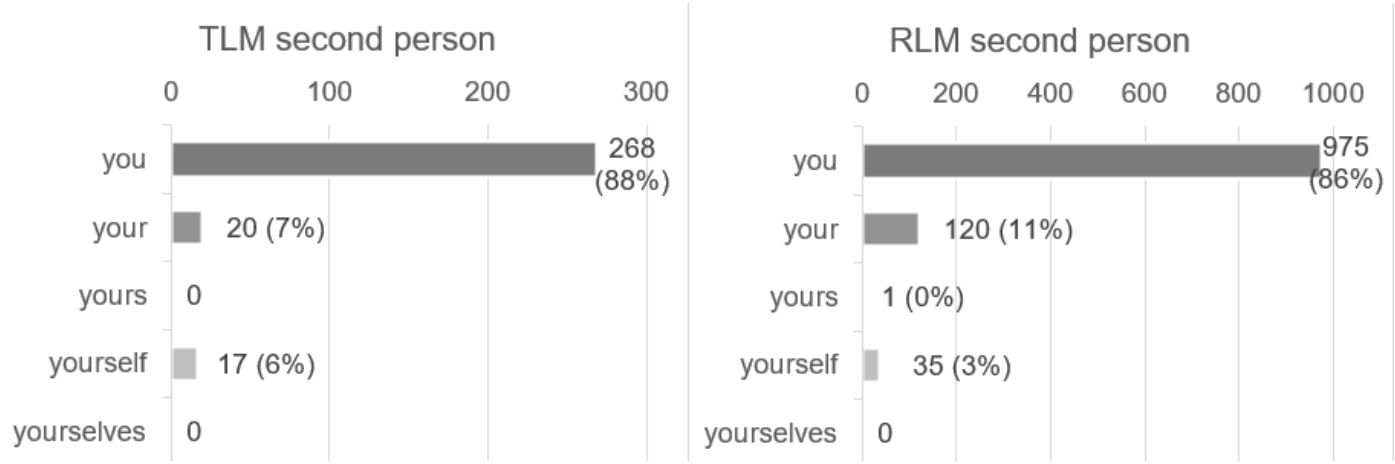

Figure 7.3: Second person pronoun usage within TLM and RLM

Second person singular *you* is the most used second person pronoun across TLM and RLM, and RLM informants use the possessive *your* to the singular reflexive *yourself* proportionally more than in TLM. Singular reflexive personal pronouns are a key PoS tag, across TLM and RLM. These are *yourself, itself, myself, himself and herself*. Since the proportional use of *yourself* is also different across the corpora, I examine some representative examples of this term in context, with quotes 7.29 to 7.31 from TLM followed by quotes 7.32 to 7.34 from RLM:

TLM quote 7.29: . wrapping it up in yourself. it becomes a bit selfish
TLM quote 7.30: just being at peace with yourself and and who you are
TLM quote 7.31: go actually look after yourself because you matter
RLM quote 7.32: support of the folk around me like yourself and uhm. my bubble here
RLM quote 7.33: the fact that you have to look at yourself in the mirror every day
RLM quote 7.34: not putting pressure on yourself is is quite hard

Every use of *yourself* in TLM is used in a generic sense and reflexively, as in quotes 7.29 to 7.31 whereby the *yourself* is a general rather than specific individual. Most of the uses of *yourself* in RLM were also reflexive and generic, with the exception of quote 7.32 from RLM, where the informant was referring to the second person reflexive *yourself* (in this case, me as the interviewer). The similarity in use of *yourself* indicates that it is the volume rather than the context of this reflexive term

that differs between TLM and RLM. Previous studies had found no patterns in the use of *yourself* in depressed data, my data also shows that the use of this term is not distinctive in my low mood data.

Existing research into pronoun use and mental health has found patterns in the use of first person pronouns, but not second or third (Edwards & Holtzman, 2017). As with the previous section on second person pronoun usage, it is pertinent to thoroughly explore my data and therefore I also consider use of third person pronouns. I found that third person pronoun usage has the greatest proportional difference between the TLM and RLM corpora:

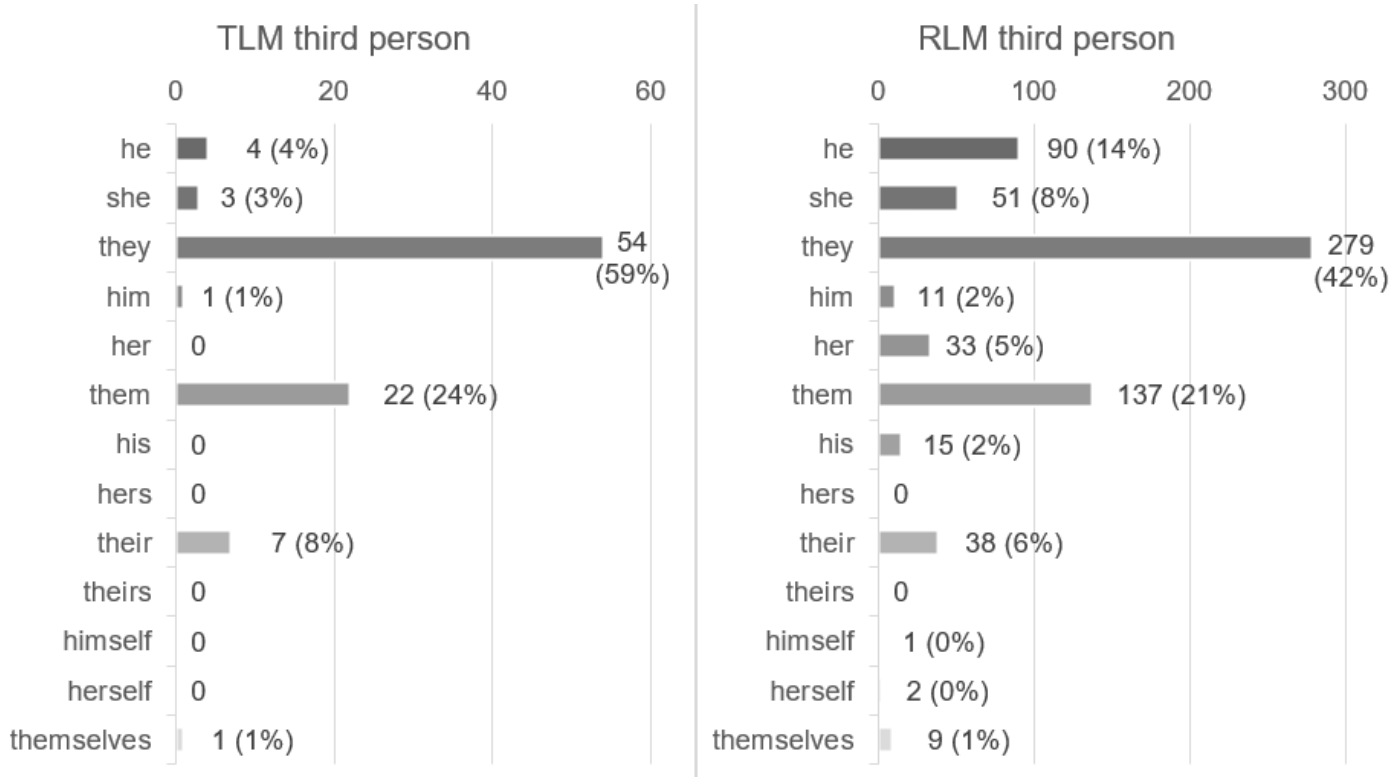

Figure 7.4: Third person pronoun usage in TLM and RLM

This comparison of third person pronoun use indicates a greater degree of lexical diversity in the RLM corpus when compared to TLM. A further exploration of the precise terms in TLM frequency order will enable a greater understanding of the lexical diversity, each figure is a percentage of total third person pronoun usage.

3rd person pronoun	TLM	RLM
it	77.1%	61.5%
they	13.1%	15.9%
them	5.4%	7.8%
their	1.7%	2.2%
he	1.0%	5.1%
she	0.7%	2.9%
him	0.2%	0.6%
themselves	0.2%	0.5%
its	0.2%	0.3%
itself	0.2%	0.2%
her	0.0%	1.9%
his	0.0%	0.9%
herself	0.0%	0.1%
himself	0.0%	0.1%

Table 7.6 Proportional 3rd person pronoun use in TLM and RLM

Neither TLM nor RLM use *theirs* or *hers*, and TLM features no uses of *her*, *his*, *herself* and *himself*. Third person singular objective personal pronouns such as *he* and *she* were a negatively key PoS tag in TLM compared to RLM. The decreased likelihood of TLM informants using third person singular objective as well as third person singular reflexive, possessive and determiner personal pronouns when compared to RLM informants is another pattern. Specifically, as discussed in section 6.3, the low mood informants refer to generalities rather than specifics. In this case, the TLM informants are less likely than the RLM informants to talk about individual people, which requires these gendered third person singular pronouns such as *he* and *herself*. Through this section on personal pronouns, we see that TLM informants instead talk about unspecific groupings, groups that they belong to (*we*), or themselves (*me*).

7.6 Modal verbs

The final key PoS tag from the comparison of TLM with RLM is *auxiliary modals*, which are also known as modal verbs. Heritage and Baker describe modal verbs as “doing words” “of necessity or possibility” (2022, p.564), and modality in discourse analysis encompasses “possibility, necessity, ability, obligation, permission and hypotheticality” (Collins, 2009, p.11). There are three main types of modal verbs: deontic which is based in duty e.g. *must*; epistemic which reveals levels of confidence e.g *may*; and the final category boulomaic which revolves around desire e.g. *should*

(Simpson, 1993, p.44). Some modal verbs may fall into more than one category based on context and therefore require close analysis to categorise. In Table 7.7 I list all modal verbs in TLM and RLM in order of TLM frequency, each number has been normalised per thousand words.

Modal verb	TLM norm.	RLM norm
Would	4.54	1.80
Can	4.33	3.05
Will	1.62	0.86
Could	1.30	1.45
ca (n't)	0.76	0.81
Should	0.65	0.81
Might	0.65	0.79
'll	0.54	0.81
wo (n't)	0.22	0.10
'd	0.11	0.71
May	0.11	0.12
Must	0.00	0.07
Shall	0.00	0.07
Need	0.00	0.05

Table 7.7: Modal verbs in TLM and RLM (normalised per 1,000 words)

The TLM informants do not use *must*, *shall* or *need*, unlike the RLM informants. As with the PoS tags *prepositions* and *general adjectives*, the less varied use of *modal verbs* further supports the trend of reduced lexical diversity in TLM when compared to RLM. With the following qualitative analysis, looking at the use of these terms in context, I explore this ongoing trend. This exploration is to better understand its possible causes, as well as to identify any other markers that may characterise a low-mood psycholect.

In examining the use of these terms in the RLM corpus, the three instances of *must* are used in a deontic sense, the informants asserting a sense of duty. The use of *shall* in RLM is epistemic, the informants ask for permission as in the following example where the informant asks: *shall I read this to you?* The RLM modal instances of *need* are boulomaic, expressing a desire for knowledge or things that are important to them respectively.

The modal verb that the TLM informants use the most is *would*, and this is used proportionally more frequently in TLM than RLM. In the TLM corpus, *would* is mostly used hypothetically, which fits under the category of epistemic since the context for this word is possibility (see section 6.5 for

keyword discussion of *would*). TLM informants also use *would* in a boulomaic sense, to express desire, and in a different epistemic sense, to express under confidence. In the RLM corpus, *would* is also epistemic and boulomaic, but is also used to refer back to the past, as with an informant who said: *my mum would say to me*. I include one example of *would* in epistemic and boulomaic senses in TLM (quotes 7.35 and 7.36) and RLM (quotes 7.37 and 7.38) in Table 7.8:

TLM	Epistemic	Q. 7.35	that *would* be a significant way to enhance. your feeling
	Boulomaic	Q. 7.36	I'm the person I *would* like to be
RLM	Epistemic	Q. 7.37	guinea pigs *would* be. like the cop out answer for me
	Boulomaic	Q. 7.38	as much as I *would* like that to be the case

Table 7.8: Would epistemic and boulomaic examples in context from TLM and RLM

The similar uses of *would* across both corpora indicates that it is volume rather than differences in usage that is significant for this term.

Within discourse analysis, the use of deontic modals have been examined in relation to the construction of power (Collins, 2009, p.22). In section 2.4 I discussed that the Power Threat Meaning Framework (PTMF) (Johnstone and Boyle, 2018) is one of the proposed alternatives to a formal diagnosis of depression. This framework involves an examination of how power is enacted upon the individual, what power the individual has and how much agency the individual has to affect the power balance (Johnstone and Boyle, 2018). I therefore explore the use of deontic modal verbs through a lens of power and agency, aligned with this framework.

The deontic modal verbs from my TLM and RLM corpora are *can*, *can't*, *shall*, *should*, *may* and *must*. Each of these words has the capacity to be deontic, though they are not always used in this way. Close analysis shows that *can't*, *shall* and *may* do not have deontic meaning in my corpora. For example, the one use of *may* in TLM is: *it may never impact upon my life*, which is epistemic since the informant is talking of possibility. *Can*, *should* and *must* are all used in their deontic sense, though *must* is only used in RLM, by informants talking about duty towards personal and natural laws. In TLM, *can* is used in both an epistemic and deontic context, whilst *should* is exclusively used in a deontic sense. I therefore explore the collocates of *should* across both corpora to explore how power and agency are expressed differently by TLM and RLM informants.

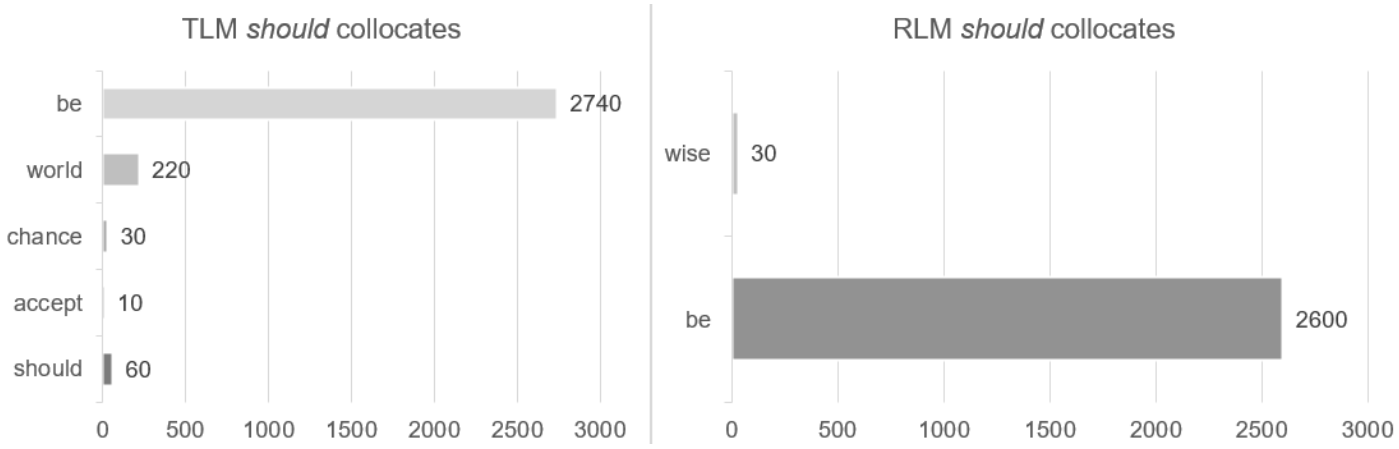

Figure 7.5: Collocates of *should* in TLM and RLM (normalised frequencies)

The most frequent collocate of *should* from TLM and RLM is *be*, and whilst there are more collocates from the TLM data, one of them is a repetition of the word itself. *Should* is collocated with itself in the TLM corpus when informants express a lack of confidence in their ability to perform their duty, as in: *I* ***should****. I* ***should*** *accept*. Since *be* is the top collocate for *should* across both corpora, I examined the use between the two. In TLM, the two terms together indicate informants feeling a sense of injustice, as in: *I don't think the world* ***should*** *always* ***be*** *uh quite as hard as it is*. In RLM, the two terms together can also indicate a sense of injustice, as in this quote from a RLM informant: *that's not what equality* ***should be*** *about;* or positivity as in this RLM quote: *everything* ***should be*** *excitement and a new discovery*. Whilst there are similarities in the use of *should* across TLM and RLM, the difference is that RLM informants are 5% more likely to treat this dutiful term with positive valence.

Through this section I explored how the informants from TLM are statistically more likely to use modal verbs when compared to RLM. Despite the higher proportion of the PoS tag *modal verbs*, there is less lexical diversity within the TLM corpus when compared to RLM. The RLM and TLM informants express their power and agency through the use of deontic modals in a similar way, though with a slightly more positive attitude in RLM.

7.7 Part of Speech tagging analysis discussion

From this inductive examination of the data, I have built upon existing studies, reinforced the two main findings of my last chapter and explored new areas of interest. In terms of building on existing studies, Bucar et al. (2021) found a greater incidence of personal pronouns and verbs in their online depression data set. I found a greater incidence of the objective first person singular personal pronoun *me*, and singular reflexive personal pronouns in TLM compared to RLM. Further to this, I found an increased use of the base form of the verb *be* and auxiliary modal verbs in the TLM corpus when compared to RLM. These two findings are similar to those of Bucar et al. (2021), which indicates that my low mood data set is sufficiently comparable to a depression data set without being directly associated with the term *depression*.

A difference in lexical density and diversity was indicated for the grouped data in section 6.4, based on quantitative measures. Reduced lexical diversity in the TLM informants compared with RLM was indicated by the negative PoS tag *interjections*, the *prepositions*, the use of *general adjectives* and the use of *modal verbs*. The PoS tagging analysis of this chapter has therefore explored how this aspect of the low mood psycholect also manifests within individual part of speech tags.

My exploration of personal pronouns corroborates existing linguistic and psychological studies, as meta-analysed by Edwards and Holtzman (2017). The TLM informants were more likely to use first person personal pronouns than second and third person. Further to this, the way that personal pronouns are used shows the TLM informants talking in generalities whilst the RLM informants are more likely to talk about specific individuals. This further builds on the analysis of section 6.3, evidencing more generalised speech as a feature of the low mood psycholect, when compared to the reference group.

This chapter and the one preceding it have provided an exploration of my first research question: *What linguistic markers are prevalent in the verbal language of people experiencing low mood?* The indications are that these markers include an increased use of personal pronouns; reduced lexical diversity and a tendency towards talking about generalities rather than specifics when compared to

the reference group. Having explored this first research question, the next chapter will be an opportunity to build on this by moving on to the second research question *Do people with low mood focus on the past?* As part of this, I use some of the PoS tagging analysis from this chapter to examine the relationship between low mood, tense and temporal focus.

Chapter 8 Language and temporal focus

8.1 Introduction

In this section I explore my second research question: *Do people with low mood focus on the past?* I initially examine this using the Constituent Likelihood Automatic Word-tagging System (CLAWS) Part of Speech (PoS) tagger, which is part of *Wmatrix* (Rayson, 2009) to analyse the pilot and main study data. I then use the tense tracking diagram (TTD) system devised by Svalberg (2018) to enable replicable close reading of each data set. This was specifically to examine the temporal focus of the data, i.e. whether the informants concentrated on past, present or future in their linguistic choices. The pilot study results indicate that students who could be diagnosable with depression are more likely to have a past focus. There was no clear pattern of temporal focus based on these two measures from the main study data. Finally, I use psychological attribution theory (Stratton, 1988) as a lens through which to consider temporal focus. Through this I find that the TTD and CLAWS trackers could not identify the full spectrum of the time-based focus of the data. Having discovered this gap in frameworks for coding temporal focus, my solution is the subjective method of coding temporal focus for each individual clause.

8.2 Main study CLAWS tense based analysis

I used *Wmatrix* (Rayson, 2009) to compare the automatically generated CLAWS PoS tags from the TLM (low mood) with those of the RLM (reference) group. Table 7.1 in section 7.2 (the previous chapter) shows the results of this comparison, filtered by *Wmatrix* to only show PoS tags that appear more than 6 times and that have at least 95% significance. For this section, I removed these filters and examined both positive and negative key PoS tags. Please note that the PoS tags are those that have been assigned by CLAWS, and therefore do not take into account the nuance of uses for words such as *been*. CLAWS has designated *been* as a past tense verb, and whilst it is the past participle of *to be*, it does not necessarily indicate past tense. As an example sentence: *I have been drinking my tea and daydreaming, have I been ignoring you*? In this case, *been* is in the present perfect progressive (Biber et al., 2010, p.157). There are a higher number of some categories of

past tense PoS tags in the TLM corpus when compared to RLM. There also exists a higher number of different past tense PoS tags in RLM group, when compared to TLM. The tense based positive and negative key part of speech tags and their ranking are outlined in Table 8.1. The keyness rank refers to how characteristic or 'key' CLAWS ranked them as, with negative numbers referring to negative keywords.

Keyness rank	**Part of Speech tag**	TLM freq (per 1k)	RLM freq (per 1k)
5	Verb (present tense) {is}	37.4	32.4
23	Verb (past tense of do) {did}	1.5	0.9
27	Verb (present tense singular/plural) {are}	11.1	10.0
49	Verb (past tense) {were}	0.8	0.7
55	Verb (present tense 3rd person singular) {does}	1.4	1.3
-3	Verb (past tense) {been}	0.1	1.7
-4	Verb/noun (present tense) {doing}	0.3	1.8
-6	Verb (past tense) (lexical) {e.g. gave, worked)	4.3	7.4
-12	Verb (past tense) {had}	0.5	1.6

Table 8.1 Tense based positive and negative key PoS, normalised

The 5th most likely PoS tag from the TLM corpus when compared with the RLM is the present tense verb *is*, and that the 23rd most significant PoS tag is the past tense term *did*. The 27th, 49th and 55th key PoS tags are the present tense *are*, past tense *were* and present tense *does* respectively. These positive PoS tags are a combination of present and past, with no clear skew towards the past or present. From the negative key PoS tags, we can see the third PoS tag least likely to appear in the low mood group when considered against the comparison group is the past tense *been*. The 4th 6th and 12th negative key PoS tags are the present tense *doing*, past tense lexical verbs such as *gave* or *worked* and past tense *had* respectively. Some of the past tense negative key PoS tags are more key than the positive past tense PoS tags, but overall the information in this table gives no clear indication of whether temporal focus differs across corpora.

To examine these results more closely, the 23rd most significantly prevalent PoS tag in the low mood corpus when considered against the comparison corpus is the verb *did*, which is the past tense of *do*. This is the first example in the key PoS tag list of a past tense based PoS tags from the comparison of the PoS tags. I provide two examples of concordance lines of this term from the low mood group (TLM), quotes 8.1 and 8.2 as well as two examples for the comparison group (RLM), quotes 8.3 and 8.4.

TLM quote 8.1: an exercise I did. in psychotherapy [@laughs] training

TLM quote 8.2: white western British woman everything I did was weird I could be the most normal British.

RLM quote 8.3: Gert [pause] did you want me to elaborate on that or just

RLM quote 8.4: people… because it just be if everyone did that it would just be great world I think so

In TLM quote 8.1, the low mood informant is talking about their previous experience in the training that they were doing to become a therapist, and clearly referencing events in the past. In TLM quote 8.2 the low mood informant refers to their past experience of living in another country. RLM quote 8.3 is from a reference informant who uses the word *did* in an interrogative present tense form, which is part of their dialect: that is, what they mean is *would you like*. In RLM quote 8.4, the term is being used in the continuous sense- if everyone had done so in the past, continues to do, and will do so in the future.

Based on the reading in section 2.8 of the literature review, we understand that tense alone is not necessarily indicative of temporal focus (Halliday & Matthiessen, 2013, p187). The lack of connection between tense and temporal focus has particularly been highlighted in the Himmelstein et al. (2018) study. They found the narratives of people experiencing depression were significantly more likely to feature the use of the present tense when talking about past memories when compared with people who were not experiencing depression. My pilot study findings corroborated this, with the informants categorised as depressed (TPD) using the present tense when talking about the past. With no specific trends immediately apparent and no clear indicators of temporal focus based purely on this PoS tagging analysis of the main study data, it is necessary to look at some of the texts more closely.

8.3 Pilot CLAWS tense based analysis

Since my pilot study data set was small in size, I reasoned that this would provide an opportunity to refine my choice of methodological options. This is in addition to the refinements to data generation techniques discussed in section 3.4. I performed the pilot data analysis prior to the analysis for the main study, but am deliberately writing about the results of the pilot data analysis after the main

study analysis. The lack of chronological order for this section is so that the pilot data can provide context and potential comparison points for the true focus of this thesis, which is the main study.

During this pilot examination of temporal focus, I took the approach that I would later replicate for the main study. This approach involved comparing PoS tags for the pilot depression corpus (TPD) against the pilot reference (RPD) corpus which comprises the responses from informants not diagnosable as depressed. Based on tense alone from the PoS tags, there is a greater proportion of tags focused on the past tense in the language of my TPD informants. These PoS tags are specifically the past tense and past participle form of a lexical verb, which tend to refer to actions performed in the past (Carter & McCarthy, 2010, p.395). The other tags were around the past form of the verb *be* which refers to a state of existence (although, as discussed for the main study data, this may be an oversimplified classification by CLAWS). The past tense forms of the lexical verb from this corpus include *said*, *grew* and *thought*, and most examples of this PoS tag were based on anecdotes or feelings about their personal past as exemplified by quotes 8.5 to 8.7:

Quote 8.5: it's like. the main things of what I've said Uhm. but obviously. I don't know.

Quote 8.6: like the comfort of being where like I grew up and. also not being in a student

Quote 8.7: as like which is dark comedy and I just thought this doesn't make me feel comfortable

In quote 8.5, the informant is referring to their immediate personal past, i.e. the things that they have just spoken of as part of the interview. It may be noted that this is the past participle form of *said*, rather than past tense. This is as a result of a misclassification by CLAWS due to the graphological equivalence of past tense and past participles. Quote 8.6 is from an informant who is nostalgic for the home and comfort of their childhood, therefore a more historical personal past. In quote 8.7, the informant relates an anecdote from the last year, adding their narratorial voice to the more recent personal past. Each of these examples have a temporal focus prior to the point of speech, and therefore based in the past.

Another PoS tag more prevalent in the data from TPD informants was the past participle form of the lexical verb. Examples of this include *given*, *wanted* and *made*. An example of each of these words

in context in TPD is in quotes 8.8 to 8.10 , and once again these are used when relating past events or atmospheres.

Quote 8.8: because when you're younger you're given like you know, ten pounds a month
Quote 8.9: study something you know I've always wanted to study. You know and I like. I like
Quote 8.10: like I can apply that to friends I've made through the years. I've talked. I'm still

In TPD quote 8.8, the informant talks about pocket money in their youth as if the past is the present. This is congruent with findings that people with depression are more likely to use present focused language when talking about the past (Himmelstein et al., 2018, p. e0207814). Quote 8.9 is an informant talking about the past up to the present, that their desire for study has been a constant. In quote 8.10, the informant is talking about present application of a concept to friendships from the past. Each of these examples shows a general reference to the past up to this point with no specific time frame. From the literature review as well as direct examples from the corpus, we know that tense does not necessarily indicate temporal focus (Halliday & Matthiessen, 2013, p187).

This pilot study CLAWS based examination did show indications of a past based temporal focus for TPD informants purely from an analysis of tense PoS tags. This prompted me to use the same method to analyse my main study data as outlined in section 5.2.2. This was not a time consuming process and so the decision to use CLAWS to examine temporal focus for both data sets enabled me to search for initial similarities between the main and pilot study texts. We can conclude based on the results that there are no shared trends in tense between the main and pilot study texts. Furthermore, a CLAWS tense based analysis of the main and pilot study data does not provide sufficient clear indication of the temporal focus of these corpora. Further exploration of both data sets using a tense tracking diagram could be more beneficial in finding temporal focus than CLAWS.

8.4 Pilot study tense tracking

Concordances from past-tense PoS tags indicated by CLAWS provided an idea of how tense was used. However, the research question I am endeavouring to address in this chapter is *Do people with low mood focus on the past?*, i.e. looking at temporal focus rather than tense. Hence, I

attempted a closer reading using a different method. It was particularly helpful to focus on this smaller pilot data at first since this could help me establish potential methods for the larger main study dataset.

I annotated the pilot study transcriptions from both the pilot depressed group (TPD) and the pilot reference group (RPD) by hand. The annotations I used were based on the tense tracking diagram developed by Svalberg (2018) and discussed in the literature review section 2.8. This involves separating tense and temporal focus, For example, *I am going to the shops tomorrow* is present tense, but the focus is the future. For this, I used the following notations:

PB (Past Before)	CB (Current Before)
PN (Past Now)	CN (Current Now)
PA (Past Afterwards)	CA (Current After)

Table 2.3 Tense tracking notation (repeated from section 2.8)

Each tense-based word or clause was first categorised by past (P) or present (C) tense, then looked at in context for temporal focus- Before (B), Now (N) or After (A). To give an example of each type directly from the pilot corpus:

PB	She used to babysit me
PN	Those would be the top two things
PA	If it went wrong
CB	We both come from a really good knowledge of pop culture
CN	It is important you know to have good health
CA	So, how successful, I become in the future

Table 8.2 TTD notation examples from target corpus

This process was time consuming, and whilst this level of simplification is useful as a pedagogical tool, I sometimes found difficulties allocating for tense or for temporal focus. For example, in the event of an infinitive or conditional tense the focus could be past, present and future concurrently (also known as universal, as I discuss in 8.6). An example from the target group of where allocating tags from the TTD was difficult is in: *we're living in and even as we grew older*. The distinction between tenses and temporal foci was also different in the language of informants who did not learn English immersively before the age of 5 years, since they were more likely to match tense and temporal focus. In English as a foreign language pedagogy, tense is considered particularly difficult to teach (Svalberg, 2019, p.417). The formal acquisition of English language therefore results in a

different usage of tense, compared to the more intuitive use that is exemplified by language acquired in a colloquial context.

I counted the hand coded TTD annotations and calculated them as a percentage of the total number of categorised clauses for each informant individually. The results follow in Table 8.3, including the total number of annotations in order to provide an idea of the scale of analysis performed.

	Informant	PB	PN	PA	CB	CN	CA	Total annotations
TPD	PI14	12%	12%	11%	23%	36%	6%	94
	PI29	19%	4%	4%	8%	58%	8%	26
	PI44	15%	0%	5%	20%	58%	3%	40
RPD	PI01	9%	17%	35%	17%	22%	0%	23
	PI03	3%	7%	13%	7%	70%	0%	30
	PI20	7%	5%	2%	21%	52%	14%	44
	PI20	7%	5%	2%	21%	52%	14%	44
	PI39	17%	8%	17%	17%	29%	13%	24
	PI39	17%	8%	17%	17%	29%	13%	24
	PI41	13%	3%	5%	0%	63%	16%	38
	PI46	6%	12%	12%	12%	35%	24%	17
	PI47	0%	0%	13%	0%	53%	33%	15
	PI49	0%	4%	9%	30%	52%	4%	23

Table 8.3 Pilot temporal focus tags percentages by informant

The first three pilot study informants are those who were categorised as potentially depressed (TPD). The rest of the pilot study informants are the reference group (RPD), and the letters at the top are for the TTD annotation codes, with the total number of annotations per informant on the right hand side. This table shows that the number of annotations varied from 15 up to 94, based on both the number of words in the transcript and the number of changes of temporal focus and tense. To clarify, if an informant were to change tense and temporal focus multiple times within the same sentence, they would have a greater number of annotations than one who maintained the same focus and tense throughout a sentence of the same length.

From the annotations, most informants use each of the tenses and temporal foci. From RPD, PI01 and PI03 did not use ‘current after’ which means use of the present tense whilst talking about a point in the future. On the other hand, RPD informants PI41 and PI47 did not talk about ‘current before’ which means use of the present tense whilst talking about a point in the past (historic present). RPD informant PI47 was quite distinct, in addition to not using ‘current before’, they also did not use ‘past

before' or 'past now'. To clarify, P147 did not use the past tense for talking about the past or present. This may be due to their short responses to the prompt questions. In fact, they were the least loquacious of the pilot study informants, hence the least opportunity to change tense and focus. PI49 also did not use 'past before', but beyond this appeared to be similar to the rest of the responses. One of the informants from the TPD group, PI44, did not talk about past now, but the language of the TPD informants seemed to otherwise be approximately similar to the language of the RPD informants in the breakdown of their temporal focus and tense use, based on this annotation.

Since there was no obvious pattern from this presentation of the pilot data, I used these annotations in a different way. The following data is the same information combined by mood and separated according to tense and temporal focus:

Tense	Past (P)	Present (C)
TPD	27%	73%
RPD	27%	73%

Focus	Past (B)	Current (N)	Future (A)
TPD	32%	56%	12%
RPD	20%	54%	26%

Table 8.4 Pilot tense and temporal focus breakdown

Based on this rudimentary analysis of the data, the tense balance of the target (TPD) and reference (RPD) groups are equally balanced in terms of the present tense. This confirms the information from the literature review section 2.3, as well as the CLAWS analyses in section 8.3: i.e. that any distinction in temporal focus is based on more than the tense itself (Halliday & Matthiessen, 2013, p187). The temporal focus of the target and control informants is mostly the present, with the TPD informants having a greater likelihood of past focus than the RPD informants, whereby TPD informants had 32% of a past focus compared to 20% for the RPD informants. TPD informants were less likely to have a future focus than the RPD informants, at a ratio of 12% for TPD compared to 26% for RPD. Previous studies found that people experiencing the symptoms of depression use more present focused language when talking about past events (Himmelstein et al., 2018), which the results of this pilot study corroborate. Since this tense tracking diagram (TTD) provided an

indication of temporal focus, I decided to use the same TTD for the main study data, with amendments to the methodology as detailed at the beginning of this section for the larger data set.

8.5 Main study tense tracking

In order to enable a close reading of a representative sample of a much larger data set, I chose to use *ProtoAnt* to find the most prototypical texts. Please see methodology section 5.2.3 for further information on how I completed this process. A list of the texts in order of prototypicality enabled me to use the TTD to analyse the first nine using the same methodology as described in the literature review section 2.8 and detailed at the start of section 8.4.

In order to address my second research question: *Do people with low mood focus on the past?* I was interested in the use of past or present tense and also the temporal foci. Taking guidance from the TTD designed by Svalberg (2018) as a pedagogical tool, I separated out tense and temporal focus, as I did in the analysis of my pilot study. This provided the annotations based on tense followed by temporal focus. The tense could be past or present, and the temporal focus could be past, present or future. I used the following annotations: past before (PB); past now (PN); past after (PA); current before (CB); current now (CN) and current after (CA). I hand coded the data from each informant then calculated the percentage of the data that each code featured. *ProtoAnt* ordered all of my informants according to representativeness, and after grouping according to the TLM (low mood) and RLM (reference) corpora, the results of the hand coding are displayed in this representative order in Table 8.5:

	Informant	PB	PN	PA	CB	CN	CA
TLM	P1091					100%	
	PI175	17%			33%	50%	
	PI171		27%			73%	
	PI161	6%				94%	
	PI044				15%	85%	
RLM	PI114		10%			90%	
	PI110				14%	86%	
	PI028				4%	96%	
	PI131	5%	9%	5%	5%	77%	

Table 8.5 TTD for main study prototypical texts (% of annotations per category by informant)

There was a lot of variety of tense and temporal focus in the pilot data (Table 8.3) and less variety in this main study data (Table 8.5), potentially because my research question was designed to prevent temporal priming. Specifically, I tried to avoid asking my informants to think to the past or look forward to the future by asking them about what was important to them at the time of the interview. If I had asked about favourite memories or future plans, this would have been temporally priming. However, some of the follow-up questions I removed between pilot and main study could have been temporally leading. An example of this is: *does importance change to you?* By asking if importance had changed, it encourages the informant to think about what was formerly important. As an informant thinks back to the past, they are temporally primed to think about the past and therefore their speech patterns are affected.

Cutting potentially temporally leading questions might explain the preference in the main study data for 'current now' focus. That is, each informant's linguistic data is at least 50% 'current now', with TLM informant PI91 using exclusively this tense and temporal focus. Otherwise, there is no initial clear pattern based on this analysis, particularly in relation to the low mood group. In future studies, an interview question ideally needs to enable a focus on the past or the present, whilst retaining the lack of priming towards this.

Similar to the comparison performed on the pilot study data, I therefore chose to examine the same analysis from a different perspective. Hence, I considered past and present tense separately, as well as the *before*; *now* and *after* temporal foci separately based on an amalgamated figure for the low mood and comparison prototypical texts. The results can be seen in Table 8.6:

	Tense		**Temporal focus**		
	Past (P)	Present (C)	Past (B)	Present (N)	Future (A)
TLM	10%	90%	14%	86%	0%
RLM	7%	93%	7%	92%	1%

Table 8.6 Breakdown of TLM and RLM temporal focus based on TTD

As in the pilot study, the proportion of past to present tense in both corpora is highly similar (c. 90% to 10%), indicating no significant tense based distinction between the verbal utterances of those with low mood and those without. Once again in line with the pilot study, the temporal focus for both

groups is firmly in the present, with no future focus at all from the low mood group and very little from the comparison group. I began with the initial hypothesis that people experiencing low mood were more likely to focus on the past, and the low mood group (TLM) were more likely to focus on the past than the reference group (RLM). Looking at Table 8.6, this past focus is not the case for all of the prototypical TLM informants.

The use of this tense tracking diagram on prototypical main study texts provided an opportunity to see if there are indications of a past focus for those with low mood when compared to those who are not experiencing low mood. Whilst in general this was the case, there are exceptions to the temporal focus of the low mood informants and so this is not a pattern. From a methodological perspective, the time consuming process of looking at temporal focus itself was expedited with the use of prototypical texts, though TTDs are too time consuming for use in analysing large corpora.

8.6 Attribution theory and temporal focus

The results from the pilot study tense tracking appeared to show that the informants potentially experiencing depression were more likely to look to the past when compared to the pilot reference group. However, the hand coding from the prototypical main study texts did not show a similar temporal focus for the low mood informants when compared to the main study prototypical reference informants. This is likely due to the difference in questions, whereby my decision to avoid temporal priming resulted in a lack of opportunity for informants to consider a past or present temporal focus.

In direct response to this lack of clear temporal focus, I decided to take a different analytic approach. As part of methodology testing, I hand-coded the first response provided by every informant to the main interview question: *When you think about what is most important to you, what comes to mind?* I used Leeds attributional coding system (LACS) (Stratton, 1988) for this coding, please see section 2.4 and 5.3.1 for further information on attribution theory and its context.

Attribution theory involves the examination of how informants view impactful events, individuals and concepts. In this case, I am examining how the informants view the first thing that they described as important to them. In LACS, data is coded according to binary opposites, the informant may attribute

an important thing as stable or unstable; specific or global; internal or external; personal or universal; controllable or outside of the bounds of control (Stratton, 1988, pp.29-34).

To provide an example of LACS in practice using the main study data, I will discuss how I coded one concept. For some informants, the first response to my interview question was *family*. If an informant told me about how important family was, I read through their full response to this question to consider the context for *family* and coded accordingly. In the case of the concept of *family*, this could potentially be coded as stable, specific, internal, universal and uncontrollable. For example, one informant used the term *family* and was referring to their child exclusively: *because biologically I'm hormonally linked […] because caveman brains are clever*. In this case, they described their relationship with their child as stable, and their love in terms of the specifics of their bond. The relationship with the child was seen as coming from an internal rather than external source. Further to this, the informant spoke about the love of a parent for a child in terms of universal rather than personal sentiment, and a resultant lack of control over this feeling: *caveman brains*. As a contrast, another informant spoke about *family* in terms of generational storytellers, taking a much wider chronological scope for the term *family*.

A practical reason for analysing only the response to the first main interviewer question for the purposes of attribution theory was efficiency of analysis within the scope of the project. However, psychological theory also affirms this analytical focus, in accordance with the first instinct fallacy. When answering questions, our initial ideas for an answer are the ones we are most likely to use (Kruger et al., 2005). I could find no studies affirming or contradicting that a first response is the most important, but it does last the longest. Moreover, every informant provided a first response to the question and therefore length and scope of replies have not skewed this analysis.

In practice, the responses did not all easily lend themselves to attributional analysis. This is because not every informant spoke or reasoned in a clearly codable way. For example, the entirety of what TLM informant 175 said about family follows:

> Quote 8.12: Family. Uhm I guess for me because it's the only guaranteed thing you'll ever really have in life

In the case of informant 175, this short summary was difficult to code as stable or unstable; specific or global; internal or external; personal or universal; as well as controllable. I discuss precise details of analysing for attribution theory in section 9.3.

I persevered with attributional coding and through this saw how simplistic my approach to temporal focus had been by purely considering past, present and future. A simple approach is arguably necessary for second language pedagogy, as with the tense tracking diagram (Svalberg, 2019). However, my research question requires an examination of the nuances of temporal focus. In attributional coding (Stratton, 1988), ascribing features such as *universal* and *stable* is necessary. Temporal focus is not necessarily purely based on the past, the present or the future, the informant could be focusing on all three. Informants can consider a present situation to have constantly been the case in the past and that the same situation will always be such in the future. This universal and stable way of conceptualising is something that attribution theorists believe applies to the situational experience of depression (Sweeney et al., 1986, P.974). This seemed to fit with my recollections of temporal anecdotal experience of low mood. Hence, attribution theory appears a promising approach to temporal focus which might evidence generalisable patterns in relation to my second research question: i.e. *Do people with low mood focus more on the past?*

To bring this theoretical underpinning back to the practicalities of data analysis, I decided to look at temporal focus in terms of past, present, future and universal. I began by looking at the temporal focus for the first response only, to see if there was a clear pattern and therefore if there was merit in this line of enquiry. I re-coded all of these first responses based on the four parameters of past, present, future and universal temporal focus.

An example of a universal temporal focus from TLM informant 91:

> TLM quote 8.12: close bonds. between family and friends that I **know** that I **can** rely on. and the loyalty that **comes** from that.

In a similar vein, TLM informant 161 says:

> TLM quote 8.13: relationships. uhm with other people so. whether that **be** family whether that **be** friendships whether that **be** romantic uhm the relationships in my life **are**. the most important thing.

In quotes 8.12 and 8.13 I have put the present tense words that led to my coding as universal temporal focus in bold. From quote 8.12, *loyalty that comes from that* is present perfect progressive, whereby this loyalty from close bonds has been, is and continues to be. In quote 8.13, the infinitive is repeatedly used for *be*, since the relationships referred to are considered to have existed, to currently exist, and it is assumed that they will continue to exist. Both of the TLM informants in quotes 8.12 and 8.13 talk about the importance of general relationships, with no reference to specific individuals. Both informants also consider relationships or connections to be constantly important. Another TLM informant, 171, is more introspective than the previous two informants, as they say that of greatest importance to them:

Quote 8.14: I think to be. content in oneself to be able to look in the mirror and. be happy.

Informants 91, 161 and 171 in quotes 8.12, 8.13 and 8.14 respectively provide examples of people experiencing low mood talking about importance not purely in the past, the present or the future but universally. As a contrast to these TLM examples, RLM informant 118 talking about the present rather than universal:

Quote 8.15: for me the most important thing. **is** the people. that .I **interact** with in my life

In quote 8.15, the verb *is* has a progressive but not a perfect aspect, indicating existence up to but no clear indication beyond now. Further to this, interact is not an infinitive, it is an intransitive verb used in this context to indicate present tense only with no connection to past or future temporal focus.

Based on my coding of all first responses, the following pattern emerged:

	TLM	RLM
Past	8%	8%
Present	17%	73%
Future	0%	3%
Universal	75%	18%

Table 8.7 Temporal focus for first response in main study corpora

I performed this test to establish if this possible pattern for temporal focus was congruent across my low mood data. This analysis indicates a definite preference for a universal temporal focus for the language of my low mood informants when compared with the language of the reference informants.

8.7 Language and temporal focus discussion

The TLM informants did not have a strong temporal focus on the past when compared to the RLM informants. I first used PoS tagging to examine general trends in tenses used. This provided no clear indication of temporal focus, although this is because tense and temporal focus are not direct corollaries (Halliday & Matthiessen, 2013, p187). I also used a tense tracking diagram (TTD) designed by Svalberg (2018). This diagram enabled me to visualise the tense and temporal focus alongside one another. The results of this examination were that the TLM informants focus on past as well as the present and future in a similar way to the RLM informants.

In an examination of temporal focus, I did find that the TLM informants had a tendency to consider the current situation as the way things always will be. This discovery was based on close reading of the first response to the main interview question from each TLM and RLM informant. To clarify, I looked at how the informant spoke about what was most important to them to examine the temporal focus. Rather than purely looking at *past*, *present* and *future*, I also considered *universal*, based on Stratton's (1968) attributional coding. I found that the temporal focus of low mood informants was more likely to be universal than the reference group of informants. Based on the Himmelstein et al. (2018) study, I had expected more present focused language when talking about past events. A universal focus is different to this, since the informants spoke about events from past, present and future similarly across TLM and RLM. For the TLM informants, constancy was the defining focus, it seems as though, for them, the current situation is how it always has been and always will be.

Chapter 9 Attribution theory meets appraisal theory

9.1 Introduction

In this chapter, I explore the potential for combining the psychological framework of attribution theory with the linguistic framework of appraisal theory. Attribution theory enables systematic analysis of the psychological features that are perceived to have caused emotions (Stratton, 2003). Appraisal theory enables one to systematically examine how situational emotion is linguistically constructed (Martin & White, 2007). Therefore this chapter is testing the concept that the psychological cause and linguistic construction of emotions can be intertwined, and in this scenario, used to reveal more about the presentation of low mood. This fine grained analysis will enable me to further explore my first research question: *What linguistic markers are prevalent in the verbal language of people experiencing low mood?*

In the early days of corpus linguistics, psychology academics combined attributional coding with corpus linguistic analysis, including finding units of speech to analyse using attribution theory (Rae, 1993). Rae used corpus linguistic techniques to find the words *because* and *cos* in spoken transcripts. They considered that these synonyms for *cause* were a logical place to seek how their participants attribute causation and reason, which is the root of attribution theory. Psychologists who have used attributional coding to analyse the speech of people experiencing depression have found recurrent patterns. For example, Sweeny et al. (1986, p,1.) found that depressed patients attributed the causes of negative situations to be internal, stable and global whilst positive events were more likely to be attributed to luck and ability. This difference between negative and positive attributions is something that I explore further in this chapter, for my low mood informants.

Appraisal theory is a tool for language and discourse analysis, influenced and pioneered by research such as that conducted by Martin into the discourse around the “Stolen Generations” in Australia (Martin, 2004, p.270). Research that uses appraisal frameworks appears to be mostly centred around news stories (e.g. reports of Covid by Ye, 2021), literature (e.g. an analysis of *Oliver Twist* by Lian, 2018) or language acquisition and translation (e.g. a comparison of translations of *Mansfield Park* by Alsina et al., 2017). The only research I could find that is directly applicable to my work was

performed by Grundlingh (2018), who found this framework helpful for screening the difference between a genuine and a falsified suicide note. Grundlingh (2018) found that the categories trust, moral ethics and negative appreciation from the judgement and appreciation attitude dimension were more prevalent in the genuine suicide notes. The low mood and depression basis of this study led me to believe that the same framework would be useful in analysing the language patterns of my own informants.

Analysis chapters 6 to 8 featured a top down exploration of my data from a linguistic perspective, then a more detailed analysis of my data. At the end of the fine grained analysis of chapter 8 I introduced attribution theory. The main findings from these analysis chapters are reduced lexical diversity and a focus on the abstract or generalities rather than the concrete and specifics in TLM compared with RLM. TLM informants also have an increased linguistic presentation of self-focus when compared with RLM informants (e.g. increased use of *me*), and a partiality towards a universal temporal focus (rather than past, present or future). I could find no other research into the use of abstract or generalities in low mood or depression based data and therefore this chapter is a deeper investigation into this abstraction.

This chapter begins with a fine grained analysis of my prototypical texts using the psychological framework of attribution theory. Once I have explored the cause of emotion in these prototypical texts, I use the linguistic framework of appraisal theory, to analyse how these emotions are linguistically constructed. For my hand coding, I chose to work on topics of interest, taking inspiration from Grundlingh (2018). They describe these themes of interest as thought units. Grundlingh (2018) advises that each thought unit is defined by the concept an informant communicates rather than punctuation (which is not relevant to transcribed speech). A thought unit can be a few words or longer speech. I went through my prototypical texts to find the concepts that each informant considered important and emotionally charged. This is similar to the topics from sections 4.4 and 6.3, but I only used topics that the informants spoke about enough to be codable. To clarify, if as a topic, an informant just said *friends*, with no data as to why or the emotions around friends, I would

not be able to code this and therefore chose the next topic. Then I annotated the attribution and codes for each topic for every informant in the sample separately.

In order to combine aspects of appraisal and attribution, I will be performing case studies based on the parts of each framework that are most relevant to my research. Moors is a psychologist who wrote about how to combine different theories or frameworks in a truly integrative manner:

> Integration is more than the mere summation of the assumptions of both theories. True integration does not tolerate inconsistencies or redundancies and has the freedom to critically examine some of the assumptions in the original theories. (Moors, 2017, p.1)

The combination of attribution and appraisal will therefore provide an opportunity to use only the most relevant parts of each framework. This will be the parts that specifically enable analysis of language patterns associated with a low mood psycholect based on my analysis up to this point. This is in order to further explore the use of generalities and specifics in the verbal language of the TLM informants when compared to RLM.

9.2 Attribution analysis

Attributional coding comes from the discipline of psychology, and involves considering dimensions that characterise the causes of outcomes such as actions or emotions. I outline the framework and the analytical methods in sections 2.4 and 5.3.1. To explore the answer to my first research question about language patterns associated with low mood, I use attribution theory to understand how informants linguistically characterise the causal mechanisms of importance and their situations or moods.

The dimensions of attributional coding are: reliability, scope of outcome, area of origin, distinctiveness and level of controllability (Stratton, 2003, p.142). Each dimension has specific parameters that are graded relative to the informant, the situation and other persons or items that are involved. Every attribute is coded with the emotions behind it taken into account: there is the option to consider causes as emotionally positive, negative or neutral. There is no distinct linguistic framework for how these causes should be coded, the analyst is expected to use their subjective reasoning.

Reliability refers to how stable or unstable the cause of the emotion is considered to be. Scope of outcome is who the cause of the emotion affects, this could be either a global phenomenon or something specific to the individual experience. The area of origin refers to the source of the emotion: this could be internal or external. A distinctiveness of attribution involves the consideration of whether this situation is personal: that it could only happen to the person experiencing it, or if it's more universal. Lastly the level of controllability, is the situation around the emotion controllable or not. Please see Table 9.1 for a visual summary:

Reliability	stable/ unstable
Scope of outcome	global/specific
Area of origin	internal/ external
Distinctiveness	personal/ universal
Controllability	controllable/ uncontrollable

Table 9.1: Visual summary of Attributional codings (based on Stratton, 2003)

Stratton (2003) writes that part of attributional coding is considering each of the dimensions alongside one another. Looking at each dimension in isolation does not provide the intersection points that can cause distress. However, researchers such as Greer et al. (2004) focus on one dimension that is relevant to their area of interest rather than every dimension. In the case of Greer et al. (2004), they purely examine the *scope of outcome* dimension when considering how their informants describe symptoms (Greer et al., 2004, p.521). Based on the analysis prior to this chapter, like Greer et al. (2004), the dimension relevant to my research is *scope of outcome* which can be *global* or *specific*.

I begin by coding each attribute individually for the prototypical main study informants, then focussed on the most relevant dimension for all informants. The attributional coding for my eight most prototypical informants broken down according to dimension is in figures 9.1-9.5. I took each topic that the informant considered to be important to them and coded it according to attributional dimensions. I then combined the attributional codings according to informant and created figures to show the proportions of codings (calculated as percentages) for each attributional dimension. In these figures, the informants are first grouped according to corpus and then listed in order of prototypicality.

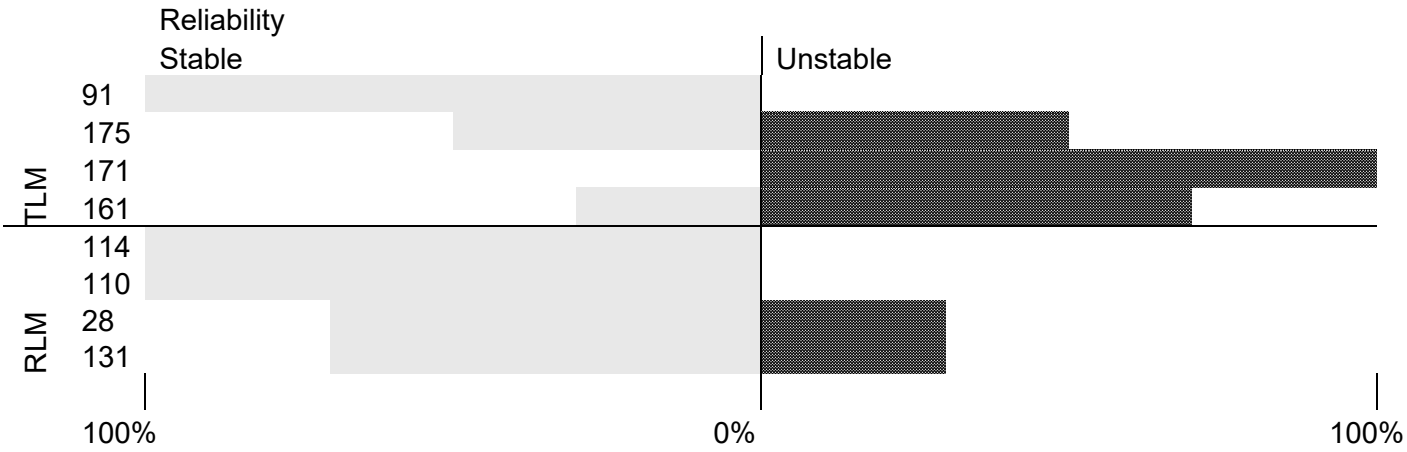

Figure 9.1: Reliability attributional coding for prototypical informants

The RLM cohort are more likely to feel that the things that they find important are reliably stable than the TLM informants. TLM informant 91 believes that all of their important things are stable, whilst the rest feel that there is an element of instability and unreliableness to the things that they consider important. An example of a stable coding is TLM informant 175, who says that their family is:

TLM quote 9.1: the only guaranteed thing you'll ever really have in life… from the moment you come into the world.

An example of unstable coding, also from informant 175 was the way they referred to their friends:

TLM quote 9.2: without them. or without frequent close interactions. it really really took a negative effect.

Within psychology, if the source of positive emotions is considered to be unreliable, this increases the risk of prolonged low mood (Sweeney et al., 1986, p.986).

In figure 9.2, I display the scope of outcome codings for my prototypical informants, once again divided into TLM and RLM and ordered by prototypicality:

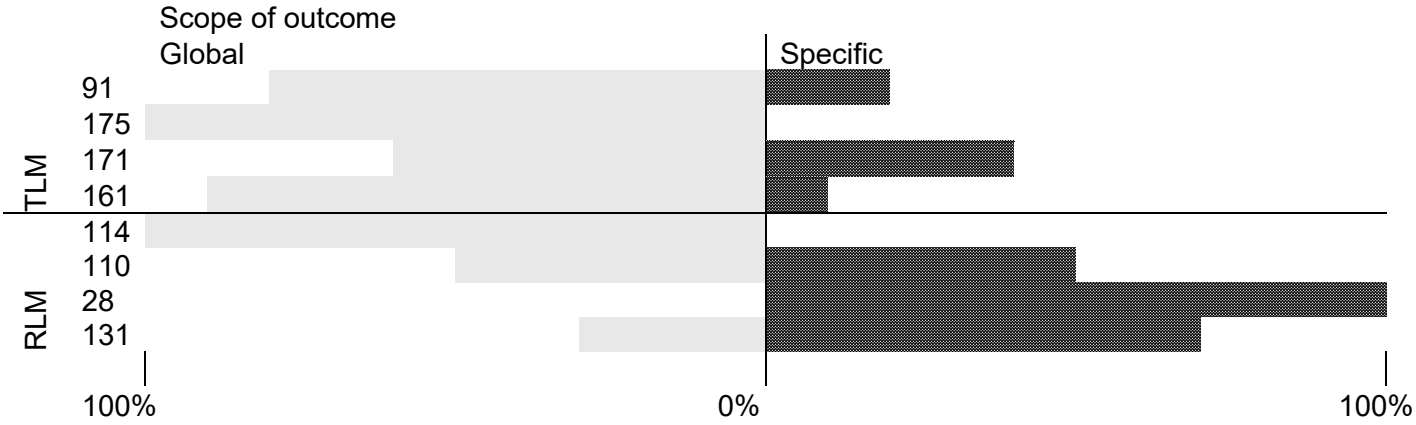

Figure 9.2: Scope of outcome attributional coding for prototypical informants

When considering the scope of the outcome of the important things to my informants, the TLM were more likely to consider the outcome as global than the RLM informants. One RLM informant had outcomes that were entirely global, whilst the rest of the RLM informants considered their outcomes to be specific in at least half of the situations. There was one area of importance that I found difficult to code for this section, and this was informant 91 saying that relationships *add… to the value of someone's life. to. to my life specifically*. When informant one talks about someone's life, this is a global or general thing, but then when they clarify by talking about their life, this changes the scope of the outcome. I therefore chose to code this example as both global and specific. This is one of two instances where I chose to double coding in my sample.

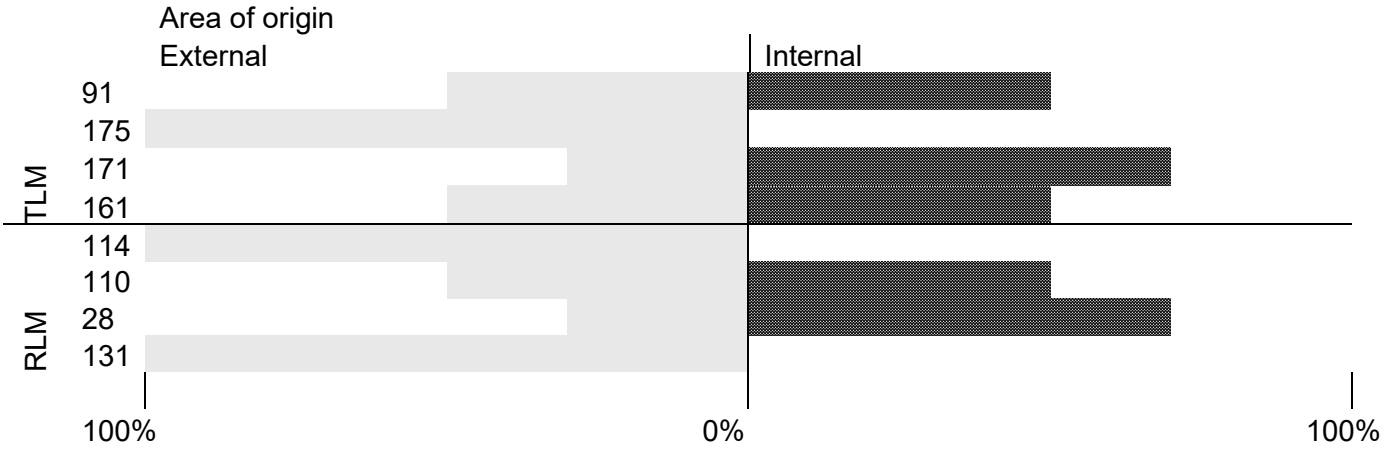

Figure 9.3: Area of origin attributional coding for prototypical informants

When considering the area of origin of their important things, two RLM and one TLM informant considered theirs to be exclusively external. The other three TLM informants and two RLM informants had a fair mixture of internal and externally attributed areas of origin. An example of an externally attributed area of origin from RLM informant 131 when talking about their family:

RLM quote 9.3: that's why I do everything else it's to make sure that. they- well all of us but particularly they can have the life that they deserve

I coded this as external, because they do not explicitly talk about the love they feel for their family, which would be more internal. Whilst *all of us* is inclusive, this is more of an afterthought, whilst *they* is the focus of the deservingness, and is more exclusionary. This exclusionary language indicates that the informant considers the external family, *they*, a motivating factor for their actions.

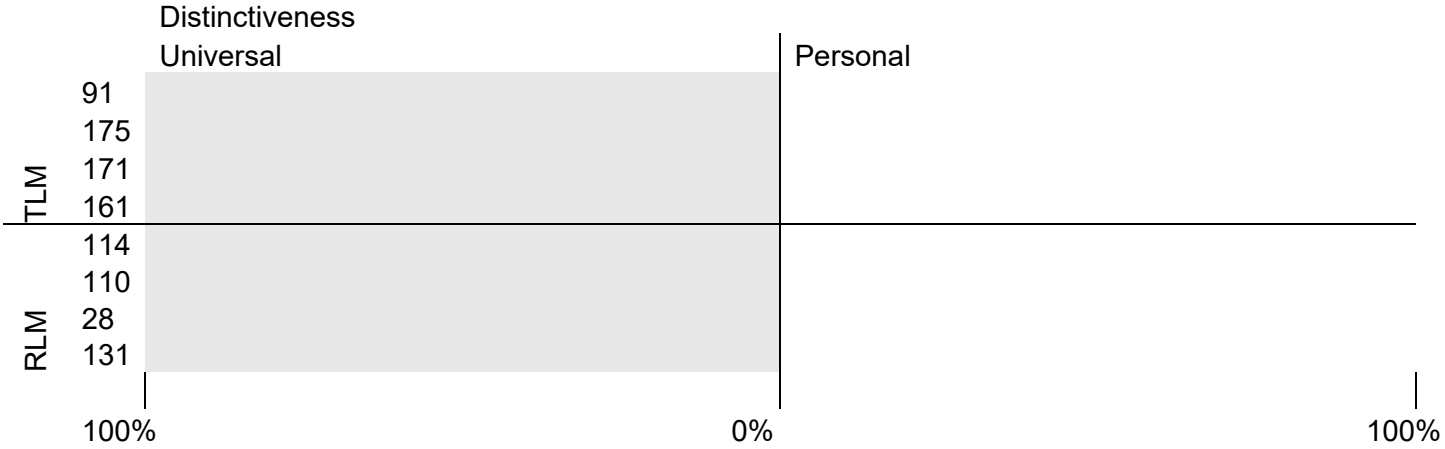

Figure 9.4: Distinctiveness attributional coding for prototypical informants

The areas of importance that the TLM and RLM informants spoke about were all universally rather than personally distinctive. From chapter 4.4, I outlined that the main topics that the TLM and RLM informants spoke about were things like family, friends and health which are all universal ideas. It is therefore logical that the distinctiveness of the important things for the informants would be universal rather than specific.

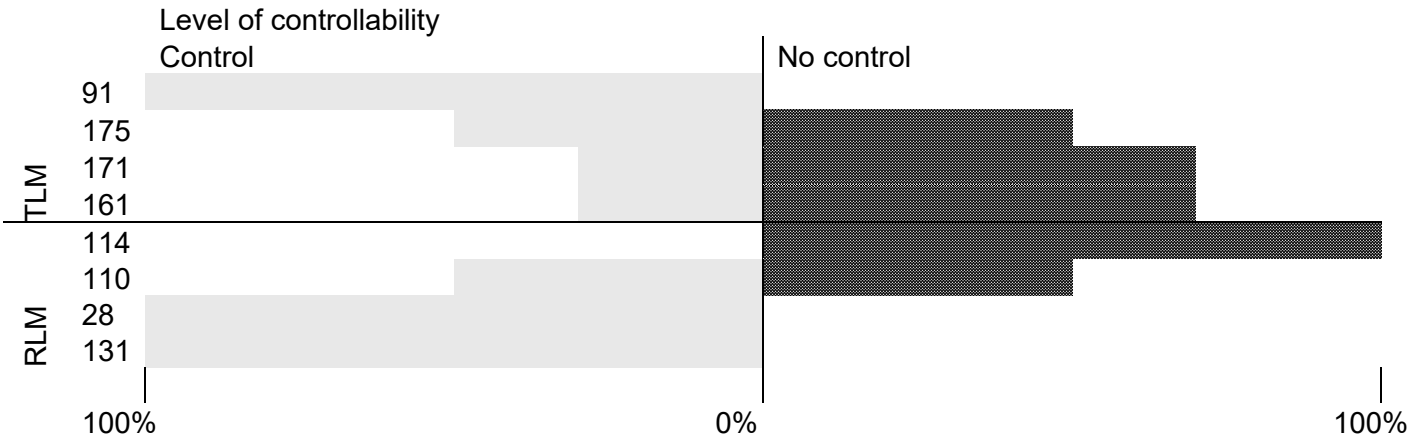

Figure 9.5: Controllability attributional coding for prototypical informants

One of the TLM informants and two of the RLM informants felt that all of their important things were controllable. One of the RLM informants described all of their important things as uncontrollable, whilst the rest of the TLM and RLM informants spoke about varying levels of control. An example of control from RLM informant 28 about the importance of self-care in the context of mental health:

RLM quote 9.4*:* if I'm looking after my mental well-being I tend to find I'm not getting ill as much I'm much healthier

This informant shows that their ability to control their mental health also exerts a measure of control on their physical health. An example of a lack of control from RLM informant 114:

RLM quote 9.5: …if friends and family aren't happy and safe… what's the point in… anyone else being. happy and safe

This informant acknowledges that it is possible for their important friends and family to be unhappy and unsafe, and therefore that they cannot actively control this.

I have used each of the attributional dimensions: reliability, scope of outcome, area of origin, distinctiveness and level of controllability (Stratton, 2003, p.142) to examine how my informants view the things that are important to them. Of these, there were similarities in attributions across the TLM and RLM informants, with the occasional dissimilarity. The clearest distinction came from the scope of outcome attributional dimension- the difference between global and specific. A concentration on global rather than specific things of importance is something that has been characteristic of my data as outlined in sections 6.3, 6.5, 7.4 and 7.5. This is therefore something to explore further in this chapter.

I analysed the scope of outcome attributions for the first concept that every informant from the full TLM and RLM main study corpora spoke about during their interview. Taking the first response to each concept is an approach I adopted in section 8.6 and found it as useful approach for analysing the full breadth and nuance of a sample of data. The results from this analysis are in figure 9.6

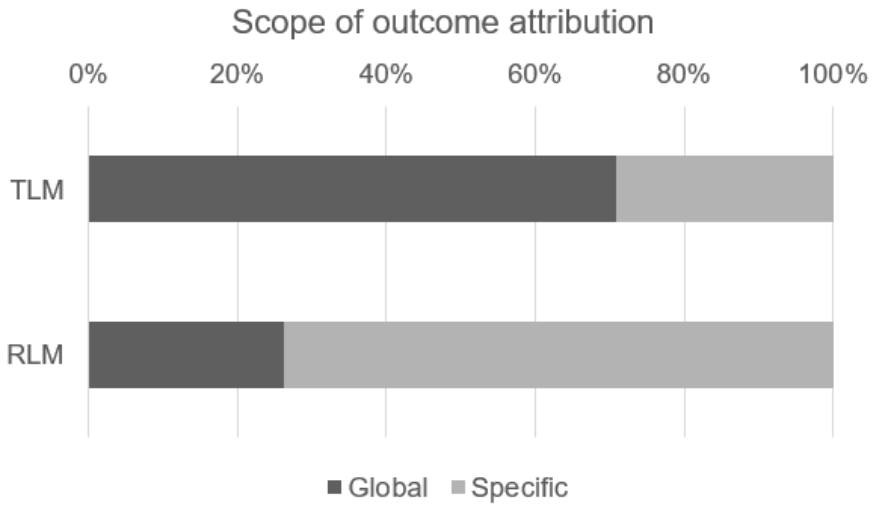

Figure 9.6: Scope of outcome attribution for all informants' first response

This figure shows that TLM informants are more likely to talk about their first topics of importance in terms of a general rather than specific scope of outcome. RLM informants are instead more likely to

talk about their first topics of importance in terms of a specific rather than general scope of outcome. Whilst the difference between TLM and RLM is notable in this example, it is not a common rule across all of the data. Using the scope of outcome attribution to analyse this data has therefore not provided a pattern exclusive to the TLM informants when compared to RLM. However, my previous linguistic analysis indicated that the way the TLM informants spoke about their topics of importance was more general than the RLM informants. This difference appears more notable when analysing language rather than the conceptual attributive scope of outcome. It therefore behoves me to consider this idea of global and specific terminology from a linguistic angle, specifically the framework of appraisal theory.

9.3 Appraisal of attributes

Appraisal theory is a linguistic framework for understanding linguistic construction of emotion; please see section 5.3.2 for a detailed description. To summarise, the constituent parts of appraisal theory are engagement, attitude, graduation and feelings (Martin & Rose, 2003, p.54). *Engagement* can be described as the source of the emotion, either from the perspective of the informant or external factors. *Attitude* refers to the positive or negative impact of the emotion based on affect, judgement or appreciation. The judgement attitude includes morality, ethics, personal capability and convention. *Graduation* is another constituent part of appraisal theory, one that accounts for the intensity and specificity of an emotion (Martin & Rose, 2003, p.54). The final part of appraisal theory is *feelings*, and this involves the classification of the components that constitute linguistic formulation of a feeling. Linguistic expression of a feeling within appraisal theory incudes affection, misery, trust, disquiet and desire.

Positive affect is societally ‘good’ emotions such as happiness, whilst negative affect refers to the corresponding societally ‘bad’ emotions such as anger (Martin & White, 2007, p.46). Feelings may be behavioural, mental or relational as well as based on something specific, or generalities (Martin & White, 2007, p.47). It is also important to grade the intensity of feelings and the level of activity or passivity that the informant experiences around this feeling (Martin & White, 2007, p.48). In terms

of specifics, emotions may be happy or unhappy; based in security or insecurity and around a sense of satisfaction or dissatisfaction (Martin & White, 2007, p.49).

Personal capabilities are how much the informant believes that they have the capacity to influence a situation. Personal normality refers to how unique an occurrence or concept is according to the informant. Focus of emotion is a part of graduation, and in the context of appraisal, focusing terms are used in much the same way as a camera lens, and can be used to soften or sharpen a concept (Martin & White, 2007, p,50).

I have noted in section 6.3 that the TLM informants seem more likely to think about things with negative valence in a specific way and things with positive valence in a general way. From the psychological framework of attribution theory, I will continue to pursue *scope of engagement* though will do so alongside the compatible *focus of emotion* dimension from linguistic appraisal theory. These are particularly compatible for exploring the concept of specifics and generality because the scope of engagement categorises the global or specific, whilst focus of emotion categorises the sharpening or imprecise words used. Alongside these compatible dimensions, I examine the attitudinal *affect* of linguistic appraisal theory, to consider the valence, personal normality and capability expressed through the informants speech. The *affect* dimension will therefore enable a linguistic understanding of how valence and power affect scope and focus of their important item. In my first worked example I will work through each part of *affect*, before using only the sections that are most relevant to the existing data. Specifically, that low mood people use general terminology for concepts with positive valence and specific terminology for concepts with negative valence. I therefore concentrate on the attributional affects of valence and specificity of the concept when working through the data in detail.

In keeping with Rae (1993), I have chosen to search the verbal utterances of my case studies for all instances of the word *because*. This is because “the role of causal expressions, used for general explanation & for assigning blame & praise, is seen as crucial to the discursive apparatus (Rae, 1993, p. 239). As an integral part of discourse, *because* is therefore an ideal place to start in interrogating texts. In order to test the viability of this combination of appraisal and attribution

frameworks, I have chosen to concentrate on one TLM text and one RLM text, beginning with TLM informant 95. TLM informant 95 uses *because* 0.8% of their interview, the most concentrated use of the term amongst the TLM informants. In order to make this comparable, I chose a RLM informant with comparable demographics and a similar interview length and concentrated use of *because*. This comparison is RLM informant 128 who is the same gender, a similar age and similar education level to TLM informant 95, and also uses *because* in 0.8% of their interview, which is the same amount. These results are not representative, but are an exploration of the concept of combining aspects of appraisal theory with an aspect of attribution theory.

9.3.1 TLM case study: informant 95

I coded each contextual instance of the word *because* for informant 95, and have included examples of these uses, as follows:

> Quote 9.6: I... have striven don't know if that's a word for a long long time for safety because I didn't have it for most of my life. and. so that's why that's very important to me and safety to me might mean different things. uhm. to what safety is for other people so safety for me is a space of my own that I can. uhm build a life in and create. in the way that I want it to look and work… and uh spend my time in a way with people and in a place or places. that feel safe to me

From attribution theory, I have coded *safety* as global rather than specific, because informant 95 begins quote 9.6 by talking about how safety as a broad concept is important to them. They progress by talking about the specifics of safety, but the global idea of safety is where they begin. As far as the graduation of focus is concerned, the terms around safety are softened rather than focused (Martin & White, 2007, p.37), including *might mean different things*, *in a way* and *that feel safe to me*. There are no precision or focused terms around this concept of safety. Regarding attitudinal affect, from their wording there has been an absence of safety, which makes this a concept with negative valence. The concept of safety is relational, something considered in relation to other things such as people and places. The informant talks about how safety is based on a general feeling rather than specific individuals or places. The only intensification modifier is the *very* in *very important*, emphasising the importance of safety to the informant. These feelings are active, particularly when the informant talks about building a life. It's not clear from this transcribed section

if the informant is specifically happy or unhappy, however the sense of insecure disquiet is clear, and a sense of dissatisfaction.

Informant 95's next use of the term *because* is not codable: they simply say *I've put myself in a corner because I'm the one who answered the question*. There is no precision or softening term in this quote; the language is neither general nor specific and there is no intensification. However, I was able to code the use of *because* that follows after:

> Quote 9.7: I'm upset partly because some of the things that you already do know about because my safety at work is. perhaps under threat because hours are being cut and I don't know if the museum is… going to survive the economic crisis that the country is currently in… but… it's a safe place in so many ways

In quote 9.7 from informant 95, they are continuing on the topic of safety. The attributional focus for quote 9.7 is specific rather than global, and in terms of appraisal, the informant uses the softening terms *partly* and *perhaps*. Considering the parts of affect that I have chosen to concentrate on, this clearly has negative valence and is based on something specific rather than a general feeling. Quote 9.7 corroborates the patterns I have found in the data so far: that negative emotions in the TLM informants involve more specific terminology and attributional focus than positive emotions.

The next section with *because* follows:

> Quote 9.8: it's a weird safe place and I kind of love that about it because that's fine uhm being weird is fine. finally embraced that and uhm… it. it is somewhere that I can-. it's like a home feeling. with people mostly that feel like home and topics and content for my day. that feels like home like I'm in control

In quote 9.8, there is a more global than specific attributional style: the informant talks about things with a global appeal: *people*, *home*, *topics* and *content*. In terms of appraisal, terms such as *finally embraced that* and *love that about it* indicate positive valence, although the latter is softened with *kind of*. Safety and home are individually named in quote 9.8 and have an attributionally global appeal, but the reference to being *in control* is appraisally specific.

Quote 9.9 is the next section that features the word *because* for informant 95:

> Quote 9.9: things that I haven't had for most of my life and that comes back to the way I was raised the the type of family I was raised in the type of… uhm. socioeconomic environment and… time period because. I was given certain messages from all of those things. that… sent me out into life. looking for. certain things

Much like the previous quotes, 9.9 has a global attributional style, talking about a *type of* family/environment/time rather than specifics. The valence as this section progresses is negative, and whilst there is no explicit negativity, beyond *I haven't had*, pragmatically, *the type of family* and *certain messages* are negative conditioning that rely on the shared understanding between interviewer and interviewee. The feelings in quote 9.9 are once again articulated generally, including terms like *certain things*. The pattern continues:

> Quote 9.10: is safety to. be who I am and express who I am and… exist how I am because… home growing up was not a safe place to do that and when I say safe I mean emotionally safe rather than physically safe

When articulating what safety means to them, the attributional style is global, because a sense of safety in self-expression is something that has global appeal. The negative valence is the lack of emotional safety: *not a safe place to be* and the generality is the concept of emotional safety. Quote 9.10 could have involved references to specific things such as words or actions taken by a named individual, but instead they remain general including *home growing up*. This coding is similar to quote 9.9, as they anaphorically reference the theme of safety.

The rest of the concordance lines for *because* from informant 95 follow the same pattern, with one exception:

> Quote 9.11: when I left my marriage I went to Bleysal… which was accidentally marvellous because. when you're in Bleysal. everything's different everything's different it's like you've gone to Mars so. as a white Western British woman… everything I did was weird

Quote 9.11 is the first point at which informant 95 changes towards positive valence, signified by the word *marvellous*. They name the specific place associated with this positivity: *Bleysal* and extol its difference, which is contextually positive. As part of my ethics protocols I changed the name of the place they are referring to, but it is clear that they reference Bleysal twice. They detail that the situation is leaving a marriage, the place is Bleysal as well as their identity as a white, western British woman. In both appraisal and attribution, their experience is more specific than global. This is an example of positive valence, and that this can be associated with specifics for this TLM informant.

9.3.2 RLM case study: informant 128

In this section, as with the previous one, I have coded each use of the word *because* for one informant for their appraisal of attributes. Specifically, I have analysed the concordance lines for the word *because* to see if they are attributed as global or specific; if they can be appraised as positive or negative valence and if they are described using specifics or a general feeling. This exploration is part of a case study designed to search for differences in language patterns for a specific TLM and a demographically comparable RLM informant.

I found coding for the RLM informant more difficult than coding for the TLM informant. The concordances for *because* seemed less clearly attributional or appraisable. The first instance follows:

> Quote 9.12: health obviously is a. a difficult one because it stops you doing things that you enjoy doing but if you've got poor health if the if you've got family and friends around you it's a lot easier to deal with that.

In terms of attribution, the scope of outcome for health in this case is specific. This is because, whilst the informant talks about the *you*, it is ambiguous but appears to be general, therefore not discussing wider consequences than to one individual. From an appraisal perspective, there is no clear positive or negative valence from this, through the rest of the interview it is clear that family and friends hold a positive valence. They use the words *difficult* and *poor* which are negative valence, then *enjoy* and *easier* which are associated with positivity. What is clear is that they value health as important, and that this is general health rather than any named health conditions.

The next instance of *because* follows:

> Quote 9.13: uhm if you don't have family and friends then what's the point in having money and health because you've got nobody to enjoy them with

In quote 9.13, money and health are described as an indirect complement to family and friends. The scope of outcome for money and health is specific benefits to individuals rather than a global outcome. There is a negative valence as the informant talks about a lack of point, and this is based on the general relationships endophorically referenced as family and friends rather than specific individuals.

Most of the co-text of *because* is similar to quotes 9.12 and 9.13, whereby informant 128 did not use as clearly codable language as informant 95. There is one exception to this, and this is a quote in which the informant talks about an abusive relationship:

> Quote 9.14: walking on egg shells because he was so. insecure and jealous and. I suppose because I lost. I lost contact with my friends and I. had… or did I just didn't have friends I had him and his family.

The scope of outcome from the situation described can be attributed as specific to the speaker because the informant is the only one referenced as being affected by their spouse. The affect is described are negative, both in terms of the experience of the speaker but also the feelings of the abuser. The situation is described by the informant in terms of general feelings and experiences such as losing contact with friends rather than specific examples of losing contact.

Based on an examination of all attributional excerpts from RLM informant 128, their speech patterns are more likely to be specific rather than global in scope of outcome for appraisal. Further to this, they are not likely to have a clear positive or negative valence as part of their interview. In the one instance where there was clear valence, this was negative and described in terms of generalities.

9.4 Attribution meets appraisal discussion

In this chapter, I examined the language of my top four TLM and RLM prototypical informants for attributional styles. I particularly concentrated on scope of outcome, and found that the TLM informants were more likely to attribute a global outcome to a concept with negative valence than the RLM informants. I further explored this by analysing the first topic of importance that each informant spoke about, using this to understand how they view the thing that is most important, based on the question they were asked in the interview. RLM informants were more likely than TLM informants to speak specifically about the things that they found important, rather than generally. This rule was not uniformly followed by all of the data and therefore required further exploration using linguistic appraisal analysis.

Appraisal analysis coding was more difficult to do for the RLM case study informant than for the TLM case study informant. When examining the thought units based on the first topic of interest, I discovered that this was the case for RLM informants when compared to TLM informants generally. Based on the research into depression from the literature review, particularly sections 2.4 and 2.5, I believe that this is due to an increase in the usage of absolute terms for my TLM informants. Al-Mosawi & Johnstone (2018) advise that depression may lead to an increase in the use of language indicating that things are either extremely good or extremely bad, with no moderation between the two. It therefore follows that my TLM informants would be more extreme in their language, when compared to the moderate (and therefore more difficult to code) language of the RLM informants.

The main findings of this chapter are building upon the analysis of sections 6.3, 6.5, 7.4 and 7.5. When talking about thought units with negative valence, TLM informants use specific language whilst RLM informants use more general language. As an example, TLM informant 194 spoke about the importance of raising money for their local food bank, when compared to RLM informants who spoke about social justice generally. Contrary to this, when talking about thought units with positive valence, TLM informants use general language whilst RLM informants talk about specifics. As an example, TLM informants are more likely to speak about family whilst RLM informants will name individual family members such as a spouse or children.

In this chapter I have combined aspects of appraisal and attribution theory to explore indications from previous analysis. I found that my low mood informants were more likely to use specific lexis when talking about things with negative valence, whilst the reference informants use general and unspecific terms around negative things. I also discovered that my low mood informants were more likely to use general and unspecific terminology when talking about things with positive valence, whilst the reference informants use specific names and lexis around positive things. In the next chapter, I draw my findings from my analysis of chapters six to nine together and discuss potential impacts.

Chapter 10 Tense and tendencies discussion

10.1 Introduction

This chapter brings together the themes of the thesis and provides an opportunity to explore their meaning as well as applications. I address the responses to my research questions: *What linguistic markers are prevalent in the verbal language of people experiencing low mood?* And *Do people with low mood focus on the past?* These research questions are specifically based on low mood rather than depression. During my psychological grounding research I decided to broaden the scope of this thesis to interrogate a *low mood* rather than *depression* psycholect, to ensure applicability. Throughout this chapter I explore the results of my investigation as well as reflecting on the methodology used for my main study. This is because my pilot study was categorised according to a depression rather than a mood based questionnaire (please see Appendices E and F for the questionnaires). Further to this, since the pilot was preparation for the main study, the ways in which it informed this work have already been covered.

In order to directly address my research questions on low mood I therefore exclusively concentrate on the TLM (target low mood) and RLM (reference, those not experiencing low mood) corpora. I explore the main themes that arose from my analysis. Addressing my first research question, I particularly explore how my informants verbally express the self, agency and power. I also consider how the language of the low mood informants was less lexically diverse than the reference informants. Finally, I draw together the analysis around TLM informants and specificity in relation to valence. In responding to my second research question, I consider my approach and how the past was not as important as the current. In addition to addressing my research questions, I look at the strengths and limitations of the project as a whole in terms of linguistic health communication based analysis. This chapter ends as I draw conclusions, consider the validity of the term *psycholect* and look to future applications for this work.

10.2 The self, agency and power

A common theme identified in research investigating the link between language and depression is a focus on the self. A majority of studies identify self-focus linguistically through an increased use of first person personal pronouns, as meta-analysed by Edwards and Holtzman (2017). In addition to a focus on the self, people experiencing depression are likely to use less agentive language (e.g. Demiray & Gençöz, 2018). Each of the existing studies into the self, agency and power are based on the language associated with clinically diagnosable mental illnesses such as depression. However, my study is based on *low mood*. I discuss my rationale in detail in section 2.4, along with the Power Threat Meaning Framework (PTMF) (Johnstone and Boyle, 2018). My analyses pertaining to the self, agency and power are particularly relevant to PTMF and the psycholect associated with low mood. These findings contribute towards addressing my first research question: *What linguistic markers are prevalent in the verbal language of people experiencing low mood?*

A commonly accepted linguistic marker associated with self-focus is increased use of first person singular personal pronouns, according to Edwards and Holtzman (2017), specifically "*I*, *I'd*, *I'll*, *I'm*, *I've*, *me*, *mine*, *my*, and *myself*" (Edwards & Holtzman, 2017, p.63). As part of the keyword comparison of TLM and RLM in section 6.5, I found that my low mood informants were characteristically more likely to use the first person singular *me* when compared to the reference group. In addition to being a keyword, *me* is also a key part of speech (PoS) tag from the comparison of TLM and RLM in section 7.5. This PoS tagging search indicated that both *first person singular* (*me*) and *singular reflexive pronouns*, (*myself*, *yourself* and *itself*) were significantly more likely to be in the verbal language of the TLM informants when compared to the RLM informants.

Unlike the meta analysis of Edwards and Holtzman (2017), I did not find increased usage of all first person pronouns (*I, my, mine, myself*), solely *me*. However, Edwards and Holtzman (2017) exclusively analysed studies that had used LIWC (Pennebaker et al., 2022) which groups together all first person pronouns. As psychologists, Edwards and Holtzman (2017) and their sources do not provide samples from their data, leaving us with no knowledge as to which personal pronouns were more prevalent. I have therefore built on their research to provide specificity, that an increased use

of *me* and *singular reflexive pronouns* are a language marker associated with low mood in my corpus. Demiray and Gençöz (2018) found that better mental health is associated with first person pronouns moving from an object to a subject position, indicating greater agency.

In their Turkish language data, Demiray and Gençöz found that patients experiencing poor mental health linguistically indicated a lack of agency during initial therapy sessions (Demiray & Gençöz, 2018, p.959). To link this to my finding, psychologists including James (2019) state that *I* is how one refers to oneself as an active or knowledgeable participant, whilst *me* indicates a view of oneself as more of an object (James. 2019, p.400). An increased use of *me* in TLM compared with RLM therefore indicates that TLM informants view themselves as objects rather than as individuals with agency. My English language data shows a parallel lack of agency with the Demiray and Gençöz (2018) transcripts from those experiencing poor mental health.

A key finding from the analysis of my corpus is that the self-focus of a psycholect associated with low mood is revealed by nuanced cues. An increased use of *me* and *singular reflexive pronouns* in TLM compared to RLM is a clear indicator, but there are examples of nuance. I found linked indicators of a subtle difference in personal focus between TLM and RLM as part of the topic comparison (see section 6.3). In this section, I coded the transcribed responses to the interview prompt: *When you think about what is most important to you, what comes to mind*? I then grouped these coded transcriptions into semantic topics. These groups were: Self; connections; society; nature; spiritual; physical; hobbies. I made a distinction between connections and society, connections are individual relationships whilst society is larger scale associations such as the desire for social justice. When examining the data grouped according to corpus, I found that the TLM informants were less likely than the RLM informants to place importance on topics that fit with the *connections* grouping. An increased concentration on distinct relationships with others rather than oneself indicates RLM informants showing a reduced focus on the self when compared to TLM informants. Further to this, TLM informants were more likely than RLM informants to talk about topics that fit into the *self* group. These topical foci combined create a balanced picture, that the TLM informants are looking to themselves at the cost of looking to the personal relationships around

them. Psychologists Roberts and Christens (2021) found that an increased sense of community and belonging was an essential component of wellbeing for people who identify as LGBT, including individuals of a range of ethnic origins. There is a nuanced self-focus evident in my TLM informants who place a reduced emphasis on distinct relationships with others when compared to RLM. A greater sense of community and a value for these connections may help the TLM informants in much the same way as it helped the LGBT informants from Roberts and Christens' study (2021, p.406).

I have considered how TLM informants look inwardly to themselves, now to look at the power and agency held within that sense of self. Johnstone and Boyle (2018, p.5) write that the dynamics of power and agency are a core component of mental health. In particular, that people who do not feel that they hold power in and for themselves, as well as agency over their lives, are more likely to experience psychological distress (Johnstone and Boyle, 2018, p.5). When Demiray & Gençöz (2018) analysed Turkish therapeutic transcripts, they found that personal pronoun placement relative to verbs changed as therapy became more successful. In particular, they found that first person pronoun placement moved from an *object* to a *subject* position, thereby indicating an increased sense of agency. In my English language data, a sense of agency is conveyed through the usage of verbs in TLM compared to RLM. Since verbs are words of agency and dynamism (Carter & McCarthy, 2010, p.301) I explored their keyness as part of my PoS tagging comparison of TLM and RLM in section 7.2. There was one positive key PoS tag associated with a verb, and this was the base form of the verb be. In context, the word *be* indicated a lack of agency and positivity for the TLM informants, for example: *uhm to* ***be*** *able to* ***be*** *proud of yourself*. In this TLM example, the informant linguistically distances themselves from the desire to be proud of oneself with the negatively dynamic modalising clause *to be able*. This informant feels so little agency that they do not hope to be proud of themselves, purely for the ability to feel this pride. An informant with more agency might say that they would like to feel proud of themselves, but this informant does not indicate any feeling of material agency.

Part of the key PoS tagging analysis in section 7.2 was an analysis of negative part of speech tags from the comparison of TLM and RLM. Negative part of speech tags are those characteristically less likely to be present in TLM when compared to RLM. A difference in verb usage indicated a reduced likelihood of agentive language in TLM when compared to RLM. The particular negative PoS tags were *doing*, *do,* and the lexical form of the *-ing* participle. Each of these verb forms articulates a sense of power and dynamism, relatively absent form TLM when compared to RLM. Both *doing* and *do* indicate a dynamic modality, an ability to do as well as the ongoing action of doing. The morphology of the lexical form of the *-ing* participle is differentiated from the adjectival *-ing* in the PoS tagging function of *Wmatrix*. This use of the -ing suffix is progressive: indicating ongoing action, as exemplified in the following quote from a RLM informant: *I enjoy leading a team of people at work*. This is a simple sentence revealing power and agency linguistically and thematically, this confidence is exhibited through their statement that does not feature any hedges or hopes.

One of the objectives of this thesis is to see how linguistic analysis can engage with psychological approaches to low mood. For example, one key area is the diagnosis and distress based model of mental health. If psychologists are to move away from a distress-based model of diagnosis based on collections of symptoms and towards a more holistic approach, one of the factors under consideration is low mood (Kamens et al. 2018, p.401). In this exploration of the self and agency, I have indicated linguistic markers associated with what I am calling a low mood *psycholect*. These are particularly relevant to Power Threat Meaning Framework (PTMF) as proposed by Johnstone and Boyle (2018). PTMF recognises “the centrality of meaning, narrative, agency and subjective experience” (Johnstone and Boyle, 2018, p.6). As my TLM informants verbalise their meaning making around what is important in their lives, they show the centrality of themselves, their lack of perceived agency and occasionally narrate their own subjective experience.

10.3 Lexical diversity

When addressing my first research question: *What linguistic markers are prevalent in the verbal language of people experiencing low mood?* I started to notice more repetition in the verbal language from TLM when compared with RLM. Many of my informants across TLM and RLM expressed themselves using a high register, displaying lexicogrammatical complexity. For example, informants from TLM and RLM used uncommon, idiomatic, Latinate, polysyllabic terms or phrases such as: *ephemeral*; *tap dancing on the precipice of nihilism*; *actualisation*; *ludicrous*; *stretches back to the dawning of time*; *maelstrom*; *there's a kind of moral imperative*; *autonomy; mastery and purpose*; *nebulous*. Whilst there was a similarity of register, I suspected that there was a difference in the diversity of words used between TLM and RLM. Looking to the literature, Park and Conway (2018) found that mental health internet forum users had reduced lexical diversity when compared to their reference users. In order to calculate lexical diversity, Park and Conway (2018, p.e121) looked at the number of unique words across their target and reference corpora. For my own data, I performed two different calculations which indicated a reduced level of lexical density and diversity in TLM compared to RLM in section 6.4. The most important is the Measure of Textual Lexical Diversity (MTLD) as developed by McCarthy and Jarvis (2010) to distinguish how lexically diverse a text is without being skewed by corpus length. The MTLD scores indicate that there is less lexical diversity overall in TLM when compared to RLM. Whilst the measure is different, the difference in lexical diversity between people experiencing poor mental health and those who are euthymic is in keeping with the work of Park and Conway (2018). To connect with my first research question, this is an indicator that a linguistic marker associated with the verbal language of people experiencing low mood is low lexical diversity, when compared to those who are not experiencing low mood.

The statistical difference in lexical diversity between TLM and RLM as a whole was also linguistically visible through the PoS tagging comparison. Within the positive PoS tags: *prepositions*, *general adjectives* and *modal verbs* there was a greater volume of each PoS tag in TLM compared to RLM but more repetition of the individual words. Further to this, interjections are a sign of complex speech (Dingemanse, 2024, p.266), and TLM informants were less likely to use interjections than the RLM informants. Finally, the TLM informants are less likely than RLM informants to use *general adverbs,*

and these descriptors add to the complexity of speech (McCarthy & Jarvis, 2010, p.313). In addition to a reduced likelihood of using general adverbs, the TLM informants are once again more likely to repeat the ones that they do use, rather than using the diversity of verbal language that the RLM informants employ.

Park and Conway (2018) did not provide any theories about the reason for the difference in lexical diversity between their mental health and reference corpus. I suggest that reduced lexical diversity in TLM compared with RLM may be due to rumination as described by Şimşek (2013). Whilst the informants verbally ruminate, they may repeat words as a reflection of the thoughts that repeat in their minds. This requires greater research to establish how prevalent and pervasive this pattern is, and if there is a link between repetition and rumination. My language data indicates that there is a greater impact on individuals experiencing low mood than is currently acknowledged in diagnostic guidelines (e.g. WHO, 2022). This is an instance whereby diagnostic labels do not cover the full experience of those who experience this embodied sense of low mood. Reduced lexical diversity could be indicative of an increase in rumination alongside associated reduction in cognition and will impact individuals on a personal, professional and academic level. This reduced lexical diversity is therefore an underexplored but viable and valuable aspect of the low mood condition.

My data analysis has reinforced the ideas posited by psychology academics such as Kamens et al. (2018), that diagnostic nomenclature such as *depression* and *anxiety* may no longer be relevant. It is more important to understand the embodied experience of an individual and look at the underlying cause rather than to loosely group symptoms under labels (Kamens et al., 2018, p.401). The reduced lexical diversity for my TLM informants when compared to my RLM informants is suited to this more nuanced understanding of mental health. A lexical diversity analysis would particularly suit outreach for alternative criteria for understanding emotional distress. There are many ways to enable mental health outreach, from communications including leaflets through to contacting patients considered at-risk (Burns, 2001, p.13). Monitoring at-risk patients for a relative reduction in lexical diversity could enable specialists to perform interventional outreach. There could be a baseline, whereby anybody with a MTLD measure below a certain figure can be considered to be experiencing

low mood. It would require further research to establish if there are set parameters of the lexical diversity of the low mood psycholect. Since many of my TLM informants did not realise that they were experiencing low mood, this would enable interventions before the low mood becomes more serious.

One of the ways of understanding the experience of poor mental health that I have discussed in section 2.4 is the Power Threat Meaning Framework (PTMF). PTMF involves ways of ameliorating unhelpful coping strategies by examining the interaction of societal, personal, interpersonal and environmental power and threat (Johnstone and Boyle, 2018, p.12). Widespread adoption of PTMF would enable practitioners to treat the emotional distress causing reduced cognition, rather than purely the symptoms associated with depression and a range of other mental health issues. Potential reduced cognition, manifested as reduced lexical diversity, has an impact on individuals on a personal, interpersonal and societal level and therefore is a good fit for the PTMF framework. My choice to investigate low mood rather than depression has enabled me to add to this conversation around mental health.

10.4 Specificity

We have explored the advantages of considering the experience of low mood on an individual level rather than grouping symptoms into diagnoses such as *depression* or *anxiety*. In a meta-analysis of studies into grouped mental disorders, Böttger et al. (2019, p.1) found that studies frequently concentrated on negative valence associated with the experience. A sense of negativity is a prevailing factor of the low mood experience, but can also be the sign of an isolated bad day rather than an ongoing mental health crisis. People experiencing low mood are more likely to be negative (Huston et al., 2019). However, they may still talk about positive things. I therefore felt it was important to look not purely at language associated with negative valence, but also how informants talk about topics with positive valence. This holistic approach to linguistic markers associated with low mood has provided a balanced picture of the low mood psycholect and insight into the low mood experience.

My TLM informants were more likely to talk about specifics when discussing topics with a negative valence. Contrary to this, when discussing topics that had a sense of positivity the same TLM informants were more likely to talk in generalities. I made an active attempt to situate this finding within the literature and to find any studies that could confirm or refute this. I could find no other studies into generalities or specifics for mental health based health communication, either in psychology or linguistics. This indicates an opportunity for other studies to explore the concepts around valence and level of specificity in the context of low mood further. My own findings indicate that people experiencing low mood talk about positive things in general terms and negative things in detail.

The main question I asked all of my informants was: *when you think about what is most important to you, what comes to mind?* When considering what is important to them, my informants had a tendency to talk about things with positive valence. During my investigation into the topics that TLM informants spoke about with when compared to RLM informants, I noticed a general vagueness from the TLM informants, in a way that could be investigated using corpus tools and linguistic frameworks. The advantage of the keyword search in section 6.5 is that this disclosed a quantifiable tendency towards abstract concepts from the TLM informants when compared to RLM informants. In particular, TLM keywords included nominalisations such as *relationships* and *interactions*, when compared to RLM informants who are more likely to name specific individuals such as their children or friends. Another keyword from TLM compared to RLM from section 6.5 was *concepts*, and I found that the TLM informants were likely to talk conceptually. This is in comparison to RLM informants who used demonstrative terminology to point to specific individuals or things that are important to them. The patterns indicated in the keyword search were further amplified by the key Part of Speech (PoS) tag search. Of particular importance was the preponderance of *general adjective* in TLM when compared to RLM. These adjectives were used by TLM informants to describe ideas rather than specifics. In the context of the positive valence of the things that are important to TLM informants, this is another indication of positivity being linguistically manifested in terms of generalities.

Relevant to the idea of specificity, as part of the PoS tagging analysis, in section 7.5 I explored use of personal pronouns in TLM compared with RLM. I found that the TLM informants did not use many third person singular pronouns (e.g. *herself*) because they don't talk about specific individuals and therefore do not need to use the pronoun form to refer to them. RLM informants used a higher proportion of third person singular pronouns when compared to the TLM informants, as they referred to specific people such as their spouse, children and so on. TLM informants used pronouns associated with general groups that they are a member of (*we*), or themselves (*me*).

When looking at keywords according to demographics in section 6.7, I found a difference in the use of the pronoun *them* between TLM women and RLM women. The women in TLM use *them* to discuss people or ideas in general, whilst the women in RLM use *them* to refer to specific people. As an example, from a woman in TLM discussing a hypothetical person:

TLM quote 10.1: being furious at someone because I feel like they took advantage of me is is is something I can… oh I can't change them

Contrasted with an example from a woman in RLM who is talking about their specific and named children:

RLM quote 10.2: there's the fact that I love them very much and and want the best for them

This is a distinct difference in the use of *them* that I discovered as a result of separating out my data according to demographics. This difference is indicative of a demographic difference, in this case gender, that is instructive for our understanding of language and low mood. Previous studies have found a difference in how people of different genders (e.g. Angst et al., 2002); sexualities (e.g. Argyriou et al., 2021) and races (e.g. Bailey et al., 2019) experience depression. This research indicates that this differential in experience can also manifest with a difference in linguistic styles. Although the informants from each demographic were given the same prompt, the low mood psycholect manifested in different ways. The counter balance to this increased granularity is that the corpus size was relatively small and therefore not generalisable. However, this does indicate that ensuring representation for each member of a demographic group, be it gender, sexuality, ethnicity, education level or class, is essential to ensure research is inclusive.

Chapter 9 is where the intersection of attribution and appraisal theory enabled me to fully explore the concept of generality and specificity. Through the close analysis in section 9.3, it became clear that the generalities that TLM informants use when talking about topics of positive valence had a counterpoint. The language of TLM informants is different to RLM informants when they talk about topics with negative valence. TLM informants are more likely than RLM informants to talk about specifics. Here are two examples from the same TLM informant, where quote 10.3 has positive valence and quote 10.4 is negative.

> Quote 10.3: These concepts of honesty integrity the beliefs that you have like you hold to your person uhm the things that make you happy that allow you to have the energy to believe

> Quote 10.4: It is important. I do not go back down. the addictive path it is there are days when it is very not easy

Quote 10.3 is the only time that this TLM informant talks about happiness, and they do so around concepts (e.g. *honesty*, *integrity*) rather than anything specific. They also equate positive valence with a sense of dynamism, as in *energy to believe*. In quote 10.4 the TLM informant has been speaking about a drug addiction and the negative impact this has had on their life and they therefore name a specific thing that they want to avoid. Each of the TLM informants follows a similar pattern. Whilst the topics of importance may change, the generalities remain tied to positive valence and the specifics are associated with negative valence.

Cognitive related therapies such as Cognitive Behavioural Therapy and Rational Emotive Therapy (David et al., 2018) are used to address biased thinking such as that found in individuals experiencing low mood. The results of studies showing an overall negative valence have been used to inform this type of therapy (Böttger et al., 2019, p.2). My corpus offers evidence of the entrenched tendency to articulate negative things in specific terms, which is a more nuanced feature of the language of low mood. Further to this, there are indications that low mood informants use generalised language as part of their positive thinking. These findings help to validate and may be used to strengthen the strategies of CBT and related therapeutic approaches.

10.5 Temporal focus

When embarking upon this PhD, I theorised that people experiencing low mood look to the past rather than the present or the future. This was largely based on anecdotal evidence as well as reading published first-hand accounts of depression (e.g. Soloman, 2001). My second research question was designed to test this hypothesis: *Do people with low mood focus on the past*? Here, my main corpus data shows no evidence to support this kind of distinction in temporal focus. This confirms that the perception of low mood from lay people, both observers and experiencers, can be flawed. However, the results highlight another feature relating to temporality, which may not have been identified without my initial hypothesis. This is that the TLM informants were more likely to consider temporality as a constant than the RLM informants.

Whilst the title of this thesis refers to tense, the nuances of English grammar mean that tense alone is not enough to understand temporal focus (Halliday & Matthiessen, 2013, p187). I therefore constructed my research question around a focus on the past rather than the past tense. After attempting to base this analysis on a tense based PoS tags comparison, or using the framework of a pedagogical Tense Tracking Diagram (TTD) (Svalberg, 2019), I finally hand coded my data to enable an analysis of the nuances of temporal focus. Whilst hand coding is more subjective than a TTD or PoS tags search, both of these approaches were too simplistic to capture the full range of potential temporal foci.

A key book for my understanding of the embodied experience of depression is *The Noonday Demon: An Atlas of Depression* (Solomon, 2001). This book is a personal and curated collection of essays designed to give a holistic guide to depression from lived experience. As discussed in my introduction, a pivotal quote that led me to my initial hypothesis about low mood and the past is from this book “The worst of depression lies in a present moment that cannot escape the past it idealises or deplores” (Solomon, 2001, p.99). My second research question focused on the concept of an idealised or deplored past, but what my research has indicated is that the pertinent part is “a present moment that cannot escape” (Solomon, 2001, p.99). The inescapable moment is construed by my

TLM informants though a sense of universality, that the feelings of the present are the stable constant in their lives.

An example of an inescapable moment comes from TLM informant 152:

> Quote 10.5: it is from old trauma uhm it is important to me because it is elusive and it doesn't happen very often and I wish that it did more

In quote 10.5, the *it* the informant refers to is their desire for happiness, and they refer to past trauma, current importance, recent elusiveness and future wishes. This combination of temporal foci results in a sense of the universal reference point, spanning past, present and future. The nuances of the English language create space for ambiguity in the interpretation of overall temporal focus. However, the rest of the context of this quote indicates a universality. This informant expresses hope for general happiness (*I wish*), rather than a practical digest of specific steps to help towards this emotion. This vague hope without precision is an expansion of the general vs specific theme discussed in the previous section. This adds to the picture of the linguistic articulation of low mood and the embodied experience that is the source of this verbal language. The sense of universality in low mood informants is something that could be useful for therapeutic application. As part of talking therapies, patients could be encouraged to consider that their current situation will change, and that they are not therefore trapped in this “present moment that cannot escape” (Solomon, 2001, p.99).

10.6 Strengths and limitations

In examining the strengths and weaknesses of this project, I have chosen to use the criteria set in *Corpus linguistics for Health Communication: A Guide For Research* (Brookes & Collins, 2023). Whilst there are other criteria I could have used, I felt that these particularly suited the methodology and focus of my project. Through this section, I consider the benefits and drawbacks of how my research questions were answered according to the design of the project, the data generation methods and analysis.

At the conception of this project, my intent was to use therapeutic transcripts to examine the language of patients experiencing major depressive disorder. However, "ethical considerations are a crucial part of any corpus study" (Brookes & Collins, 2023, p.62). The ethical implications of obtaining and analysing transcripts from patients during treatment were both challenging and impractical. To qualify, I deemed therapeutic transcripts unnecessary since they are based on the label of *depression* as it is currently diagnosed. Further to this, it is difficult to obtain linguistic data directly from a clinical setting (Brookes & Collins, 2023, p.122).

Corpus linguists investigating health communication are less likely to choose to use spoken rather than written corpora (Brookes & Collins, 2023, p.109). This is due to the relative higher cost in terms of time in particular for recruiting, interviewing, and transcribing (Brookes & Collins, 2023, p.48). My design choice to use spoken data was both a strength and a limitation; it provided fruitful data but not enough to enable my results to be fully generalisable. In this context being generalisable or representative is for my informants to reflect “full range or variability in a population” (Biber, 1993, p. 243). The modest size of my corpus means that the results are not fully representative in terms of either the informant demographics or the size of the corpus. There does not appear to be a corpus currently available that is fully representative and of a sufficient size. Even if my corpus results can only be taken as an indication of the specific population I interviewed, my investigation has clearly highlighted the potential benefits of this approach. This is given the contrasting data with previous studies, including my own pilot study, and the greater insights into the linguistic preferences of individuals with low mood on an individualised and contextualised basis.

A strength of corpus linguistics is that it works well with naturally occurring data such as that collected through interviews (Brookes & Collins, 2023, p.110). Like many health communication corpora, mine was designed and built specifically for this project (Brookes & Collins, 2023, p.40). When using corpus linguistics, greater volume provides a greater opportunity for finding language patterns (Brookes & Collins, 2023, p.19). My corpus was about half of the 100,000 words suggested as suitable in size for the generalisable study of prosodic features (Brookes & Collins, 2023, p.46).

Balancing generalisability with the time involved in compiling spoken word corpora, Brookes and Collins (2023, p.50) advise not to overextend oneself over the size of a spoken corpus.

Generalisability includes balanced representation of data including demographic groups (Brookes & Collins, 2023, p.46). The popular data source of health communication data for corpus studies is online fora and social media sites, which are “relatively anonymous” (Brookes & Collins, 2023, p.44). Anonymity would also have been an issue if I had used clinical transcripts, as my original plan had been. One of the great strengths of this project is the control and transparency of the data generation process, meaning that the demographic balance of the data is explicit. A limitation is in the range of demographic data, particularly including non-graduates; same-sex attracted; non binary; black and minority ethnic group informants. However, due to the strength in the data generation process, I am actively aware of the gaps in my informant demographics as a limitation, and have attempted to take these into account in my analysis and interpretations.

Societal attitudes towards the intersection of the demographic groups we belong to have an impact on mental health (Holden et al., 2012). The embodied experience of each aspect of our identity contributes towards our mood level (Roberts & Christens, 2021). Without taking an essentialist view, it is important to be aware of contributory factors in order to ensure fully inclusive amelioration and outreach strategies for low mood. My data suggests that expanding the research in the future, to develop a corpus of similarly elicited and categorised data could provide new insights into gendered, age-based, sexuality, education level and race based language associated with low mood.

The interview stage of the data generation process was a strength in terms of fidelity, context and reflexivity. When collecting natural speech, it is essential to record the data faithfully the first time, since it is not possible to accurately reproduce authentic responses (Brookes & Collins, 2023, p.49). During my data generation process, in addition to the ethical process of ensuring consent for recording, I made certain that the recording process was working and that I had secure back-ups for each interview. Context of each interview is also important to the collection of spoken data for health communication (Brookes & Collins, 2023, p.50). My data generation protocols meant that I had contextual demographic details as well as a secure log of the interview conditions themselves.

This was due to my reflexive diary which provided context for a range of aspects of the interview including my own positionality. This reflexivity allowed me to consider how I “impact the participants and the research and how the researcher might be impacted by the research process itself" (Brookes & Collins, 2023, p.67). I have reflected on how the research interviews impacted me during section 3.6, particularly the difficulty of maintaining the separation of interviewer rather than friend. The impact that I had on my informants includes the opportunity for them to feel a sense of connection with me, for them to be useful in helping with my research and for them to have the opportunity to talk for as long as they wished about things that they value.

My transcription protocols were based on those devised by Collins and Hardie (2022) to be specifically suitable for corpus linguistic analysis. These transcription guidelines (Appendix E) are another strength of this project, since they were carefully prepared to be used in this health communication analysis. Important components include textual mark ups such as laughter (Brookes & Collins, 2023, p.59) and de-identification (Brookes & Collins, 2023, p.62). In this context, de-identification was the removal of any identifiers, including my use of a random name generator to anonymise the transcripts.

To perform some forms of corpus linguistic analysis of health data, it is necessary to have a target corpus and a reference corpus. Brookes and Collins (2023, p.82) write that a "reference corpus should, ideally, represent language that is matched to that in our target corpus” and specify that this matching should cover both register and genre (Brookes & Collins, 2023, p.82). This is one of the largest strengths of this study, that my reference corpus (RLM) was generated in precisely the same way as my target corpus (TLM). Not only are the corpora directly comparable in terms of the data generation, but they also have precisely the same level of metadata available on the informants themselves. At the time of interview I did not know what the mood of my informants would be, and therefore there was no unconscious bias towards the target or reference group. The comparability of the corpora and the informant demographics are distinctly ideal for a health communication project.

Corpus linguistics for health communication has many strengths, but sometimes it becomes necessary to choose a sample from a corpus to examine closely (Brookes & Collins, 2023, p.104). I chose to use *ProtoAnt* (Anthony & Baker, 2023) to find the texts that were most representative of my corpus. This software was released the same year as *Corpus linguistics for Health Communication: A Guide For Research* (Brookes & Collins, 2023) and therefore is not mentioned in that book. I used this software to choose representative texts rather than alternatives such as random sampling. This is the first such application of this software with spoken data corpus in the health communication field that I am aware of. The use of corpus linguistics to choose which individual interviews to concentrate on enabled an empirical method of sampling.

I built my dataset to be compatible with corpus linguistic methods, but still needed to choose which software to use in my analysis. There are three types of corpus linguistic software, each of which has different aptitudes and these are "Online, offline, and DIY tools" (Brookes & Collins, 2023, p.78). I primarily used *AntConc* (Anthony, 2010) which is an offline tool, to perform my data analysis. This is because offline tools are more transparent in their workings than online tools, whilst requiring less coding expertise than DIY tools (Brookes & Collins, 2023, p.79). I also used the online tool of *Wmatrix* (Rayson, 2009) for the parts of analysis that required access to the BNC spoken sample, since only the wordlist rather than the full file is available for use with *AntConc*. In addition, I used *Wmatrix* for the constituent likelihood automatic word-tagging system (CLAWS) as a time saving device for part of speech tagging. The choice to use both online and offline corpus linguistic software was a strength of this project, enabling me to use the correct software for each task.

Corpus linguistic software provides many analytical options which can be used individually or in combination based on the needs of the project (Brookes & Collins, 2023, p.104). In order to examine patterns of language, I used keywords which enabled “an inductive ‘way in’ to the corpus that typically precedes and guides more qualitative analysis" (Brookes & Collins, 2023, p.104). An exploratory analysis using keywords from my directly comparable corpora is one of the strengths of the project. However, the level of fine-grained sub-corpus analysis that would have been helpful in terms of demographic analyses and so on was limited by the size of my corpus. Building upon the

keyword-based thematic analysis with contextual information from concordances is a way of clearly understanding what relevant themes mean in use (Brookes & Collins, 2023, p.86). The collocate lines that I include throughout the analytical chapters enable you, the reader, to see this context and enable greater transparency of my analytical process.

When examining the part of speech tags that were of particular interest, if there was no clear pattern of use, I performed collocation analysis. This is another strength of my project, since understanding the lexical company of regularly used part of speech tags enables us to “learn more about a word’s meanings and patterns of use” (Brookes & Collins, 2023, p.91). Since my first research question was: *What linguistic markers are prevalent in the verbal language of people experiencing low mood?* this insight into patterns of use was a strength.

This project features a great number of strengths, some of which have directly led to limitations. An example of this is the strength of the controlled data generation process leading to corpora that are too modest to be generalisable. On reflection, this work has more strengths than limitations and the process and results of this health based analysis “as articulated in micro contexts of interaction, inevitably intersect with macro level social and political discourses" (Brookes & Collins, 2023, p.239).

10.7 Conclusion

The simple elegance of my thesis title and research questions belies the complexity of both the embodied experience of mental health and of the English language. The experience of low mood is part of intersecting factors around socialisation, cognition, power and levels of support. In response to my first research question: *What linguistic markers are prevalent in the verbal language of people experiencing low mood?* there are distinctive features associated with the low mood psycholect, and two of these are visible through simple searches. An increased use of the first person pronoun *me* in TLM compared to RLM can be found through a keyword search. To establish the lexical diversity of the verbal language of TLM against RLM, it is possible to paste the corpora into a website (Reuneker, 2017). Due to the ability to search for both of these features, these could be used to help mental health professionals with outreach for people experiencing low mood.

In response to my second research question: *Do people with low mood focus on the past?* The simple answer is that they do not, but they are more likely to believe that their current experience is temporally universal; that is, that the way things are at present is how they always have been, and always will be. Whilst this finding may not have immediate health communication applications, it does demonstrate one of the difficulties of the embodied low mood experience. This sense of temporal universalness also suggests an area of focus for future clinical approaches.

The final part of my findings are around agency, power and generalities against specifics in terms of positive and negative valence. These findings support the idea of a more complex reading of language and experience than can be covered by my research questions. The lack of perceived power and agency that my TLM informants experience concurs with the psychological insight of Johnstone and Boyle (2018). This is particularly important at the stage where psychology may be moving towards frameworks based on power and agency such as the PTMF (Johnstone and Boyle, 2018). My findings around generalities and positive valence, alongside specifics and negative valence are distinct. Not only is this an insight into the embodied experience of low mood, but could also be used for therapeutic application.

In table 10.1, I summarise each of the linguistic tendencies of a low mood psycholect and how I found them:

Psycholectical tendency	Analytical source
Increased use of first person pronoun *me*	Keywords
Increased use of first person singular reflexive pronouns	PoS tag search
Less agentive use of verbs	PoS tag search
Reduced lexical diversity	MTLD (McCarthy & Jarvis, 2010)
Universal temporal focus	Hand coding (including use of *present perfect progressive* or *infinitives)*
Topics with positive valence considered in general terms	Topics hand coding and key PoS tag *adverbs* (describing general/unspecific relationships)
Negative valence considered in terms of specifics	Attribution and appraisal hand coding

Table 10.1: Tendencies of low mood psycholect

At the beginning of this thesis I introduced the term *psycholect*, and suggested that it could be used to indicate lexicogrammatical patterns that are distinct to psychopathologies. I have been exploring

the possibility of a definable low mood psycholect throughout this thesis. My analysis indicates that people experiencing low mood do have a tendency towards distinctive spoken language features. Whilst these language features may be nuanced at times, I have made a case for the existence of, and features of, the low mood psycholect. There is an opportunity for further research in this area, to both reinforce and expand our current understanding of the low mood psycholect.

This project has created a foundation for future studies, as well as cementing existing approaches and findings. In terms of existing approaches, data generation designed around an intersectional approach is essential for inclusive effective approaches to health. From a reflexive perspective, this study has also shown the benefit of performing mental health research with a softer focus. Rather than using the diagnostic term *depression*, this study has shown that concentrating instead on low mood is a fruitful approach.

References

Abreu, R. L. ; Townsend, D. ; Mitchell, Y'L. ; Ward, J. ; Audette, L. ; Gonzalez, K. A. 2022. LGBTQ Qualitative and Mixed Methods Research in Counseling Psychology: A Content Analysis. *The Counseling psychologist* 50 (5), pp.708-737 [Accessed 4 March 2022] Available from: DOI 10.1177/00110000221092481

Adil Jaafar, E. & Abdul-Salam Jasim, H. 2022. A corpus-based stylistic analysis of online suicide notes retrieved from Reddit. *Cogent arts & humanities*. **9**(1). [Accessed 5 June 2025] Available from: DOI: 10.1080/23311983.2022.2047434

Adolphs, S. & Carter, R. 2013. *Spoken corpus linguistics: from monomodal to multimodal*. Abingdon: Routledge

Al-Mosaiwi, M. 2018. *People with depression use language differently – here's how to spot it*. The Independent. 5th Feb 2018. [Available at: https://www.independent.co.uk/life-style/health-and-families/depression-language-differently-mental-health-kurt-cobain-a8194731.html]

Al-Mosaiwi, M. & Johnstone, T. 2018. In an Absolute State: Elevated Use of Absolutist Words Is a Marker Specific to Anxiety, Depression, and Suicidal Ideation. *Clinical psychological science*. 6(4), pp.529-542. [Accessed 7 November 2021] Available from: DOI 10.1177/2167702617747074

Al-Sharif, R. 2021. Critical realism and attribution theory in qualitative research. *Qualitative research in organizations and management* 16(1) pp127-144 [Accessed 18 July 2021] Available from: DOI 10.1108/QROM-04-2020-1919

Allan, K. 2016. *The Routledge handbook of linguistics*. London: Routledge.

Alsina, V., Espunya, A. & Naro, M.W., 2017. An appraisal theory approach to point of view in Mansfield park and its translations. *International Journal of Literary Linguistics*, *6*(1). [Accessed 2 June 2025] Available from: https://doi.org/10.15462/ijll.v6i1.103

Amaireh, H. A. 2023. Biden's Rhetoric: A Corpus-Based Study of the Political Speeches of the American President Joe Biden. *Theory and practice in language studies,* 13**,** 728-735. [Accessed 22 February 2025] Available from: DOI: 10.17507/tpls.1303.22

Angold, A. & Costello, E. 1987. Mood and Feelings Questionnaire (MFQ*). Developmental Epidemiology Program;* North Carolina: Duke University [Accessed: 21 November 2021] Available from: DOI 10.1037/t15197-000

Anthony, L. 2010. *AntConc*. [Computer software]. Tokyo, Japan: Waseda University.

Anthony, L. & Baker, P. 2023. *ProtoAnt* (Version 1.3.0). [Computer software]. Tokyo, Japan: Waseda University.

APA. 2013. *Diagnostic and statistical manual of mental disorders (DSM-5).* 5th ed. Washington DC: American Psychiatric Publishing

Archibald, M. M.; Ambagtsheer, R. C.; Mavourneen, C. G. & Lawless, M. 2019. Using Zoom Videoconferencing for Qualitative Data Collection: Perceptions and Experiences of Researchers and Participants. *International Journal of Qualitative Methods* 18, pp.1-8 [Accessed 12 August 2022] Available from: DOI: 10.1177/1609406919874596

Argyriou, A., Goldsmith, K. A., & Rimes, K. A. (2021). Mediators of the disparities in depression between sexual minority and heterosexual individuals: A systematic review. *Archives of Sexual Behavior*, *50*(3), 925-959. [Accessed 7 October 2024] Available from: DOI 10.1007/s10508-020-01862-0

Attardo, S. & Pickering, L. 2023. *Eye Tracking in Linguistics*. London: Bloomsbury Publishing.

Azure. 2022. *Microsoft teams transcript VTT cleaner*. [online] [Last accessed 12.01.2023] Available from: https://amsglob0cdnstream13.azureedge.net/vttcleaner/CleanVTT.html

BAAL. 2021. *Recommendations on Good Practice in Applied Linguistics*. 4th Edition. Available at www.baal.org.uk

Bailey, R.K.; Mokonogho, J. & Kumar, A. 2019. Racial and ethnic differences in depression: current perspectives. *Neuropsychiatric disease and treatment*. 15, pp.603-609. [Accessed 18 January 2021] DOI: 10.2147/NDT.S128584

Baker, P.; Hardie, A. & McEnery, T. 2006. *A Glossary of Corpus Linguistics.* Edinburgh: Edinburgh University Press.

Baker, P. & Levon, E. 2015. Picking the right cherries? A comparison of corpus-based and qualitative analyses of news articles about masculinity. *Discourse & Communication*, 9(2), pp.221-236 [Accessed 3 November 2024] Available from: DOI: 10.1177/1750481314568542

Ballard, E. D.; Henter, I. D. & Zarate, C. A. 2019. Chapter 1 - The Classification of Depression: Embracing Phenotypic Heterogeneity in the Era of the RDoC. Quevedo, J.; Carvalho, A. F. & Zarate, C. A. (eds.) *Neurobiology of Depression: Road to Novel Therapeutics*. Cambridge MA: Academic Press.

Barker, H. & Brown, M. 2011. *Howard Barker interviews, 1980-2010 : conversations in catastrophe*. Bristol [England]: Intellect.

Barratt, C. 2022*. "It doesn't count here": super-diverse community experiences in a local history museum*. [Accessed 28 February 2025]. Available from: https://figshare.le.ac.uk/articles/thesis/_It_doesn_t_count_here_super-diverse_community_experiences_in_a_local_history_museum/18318965/1.

Bayer, J.M.M., Spark, J., Krcmar, M., Formica, M., Gwyther, K., Srivastava, A., Selloni, A., Cotter, M., Hartmann, J., Polari, A., Bilgrami, Z.R., Sarac, C., Lu, A., Yung, A.R., McGowan, A., McGorry, P., Shah, J.L., Cecchi, G.A., Mizrahi, R., Nelson, B. and Corcoran, C.M. 2023. The SPEAK study rationale and design: A linguistic corpus-based approach to understanding thought disorder. *Schizophrenia research*. **259**, pp.80–87. [Accessed 4 June 2025] Available from: DOI: 10.1016/j.schres.2022.12.048

Beauboeuf-Lafontant, T. 2007. You have to show strength: An exploration of gender, race, and depression. *Gender & Society,* 21**,** 28-51. [Accessed 2 February 2021] Available from: DOI: 10.1177/0891243206294108

Bednarek, M. 2008. *Emotion Talk Across Corpora*. Basingstoke: Palgrave Macmillan UK.

Bertolote, J. M. & Fleischmann, A. 2002. Suicide and psychiatric diagnosis: a worldwide perspective. *World psychiatry,* 1**,** 181-185. [Accessed 23 April 2025] Available from: PMID: 16946849

Bernaras, E.; Jaureguizar, J. & Garaigordobil, M. 2019. Child and Adolescent Depression: A review of Theories, Evaluations Instruments, Prevention Programs and Treatments. *Frontiers in Psychology*, 10, p.543 [Accessed 23 March 2023] Available from: DOI: 10.3389/fpsyg.2019.00543

Bernard, J.D.; Baddeley, J.L.; Rodriguez, B.F. and Burke, P.A. 2016. Depression, Language, and Affect: An Examination of the Influence of Baseline Depression and Affect Induction on Language. *Journal of language and social psychology*. 35(3), pp.317-326. [Accessed 20 November 2020] Available from: DOI: 10.1177/0261927X15589186

Bhavnani, K.-K.; Chua, P. & Collins, D. 2020. Critical Approaches to Qualitative Research. In Leavy, P. (Ed) *The Oxford Handbook of Qualitative Research*. Oxford: Oxford University Press pp. 243-262

Biber, D. (1993). Representativeness in corpus design. Literary and Linguistic Computing, 8(4), 243–257. [Accessed 27 June 2025] Available from: https://doi.org/10.1093/llc/8.4.243

Biber, D., Conrad, S. & Leech, G. 2010. *Longman Student Grammar of Spoken and Written English.* New Delhi: Dorling Kindersley by arrangement with Pearson Education Limited

Blakey, S.; Yi, J.Y.; Calhoun, P.S.; Beckham, J.C. & Elbogen, E.B. 2019. Why do trauma survivors become depressed? Testing the behavioral model of depression in a nationally representative sample. *Psychiatry research*, 272, pp.587-594 [Accessed 28 October 2020] Available from: DOI: 10.1016/j.psychres.2018.12.150

Block, P.E.R. & Grund, T. 2014. Multidimensional homophily in friendship networks. *Network science* (Cambridge University Press). 2(2), pp.189-212. [Accessed 16 June 2023] Available from: DOI: 10.1017/nws.2014.17

BNC Consortium. 2007. *The British National Corpus, XML Edition*. Oxford Text Archive, http://hdl.handle.net/20.500.12024/2554.

Böttger, S.J., Förstner, B.R., Szalek, L., Koller-Schlaud, K., Rapp, M.A. and Tschorn, M., 2023. Mood and anxiety disorders within the Research Domain Criteria framework of Positive and

Negative Valence Systems: a scoping review. *Frontiers in Human Neuroscience*, *17*, p.1184978. [Accessed 30 June 2025] Available from: 10.3389/fnhum.2023.1184978

Bourdieu, P. 2021. *Forms of capital.* Cambridge, UK: Polity Press.

Brinkmann, S. & Kvale, S. 2018. *Doing interviews*. London: SAGE Publications Ltd.

Brookes, G. & Collins, L. C. 2023. *Corpus Linguistics for Health Communication : A Guide for Research*. London: Taylor & Francis.

Brown, L.H., Strauman, T., Barrantes-Vidal, N., Silvia, P.J. & Kwapil, T.R. 2011. An Experience-Sampling Study of Depressive Symptoms and Their Social Context. *The journal of nervous and mental disease*. 199(6), pp.403-409. [Accessed 15 November 2023] Available from: DOI: 10.1097/NMD.0b013e31821cd24b

Bryman, A. 2012. *Social research methods*. Fourth edition. Oxford: Oxford University Press.

Browne, K. & Nash, C.J. 2016. *Queer methods and methodologies : intersecting queer theories and social science research*. London: Taylor & Francis.

Bucholtz, M. & Hall, K. 2005. Identity and interaction: a sociocultural linguistic approach. *Discourse studies,* 7**,** 585-614. [Accessed 7 February 2024] Available from: DOI: 10.1177/1461445605054407

Bucur, A.-M., Podină, I. R. & Dinu, L. P. 2021. *A Psychologically Informed Part-of-Speech Analysis of Depression in Social Media*. Open access. [Accessed 30 March 2024] Available from: DOI: 10.48550/arxiv.2108.00279

Burns, T. 2001. To outreach or not to outreach. *The Journal of Forensic Psychiatry* 12 (1) pp.13-17. [Accessed 27 July 2025] Available from: DOI: 10.1080/09585180010027824

Buysse, L. 2017. The pragmatic marker *you know* in learner Englishes. *Journal of Pragmatics*, 121, 40-57. [Accessed 26 April 2025] Available from: DOI: 10.1016/j.pragma.2017.09.010

Campbell, M (1996) *Behind the Name random name generator*. [online] [Accessed 12/01/2025] Available from: https://www.behindthename.com/random/

Carter, R. & McCarthy, M. 2010. *Cambridge Grammar of English*. Cambridge: Cambridge University Press.

Chambers, J.K. 1995. *Sociolinguistic theory: linguistic variation and its social significance*. Cambridge, MA: Blackwell.

Charteris-Black, J. & Seale, C. 2010. *Gender and the Language of Illness.* London Basingstoke: Palgrave Macmillan.

Collins, L. & Hardie, A. 2022. *Making use of transcription data from qualitative research within a corpus-linguistic paradigm: Issues, experiences, and recommendations.* Corpora (Online). 17(1), pp.123–135. [Accessed 18 July 2022]. Available from: DOI: 10.3366/cor.2022.0237

Collins, P. 2009. *Modals and quasi-modals in English,* Amsterdam: Rodopi.

Cook, G. (2011) Discourse Analysis in: Simpson, J. (ed.) *The Routledge Handbook of Applied Linguistics*. Oxford: Routledge. pp.431-444

Copland, F., Creese, A., Rock, F. & Shaw, S. 2015. *Linguistic ethnography : collecting, analysing and presenting data*. Los Angeles: SAGE.

Coventry, K.R. & Garrod, S.C. 2004. *Saying, seeing, and acting : the psychological semantics of spatial prepositions*. Hove: Psychology Press.

Crawford, F. W., Aronow, P. M., Zeng, L. & Li, J. 2018. Identification of Homophily and Preferential Recruitment in Respondent-Driven Sampling. *American journal of epidemiology,* 187**,** 153-160. [Accessed 23 April 2025] Available from: DOI: 10.1093/aje/kwx208

Crawford, P., Brown, B. & Harvey, K. 2014. Corpus linguistics and evidence-based health communication. *In:* Hamilton, H. Wen-Ying, S C (ed.) *The Routledge Handbook of Language and Health Communication.* Abingdon: Routledge.

Culpeper, J. (2009). Keyness: Words, parts-of-speech and semantic categories in the character-talk of Shakespeare's Romeo and Juliet. *International journal of corpus linguistics* 14(1): pp.29-59 [Accessed 17 November 2011] Available from: DOI: 10.1075/ijcl.14.1.03cul

D'Hondt, S. & May, A. 2022. Engaging with the field while studying language in the legal process: Windows of engagement and normative moorings. *Journal of pragmatics.* 199, pp.1-5. [Accessed 19 September 2023] Available from: DOI: 10.1016/j.pragma.2022.05.012

Danesi, M. 1981. Idiolect, dialect, sociolect: What about 'psycholect'? In Gutwinski, W. & Jolly, G. (eds.), *The Eighth LACUS Forum 1981*, pp.319–325. Columbia, S.C.: Hornbeam Press. [Accessed 11 May 2025] Available from: ISBN 0-917496-22-1

David, D., Cotet, C., Matu, S., Mogoase, C. and Stefan, S., 2018. 50 years of rational-emotive and cognitive-behavioral therapy: A systematic review and meta-analysis. *Journal of clinical psychology*, *74*(3), pp.304-318. [Accessed 27 June 2025] Available from: DOI: 10.1002/jclp.22514

Davies, M. 2004. *British National Corpus.* [Online]. [Accessed 15.06.22]. Available from: https://www.english-corpora.org/bnc/

De Choudhury, M., Counts, S. & Horovitz, E. 2013. Social media as a measurement tool of depression in populations. *Proceedings of the 5th annual ACM web science conference*. pp.47-56. [Accessed 9 August 2021] Available from: DOI: 10.1145/2464464.2464480

Demiray, Ç. K. & Gençöz, T. 2018. Linguistic Reflections on Psychotherapy: Change in Usage of The First Person Pronoun in Information Structure Positions. *Journal of psycholinguistic research*, 47, pp.959-973. [Accessed 25 November 2020] Available from: DOI: 10.1007/s10936-018-9569-4

Demjén, Z. 2020. *Applying Linguistics in Illness and Healthcare Contexts*. London: Bloomsbury Publishing.

Dennett, Daniel C. (1992) The Self as a Center of Narrative Gravity. In: F. Kessel, P. Cole and D. Johnson (eds.) *Self and Consciousness: Multiple Perspectives*. Hillsdale, NJ: Erlbaum. Danish translation, *"Selvet som fortællingens tyngdepunkt,"* Philosophia 15 pp.275-88, 1986

Deuber, D. 2010. Modal Verb Usage at the Interface of English and a Related Creole: A Corpus-based Study of Can/Could and Will/Would in Trinidadian English. *Journal of English linguistics.* 38(2), pp.105-142. [Accessed 10 October 2023] Available from: DOI: 10.1177/0075424209348151

Dingemanse, M. 2024. Interjections at the Heart of Language. *Annual review of linguistics,* 10**,** 257-277. [Accessed 10 April 2025] Available from: DOI: 10.1146/annurev-linguistics-031422-124743

Dörnyei, Z. 2007. *Research methods in applied linguistics : quantitative qualitative, and mixed methodologies*. Oxford: Oxford University Press.

Douglas, F. 2009. *Scottish newspapers, language and identity*. Edinburgh: Edinburgh University Press.

Edwards, T.M. & Holtzman, N.S. 2017. A meta-analysis of correlations between depression and first person singular pronoun use*. Journal of research in personality*. 68, pp.63-68. [Accessed 25 November 2023] Available from: DOI: 10.1016/j.jrp.2017.02.005

Egbert, J. & Biber, D. 2019. Incorporating text dispersion into keyword analyses. *Corpora* 14(1): pp.77-104. [Accessed 23 January 2023] Available from: DOI: 10.3366/cor.2019.0162

Esmaeelpour, E. & Sasani, F. 2016. The Effect of Bipolar Mood Disorder on Sadegh Hedayat's Letters. *Journal of psycholinguistic research,* 45**,** pp.367-377. [Accessed 10 December 2020] Available from: DOI: 10.1007/s10936-015-9355-5

Evans, C.; Mellor-Clark, J.; Margison, F.; Barkham, M.; Audin, K.; Connell, J. & McGrath, G. 2000. *CORE: Clinical Outcomes in Routine Evaluation*. Journal of mental health (Abingdon, England), 2000-06, Vol.9 (3), p.247-255. [Accessed 9 January 2022] Available from: DOI: 10.1080/713680250

Evans, M.A. 2017. Tudor women writing: Multimodal style and identity in the English letters and prose of Queen Katherine Parr and Princess Elizabeth. *The Pragmatics and Stylistics of Identity Construction and characterisation (Studies in Variation, Contacts and Change in English).* 17, pp.1797-4453. [Accessed 8 August 2021] Available from: https://varieng.helsinki.fi/series/volumes/17/evans/

Fava, G. A. & Guidi, J. 2020. The pursuit of euthymia. *World psychiatry*, 19(1) pp.40-50 [Accessed 15 August 2021] Available from: DOI: 10.1002/wps.20698

Fludernik, M. 2003. Chronology, time, tense and experientiality in narrative. *Language and literature (Harlow, England),* 12**,** 117-134. [Accessed 25 February 2025} Available from: DOI: 10.1177/0963947003012002295

Fludernik, M. 2009. *An introduction to narratology*. London: Routledge.

Gabrielatos, C. & Marchi, A. 2012. Keyness: Appropriate metrics and practical issues. *CADS International Conference 2012*. University of Bologna: Italy.

Gage, J.T. & Perelman, C.M. 2011. *The promise of reason: studies in The new rhetoric*. Carbondale: Southern Illinois University Press.

Galasiński, D. 2008. *Men's Discourses of Depression*. 1st ed. Basingstoke: Palgrave Macmillan.

Garside, R. 1987. The CLAWS Word-tagging System. In: Garside, R.; Leech, G. & Sampson, G. (eds), *The Computational Analysis of English: A Corpus-based Approach*. London: Longman

Greer, J., Halgin, R. & Harvey, E. 2004. Global versus specific symptom attributions: predicting the recognition and treatment of psychological distress in primary care. *Journal of psychosomatic research,* 57**,** 521-527. [Accessed 26 April 2025] Available from: DOI: 10.1016/j.jpsychores.2004.02.020

Grob, R.; Schlesinger, M.; Parker, A.M.; Shaller, D.; Barre, L.R.; Martino, S.C.; Finucane, M.L.; Rybowski, L. & Cerully, J.L. 2016. Breaking Narrative Ground: Innovative Methods for Rigorously Eliciting and Assessing Patient Narratives. *Health services research*. 51(S2), pp.1248-1272. [Accessed 23 April 2021] Available from: DOI: 10.1111/1475-6773.12503

Groom, N. 2022. Critical Issues in Keyness Analysis. In: *Corpus Linguistics Summer School 2022*, 05072022, University of Birmingham.

Grundlingh, L. 2018. Exploring the possibility of using appraisal theory to determine the legitimacy of suicide notes. *Lingua*. 214, pp1-10. [Accessed 22 June 2022] Available from: DOI: 10.1016/j.lingua.2018.07.004

Haahr, M. (1998) Random.org . Available at: https://www.random.org/ [accessed multiple dates].

Halliday, M. A. K. & Matthiessen, C. M. I. M. 2004. *An introduction to functional grammar, 3rd Edition.* London: Arnold.

Halliday, M. A. K. & Matthiessen, C. M. I. M. 2013. *Halliday's introduction to functional grammar, 4th edition.* London: Routledge.

Harrington, L. 2023. Incorporating automatic speech recognition methods into the transcription of police-suspect interviews: factors affecting automatic performance. *Frontiers in Communication,* Volume 8 - 2023. [Accessed 8 August 2025] Available from; DOI: 10.3389/fcomm.2023.1165233

Harvey, K. 2012. Disclosures of depression: Using corpus linguistics methods to examine young people's online health concerns. *International journal of corpus linguistics*. 17(3), pp.349-379. [Accessed 25 October 2021] Available from: DOI: 10.1075/ijcl.17.3.03har

Heritage, F. 2022. Magical women: Representations of female characters in the Witcher video game series. *Discourse, context & media*. 49, p.100627. [Accessed 5 May 2023] Available from: DOI: 10.1016/j.dcm.2022.100627

Heritage, F. & Baker, P. 2022. How to use corpus linguistics in sociolinguistics: a case study of modal verb use, age and change over time. In: O'Keeffe, A. and
McCarthy, M. eds. *The Routledge handbook of corpus linguistics. Second edition*. ed. Abingdon: Routledge, Taylor & Francis Group.

Himmelstein, P., Barb, S., Finlayson, M. A. & Young, K. D. 2018. Linguistic analysis of the autobiographical memories of individuals with major depressive disorder. *PloS one*, 13, e0207814-e0207814. [Accessed 16 December 2020] Available from: DOI: 10.1371/journal.pone.0207814

Holden, K. B., Hall, S. P., Robinson, M., Triplett, S., Babalola, D., Plummer, V., Treadwell, H. & DiAnne Bradford, L. 2012. Psychosocial and Sociocultural Correlates of Depressive Symptoms Among Diverse African American Women. *Journal of the National Medical Association,* 104**,** pp.493-504. [Accessed 6 January 2025] Available from: DOI: 10.1016/S0027-9684(15)30215-7

Hühn, P., Meister, J. C., Pier, J. & Schmid, W. 2014. *Handbook of Narratology*. New York: De Gruyter.

Hunston, S. 2022. *Corpora in applied linguistics*. Cambridge: Cambridge University Press.

Hunt, D. & Brookes, G. 2020. *Corpus, Discourse and Mental Health*. London: Bloomsbury Publishing.

Hunt, D. & Carter, R. 2012. Seeing through The Bell Jar: Investigating Linguistic Patterns of Psychological Disorder. *The Journal of medical humanities*, 33(1), pp.27-39. [Accessed 25 October 2021] Available from: DOI: 10.1007/s10912-011-9163-3

Hunter, M. & Grant, T. 2025. Is LIWC reliable, efficient, and effective for the analysis of large online datasets in forensic and security contexts? *Applied Corpus Linguistics,* 5**,** 100118. [Accessed 3 February 2025] Available from: https://doi.org/10.1016/j.acorp.2025.100118

Huston, J.; Meier, S.; Faith, M. and Reynolds, A. 2019. Exploratory study of automated linguistic analysis for progress monitoring and outcome assessment. *Counselling and psychotherapy research*. 19(3), pp.321-328. [Accessed 26 November 2020] Available from: DOI: 10.1002/capr.12219

IBM Corp. 2020. *IBM SPSS Statistics for Windows* (Version 27.0) [Computer software]. IBM Corp

Imahori, E. 2019. Linguistic Expressions of Depressogenic Schemata. *Working papers in applied linguistics and TESOL 2019.* [Accessed 19 November 2021] Available from: DOI: 10.7916/d8-cjc9-g216

Jacobson, D. & Mustafa, N. 2019. Social Identity Map: A Reflexivity Tool for Practicing Explicit Positionality in Critical Qualitative Research*. International journal of qualitative methods.* 18, p160940691987007. [Accessed 4 August 2021] Available from: DOI: 10.1177/1609406919870075

William James 2019. *The Principles of Psychology*. Web: Laurus Book Society.

Jaworska, S. & Ryan, K. 2018. Gender and the language of pain in chronic and terminal illness: A corpus-based discourse analysis of patients' narratives. *Social science & medicine,* 215, pp.107-114. [Accessed 25 October 2021] Available from: DOI: 10.1016/j.socscimed.2018.09.002

Johnstone, L. with Boyle, M., Cromby, J.; Dillon, J.; Harper, D.; Kinderman, P.; Longden, E.; Pilgrim, D. & Read, J. 2018. *The Power Threat Meaning Framework: Overview*. London: British Psychological Society.

Jones, L. 2016. Language and Gender identities. In: Preece, S. ed. *The Routledge Handbook of Language and Identity.* Oxford: Taylor & Francis Group, pp.210- 224.

Junco, N. 2024. *Perception and production of singular They in British English*. The University of Edinburgh. [Accessed 10 June 2025] Available from: https://core.ac.uk/download/pdf/596421879.pdf

Kamens, S. R., Cosgrove, L., Peters, S. M., Jones, N., Flanagan, E., Longden, E., Schulz, S., Robbins, B. D., Olsen, S., Miller, R. & Lichtenberg, P. 2018. Standards and Guidelines for the Development of Diagnostic Nomenclatures and Alternatives in Mental Health Research and Practice. *The Journal of humanistic psychology,* 59**,** 401-427. [Accessed 5 November 2020] Available from: DOI: 10.1177/0022167818763862

Karmen, C.; Hsiung, R.C. & Wetter, T. 2015. Screening Internet forum participants for depression symptoms by assembling and enhancing multiple NLP methods. *Computer methods and programs in biomedicine*. 120(1), pp.27-36. [Accessed 28 November 2020] Available from: DOI: 10.1016/j.cmpb.2015.03.008

Kilgarriff, A.; Rychlý, P.; Smrž, P. & Tugwell, D. 2004. *The Sketch Engine*. [Online]. Available from: http://www.sketchengine.eu

King, Bryan H. 2016. Psychiatric comorbidities in neurodevelopmental disorders. *Current Opinion in Neurology* 29(2), pp. 113-117. [Accessed 18 June 2025] Available from: DOI: 10.1097/WCO.0000000000000299

Koester, A. 2010. Building small specialised corpora in O'Keeffe, A., McCarthy, M. (Eds.), *The Routledge Handbook of Corpus Linguistics*. Abingdon, UK: Routledge

Kruger, J., Wirtz, D., Miller, D. T. & Devine, P. 2005. Counterfactual Thinking and the First Instinct Fallacy. *Journal of personality and social psychology,* 88**,** pp.725-735. [Accessed 21 November 2024] Available from: DOI: 10.1037/0022-3514.88.5.725

Kyle, K. 2019. Measuring Lexical Richness. In: Stuart, W. (ed.) *The Routledge Handbook of Vocabulary Studies*. London: Routledge.

Labov, W. 2013. *The Language of Life and Death: The Transformation of Experience in Oral Narrative*. Cambridge: Cambridge University Press

Lakoff, R.T. & Bucholtz, M. 2004. *Language and woman's place.* Revised and expanded edition. ed. New York: Oxford University Press.

Lazard, L. & McAvoy, J. 2017. Doing reflexivity in psychological research: What's the point? What's the practice? *Qualitative research in psychology*. Vol.17 (2), pp.159-177. [Accessed 14 June 2022] Available from: DOI: 10.1080/14780887.2017.1400144

Le Glaz, A., Haralambous, Y., Kim-Dufor, D.-H., Lenca, P., Billot, R., Ryan, T. C., Marsh, J., Devylder, J., Walter, M., Berrouiguet, S. & Lemey, C. 2021. Machine Learning and Natural Language Processing in Mental Health: Systematic Review. *Journal of medical Internet research,* 23**,** e15708-e15708. [Accessed 1 April 2025] Available from: DOI: 10.2196/15708

Leach, L. S., Christensen, H. & Mackinnon, A. J. 2008. Gender Differences in the Endorsement of Symptoms for Depression and Anxiety: Are Gender-Biased Items Responsible? *The journal of nervous and mental disease,* 196**,** 128-135. [Accessed 4 January 2021] Available from DOI: 10.1097/NMD.0b013e318162aa63

Leech, G. N. 2004 *Meaning and the English Verb*. London: Pearson Education

Lorenzo-Dus, N. 2021. 'It's the subtle language that gets to you'. Understanding and managing researcher exposure to online child sexual grooming content. In: *Focus on the researcher: dealing with depressing data seminar*, 14.05.2021: BAAL SIG: Language and New Media.

Leis, A.; Ronzano, F.; Mayer, M. A.; Furlong, L. I. & Sanz, F. 2019. Detecting Signs of Depression in Tweets in Spanish: Behavioral and Linguistic Analysis. *Journal of medical Internet research*, 21, e14199-e14199. [Accessed 18 November 2020] Available from: DOI: 10.2196/14199

Lian, M., 2018. Analysis of Dickens' Critique and Humanity Spirit in Oliver Twist Based on the Appraisal Theory. *Theory and Practice in Language Studies*, *8*(8), pp.1050-1054. [Accessed 2 June 2025] Available from: DOI: http://dx.doi.org/10.17507/tpls.0808.19

Love, R., Dembry, C., Hardie, A., Brezina, V. & McEnery, T. 2017. The Spoken BNC2014: Designing and Building a Spoken Corpus of Everyday Conversations. *International Journal of Corpus Linguistics,* 22**,** 319-344. [Accessed 24 November 2021] Available from: DOI: 10.1075/ijcl.22.3.02lov

Macken-Horarik, M. & Isaac, A. R. 2014. Appraising Appraisal. *In:* Thompson, G. & Alba-Juez, L. (eds.) *Evaluation in Context.* The Netherlands: John Benjamins Publishing Company.

Madill, A. & Gough, B. 2008. Qualitative research and its place in psychological science. *Psychological Methods*. 13(3) pp.254-271. [Accessed 25 May 2023] Available from: DOI: 10.1037/a0013220

Maji, S. 2018. Society and ‘good woman’: A critical review of gender difference in depression. *International Journal of Social Psychiatry*, 64, pp.396-405. [Accessed 23 December 2020] Available from: DOI: 10.1177/0020764018765023

Martin, J. R. 2004. Sense and sensibility: texturing evaluation. *In:* Joseph, F. (ed.) *Language, Education and Discourse : Functional Approaches.* London: Bloomsbury Publishing.

Martin, J.R. & Rose, D. 2003. *Working with Discourse: Meaning beyond the Clause*. London: Continuum.

Martin, J. R. & White, P. R. R. 2007. *The language of evaluation : appraisal in English*. Basingstoke: Palgrave Macmillan.

Maslow, A. H. 1943. A theory of human motivation. *Psychological review,* 50**,** 370-396. [Accessed 22 March 2025] Available from: DOI: 10.1037/h0054346

Matthiessen, C. M & Halliday, M.A.K. 1997 Systemic Functional Grammar: A First Step into the Theory. [online] Available at: https://www.alvinleong.info/sfg/sfgintro.html

McCarthy, P. M. & Jarvis, S. 2010. MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior research methods,* 42**,** 381-392. [Accessed 24 March 2025] Available from: DOI: 10.3758/BRM.42.2.381

McEnery, T. & Hardie, A. 2011. *Corpus Linguistics: Method, Theory and Practice*. Cambridge: Cambridge University Press.

McLeod, S. 2023. *Psychological Theories Of Depression*. [Online]. [Accessed 18 April 2023]. Available from: https://www.simplypsychology.org/depression.html

Molendijk, M.L.; Bamelis, L.; van Emmerik, A.A.P.; Arntz, A.; Haringsma, R. & Spinhoven, P. 2010. Word use of outpatients with a personality disorder and concurrent or previous major depressive disorder. *Behaviour research and therapy*. 48(1), pp.44-51. [Accessed 26 November 2020] Available from: DOI: 10.1016/j.brat.2009.09.007

Moors, A. 2017. Integration of two skeptical emotion theories: Dimensional appraisal theory and Russell's psychological construction theory. *Psychological Inquiry,* 28**,** 1-19. [Accesed 1 June 2025] Available from: http://dx.doi.org.10.1080/1047840X.2017.1235900

Morgenroth, T. & Ryan, M. K. 2021. The Effects of Gender Trouble: An Integrative Theoretical Framework of the Perpetuation and Disruption of the Gender/Sex Binary. *Perspectives on Psychological Science,* 16**,** 1113-1142. [Accessed 1 September 2024] Available from: DOI: 10.1177/1745691620902442

Motschenbacher, H. 2018. Corpus linguistics in language and sexuality studies: Taking stock and looking ahead. *Journal of Language and Sexuality,* 7**,** 145-174. [Accessed 3 October 2018] Available from: DOI: 10.1177/0957926513486069

Neuman, Y., Cohen, Y., Assaf, D. & Kedma, G. 2012. Proactive screening for depression through metaphorical and automatic text analysis. *Artificial intelligence in medicine,* 56**,** 19-25. [Accessed 9.08.21]. Available from: DOI: 10.1016/j.artmed.2012.06.001

Newell, E. E., McCoy, S. K., Newman, M. L., Wellman, J. D. & Gardner, S. K. 2018. You Sound So Down: Capturing Depressed Affect Through Depressed Language. *Journal of language and social*

psychology, 37**,** 451-474. [Accessed 23 November 2020] Available from: DOI: 10.1177/0261927X17731123

NHS Digital statistics. 2021. *Health Survey England Additional Analyses - Health and health-related behaviours of Lesbian, Gay and Bisexual adults.* [Online]. [Accessed 10.01.22]. Available from: https://digital.nhs.uk/news/2021/lgb-health-statistics

O'Donnell, M. 2008. The UAM CorpusTool: Software for corpus annotation and exploration. In Callejas, C.M.B., Sánchez, J.F.F., Ibáñez, J.R., Sánchez, M.E.G., Cortés de los Ríos, E., Ramiro, S.S., Martínez, S.C., Honeyman, N.P. & Márquez, B.C. (eds) *Applied Linguistics Now: Understanding Language and Mind* / La Lingüística Aplicada Hoy: Comprendiendo el Lenguaje y la Mente, 1433–1447. Almería: Universidad de Almería

O'Donnell, M. 2014. Exploring identity through Appraisal Theory: A corpus annotation methodology. *Linguistics and the Human Sciences*, 9, pp.1-15. [Accessed 12 August 2024] Available from: DOI: 10.1558/lhs.v9i1.95

O'Keeffe, A. & McCarthy, M. 2010. *The Routledge handbook of corpus linguistics*. Abingdon: Routledge.

Oates, J.C., D.; Fisher, M.; Goodson, S.; Beth, H.; Kwiatkowski, R.; Prutton, K.; Reeves, D.; Wainwright, T. 2021. *BPS Code of Human Research Ethics*. The British Psychological Society.

Odea, B.; Boonstra, T.W.; Larsen, M.E.; Nguyen, T.; Venkatesh, S. & Christensen, H. 2018. The relationship between linguistic expression and symptoms of depression, anxiety, and suicidal thoughts: A longitudinal study of blog content*. PLoS One.*16(5), e0251787 [Accessed 27 October 2021] Available from: DOI: 10.48550/arxiv.1811.02750

Orvell, A., Galmen, S. & Kross, E. (2022). "What "you" and "we" say about me: How small shifts in language reveal and empower fundamental shifts in perspective." *Social and personality psychology compass* 16(5) [Accessed 26 July 2022] Available from: DOI: 10.1111/spc3.12665

Otte, C. 2008. Incomplete remission in depression: role of psychiatric and somatic comorbidity. *Dialogues in Clinical Neuroscience* 10 (4), pp.453-460. [Accessed 26 October 2022] Available from: DOI: 10.31887/DCNS.2008.10.4/cotte

Oxford English Dictionary. 2023. Cisgender adj. & n. *Oxford English Dictionary.* Oxford: Oxford University Press.

Oxford English Dictionary. 2023. Easy, adj., adv., int., n. *Oxford English Dictionary.* Oxford: Oxford University Press.

Oxford English Dictionary. 2023. Just, adv. *Oxford English Dictionary.* Oxford: Oxford University Press.

Oxford English Dictionary. 2023. Would, v., n. *Oxford English Dictionary.* Oxford: Oxford University Press.

Oxford English Dictionary, 2024. Anhedonia (n.), Etymology. *Oxford English Dictionary.* Oxford: Oxford University Press.

Oxford English Dictionary. 2024. Change, n., v. *Oxford English Dictionary.* Oxford: Oxford University Press.

Oxford English Dictionary. 2024. Linguistics, n. *Oxford English Dictionary.* Oxford: Oxford University Press.

Oxford English Dictionary. 2024. phenomenology, n., sense 2. *Oxford English Dictionary.* Oxford: Oxford University Press.

Oxford English Dictionary. 2024. speech act (n.), sense 2. *Oxford English Dictionary.* Oxford: Oxford University Press.

Oxford English Dictionary. 2025. be, v. *Oxford English Dictionary.* Oxford: Oxford University Press.

Oxford English Dictionary. 2025. have, v. *Oxford English Dictionary.* Oxford: Oxford University Press.

Park, A. & Conway, M. 2018. Harnessing Reddit to Understand the Written-Communication Challenges Experienced by Individuals With Mental Health Disorders: Analysis of Texts From Mental Health Communities. *Journal of medical Internet research,* 20**,** e121-e121. [Accessed 24 March 2025] Available from: DOI: 10.2196/jmir.8219

Park, G., Schwartz, H. A., Sap, M., Kern, M. L., Weingarten, E., Eichstaedt, J. C., Berger, J., Stillwell, D. J., Kosinski, M., Ungar, L. H. & Seligman, M. E. P. 2017. Living in the Past, Present, and Future: Measuring Temporal Orientation With Language. *Journal of personality,* 85**,** 270-280. [Accessed 16 December 2020} Available from: DOI: 10.1111/jopy.12239

Parker, G. & Brotchie, H. 2010. Gender differences in depression. *International review of psychiatry (Abingdon, England),* 22**,** 429-436. [Accessed 28 August 2024] Available from: DOI: 10.3109/09540261.2010.492391

Pennebaker, J.W. 2011. *The secret life of pronouns: what our words say about us.* New York: London: Bloomsbury.

Pennebaker, J. W.; Boyd, R. L.; Booth, R. J.; Ashokkumar, A. & Francis, M. E. 2022. *Linguistic Inquiry and Word Count: LIWC-22*. Pennebaker Conglomerates

PHE 2020. Public Health England Common Mental Health Disorders Prevalence. *Public Health England*. Retrieved January 10, 2022 from https://fingertips.phe.org.uk/profile-group/mental-health/

Pilgrim, D. 2020. A Critical Realist Reflection on the Power Threat Meaning Framework. *Journal of constructivist psychology*, 12(1) pp.1-13. [Accessed 27 October 2021] Available from: DOI: 10.1080/10720537.2020.1773359

Pintea, S. & Gatea, A. 2021. The Relationship Between Self-Silencing and Depression: A Meta-Analysis. *Journal of social and clinical psychology,* 40**,** 333-358. [Accessed 18 March 2021] Available from: DOI: 10.1521/jscp.2021.40.4.333

Power, M. J. 2004. *Mood disorders a handbook of science and practice*. New Jersey: J. Wiley.

Quevedo, J. C.; Ferrer, A. & Zarate, C. A. 2019. *Neurobiology of depression: road to novel therapeutics*, Cambridge MA: Academic Press, An imprint of Elsevier.

Rae, J. P. 1993. Can Causal Connectives be Used to Find Causal Expressions in Naturally Occurring Discourse? *Journal of language and social psychology*, 12, pp.239-245. [Accessed 13 August 2024] Available from: DOI: 10.1177/0261927X93123005

Rawlings, G.H. & Bains, M. 2020. Experiences of depression in older adults from non-white and ethnic minority groups: A thematic synthesis of qualitative studies. *Journal of affective disorders*. 266, pp.341-348. [Accessed 28 July 2022] Available from: DOI: 10.1016/j.jad.2020.01.165

Rayson, P. 2009. *Wmatrix: a web-based corpus processing environment*, Computing Department, Lancaster University. Available at [http://ucrel.lancs.ac.uk/wmatrix/]

Reuneker, A. (2017). *Lexical Diversity Measurements*. [Accessed 24 March, 2025] Available from: https://www.reuneker.nl/files/ld.

Rezaii, N., Walker, E. & Wolff, P. 2019. A machine learning approach to predicting psychosis using semantic density and latent content analysis. *NPJ schizophrenia*, 5, pp.9-12. [Accessed 30 July 2024] Available from: DOI: 10.1038/s41537-019-0077-9

Roberts, L. M. & Christens, B. D. (2021). Pathways to Well-being among LGBT adults: Sociopolitical Involvement, Family Support, Outness, and Community Connectedness with Race/Ethnicity as a Moderator. *American Journal of Community Psychology*, 67, pp.405-418. [Accessed 2 October 2024] Available from: DOI: 10.1002/ajcp.12482

Robertson, D. 2010. *The philosophy of cognitive-behavioural therapy (CBT): stoic philosophy as rational and cognitive psychotherapy.* Oxford: Taylor & Francis Group

Robinson, N. & Villasenor, V. (eds.) *Desire, Love, Identity: From the Nottinghamshire LGBTQ community.* Nottingham: Global Worlds Press

Rosenbach, C. & Renneberg, B. 2014. Remembering rejection: Specificity and linguistic styles of autobiographical memories in borderline personality disorder and depression. *Journal of behavior*

therapy and experimental psychiatry. 46, pp.85-92. [Accessed: 19 November 2021] Available from: DOI: 10.1016/j.jbtep.2014.09.002

Scott, M. 2020. *WordSmith Tools*. 2010. Stroud: Lexical Analysis Software.

Semino, E.; Demjén, Z. & Demmen, J. 2018. An Integrated Approach to Metaphor and Framing in Cognition, Discourse, and Practice, with an Application to Metaphors for Cancer. *Applied linguistics*. 39(5), pp.625-645. [Accessed 13 July 2023] Available from: DOI: 10.1093/applin/amw028

Simpson, J. (ed.) 2011. *The Routledge Handbook of Applied Linguistics*. Oxford: Routledge.

Simpson, P. 1993. *Language, ideology and point of view*. Abingdon: Routledge.

Sims-Schouten, W. & Riley, S. 2019. Presenting Critical Realist Discourse Analysis as a Tool for Making Sense of Service Users' Accounts of Their Mental Health Problems. *Qualitative health research,* 29**,** 1016-1028. [Accessed 9 April 2021] Available from: DOI: 10.1177/1049732318818824

Şimşek, Ö.F. 2013. The Relationship Between Language Use and Depression: Illuminating the Importance of Self-Reflection, Self-Rumination, and the Need for Absolute Truth. *The Journal of general psychology.* 140(1), pp.29-44. [Accessed 16 December 2020] Available from: DOI: 10.1080/00221309.2012.713407

Şimşek, Ö.F. & Kuzucu, Y. 2012. The gap that makes us desperate: Paths from language to mental health. *International journal of psychology.* 47(6), pp.467-477. [Accessed 15 December 2020] Available from: DOI: 10.1080/00207594.2011.645479

Sinclair, J. 2004. *Trust the Text : Language, Corpus and Discourse*. 1st ed. ed. Carter, R. Florence: Taylor and Francis.

Smail, D. (2005). *Power, interest and psychology: Elements of a social materialist understanding of distress*. Ross-on-Wye: PCCS Books.

Smirnova, D.; Cumming, P.; Sloeva, E.; Kuvshinova, N.; Romanov, D. & Nosachev, G. 2018. Language Patterns Discriminate Mild Depression From Normal Sadness and Euthymic State.

Frontiers in psychiatry. 9, pp.105-105. [Accessed 22 March 2021] Available from: DOI: 10.3389/fpsyt.2018.00105

Smith, K. 2019. *Facebook statistics, Brandwatch*. Available at: https://www.brandwatch.com/blog/facebook-statistics [Accessed: 22 April 2023].

Smith, N. 2020. Conversationalization and democratization in a radio chat show: a grammar-led investigation. *Language Sciences*, 79, pp.101-269. [Accessed 12 January 2021] Available from: DOI 10.1016/j.langsci.2020.101269

Smith, N. & Waters, C. 2019. Variation and change in a specialized register: A comparison of random and sociolinguistic sampling outcomes in Desert Island Discs. *International journal of corpus linguistics*. 24(2), pp.169-201. [Accessed 26 January 2023] Available from: DOI: 10.1075/ijcl.17117.smi

Soames, S. 1984. Linguistics and Psychology. *Linguistics and philosophy*, 7(2), pp.155-179 [Accessed 4 November 2021] Available from: DOI: 10.1007/BF00630811

Solomon, A. 2001. *The noonday demon: An atlas of depression*, New York: Scribner/Simon & Schuster.

Starr, L. R., Davila, J. & Watson, D. 2008. Excessive Reassurance Seeking, Depression, and Interpersonal Rejection: A Meta-Analytic Review. *Journal of abnormal psychology,* 117**,** pp.762-775. [Accessed 6 January 2025] Available from: DOI: 10.1037/a0013866

Stratton, P. 1988*. Leeds attributional coding system manual*. Leeds: Leeds Family Therapy & Research Centre.

Stratton, P. 2003. Causal attributions during therapy I: Responsibility and blame. *Journal of family therapy*, 25, pp.136-160. [Accessed 6 June 2024] Available from: DOI: 10.1111/1467-6427.00241

Stockard, J. 2006. Gender Socialization. *In:* Chafetz, J. (ed.) *Handbook of the sociology of gender.* New York: Springer.

Svalberg, A. M. L. 2019. Mapping tense form and meaning for L2 learning – from theory to practice. *International Review of Applied Linguistics in Language Teaching*, 57, pp.417-445. [Accessed 18 April 2021] Available from: DOI: 10.1515/iral-2016-0105

Sweeney, P. D., Anderson, K., & Bailey, S. (1986). Attributional style in depression: A meta-analytic review. *Journal of Personality and Social Psychology, 50*(5), 974–991 [Accessed 28 January 2025] Available from: DOI: 10.1037/0022-3514.50.5.974

Szudarski, P. 2018. *Corpus Linguistics for Vocabulary: A Guide for Research*. Abingdon: Routledge.

Taavitsainen, I. 1998. Emphatic language and romantic prose: Changing functions of interjections in a sociocultural perspective. *European journal of English studies*. **2**(2), pp.195–214. [Accessed 11 August 2025] Available from: DOI: 10.1080/13825579808574413

Tackman, A.M.; Sbarra, D.A.; Carey, A.L.; Donnellan, M.B.; Horn, A.B.; Holtzman, N.S.; Edwards, T.M.S.; Pennebaker, J.W. & Mehl, M.R. 2019. Depression, Negative Emotionality, and Self-Referential Language: A Multi-Lab, Multi-Measure, and Multi-Language-Task Research Synthesis. *Journal of personality and social psychology*. 116(5), pp.817-834. [Accessed 24 November 2020] Available from: DOI: 10.1037/pspp0000187

Tadesse, M.M.; Lin, H.; Xu, B. & Yang, L. 2019. Detection of Depression-Related Posts in Reddit Social Media Forum. *IEEE access*. 7, pp.44883-44893. [Accessed 24 November 2020] Available from: DOI: 10.1109/access.2019.2909180

Thornicroft, G., Sunkel, C., Alikhon Aliev, A., Baker, S., Brohan, E., EL Chammay, R., Davies, K., Demissie, M., Duncan, J., Fekadu, W., Gronholm, P. C., Guerrero, Z., Gurung, D., Habatamu, K., Hanlon, C., Heim, E., Henderson, C., Hijazi, Z., Hoffman, C., Hosny, N., Huang, F.-X., Kline, S., Kohrt, B. A., Lempp, H., Li, J., London, E., Ma, N., Mak, W. W. S., Makhmud, A., Maulik, P. K., Milenova, M., Morales Cano, G., Ouali, U., Parry, S., Rangaswamy T., Rüsch, N., Sabri, T., Sartorius, N., Schulze, M., Stuart, H., Taylor Salisbury, T., Vera San Juan, N., Votruba, N. & Winkler, P. 2022. The Lancet Commission on ending stigma and discrimination in mental health.

The Lancet (British edition), 400**,** pp.1438-1480. [Accessed 20 January 2025] Available from: DOI: 10.1016/S0140-6736(22)01470-2

Tirvassen, R. 2019. *Sociolinguistics and the narrative turn : researching languages and society in contexts of change and transition*. Leiden, The Netherlands: Brill.

Tishler, C. L. & Bartholomae, S. 2002. The recruitment of Normal Healthy Volunteers: A Review of The Literature on the Use of Financial Incentives. *Journal of Clinical Pharmacology,* 42(4), pp.365-375. [Accessed: 23 November 2021] Available from: DOI: 10.1177/00912700222011409

Toolan, M.J. 1988. *Narrative : a critical linguistic introduction.* London: Routledge.

Traxler, M. J. 2012. *Introduction to psycholinguistics : understanding language science*. Chichester: Wiley-Blackwell.

Trifu, R., Nemes, B., Bodea-Haţegan, C. & Cozman, D. 2017. Linguistic indicators of language in major depressive disorder (MDD). An evidence based research. *Journal of Evidence-Based Psychotherapies,* 17**,** 105-128. [Accessed 27 May 2021] Available from: DOI: 10.24193/jebp.2017.1.7

Tsuchiya, K., Coffey, F. & Nakamura, K. 2023. *Multimodal Approaches to Healthcare Communication Research : Visualising Interactions for Resilient Healthcare in the UK and Japan*. London: Bloomsbury Publishing.

Umney, C. 2018. *Class Matters Inequality and Exploitation in 21st Century Britain*. London: Pluto Press

Ungless, E. L., Ross, B. & Belle, V. 2023. Potential pitfalls with automatic sentiment analysis: The example of queerphobic bias. *Social science computer review*. 41, pp.2211-2229. [Accessed 6 September 2024] Available from: DOI: 10.1177/08944393231152946

Vaughan, A. 2020. Identifying linguistic features of suicidal ideation in unthematic content on Reddit. *English Language Seminar University of Leicester.* University College London.

Wales, K. 1996. *Personal pronouns in present-day English*. Cambridge: Cambridge University Press.

Watkins, E. 2004. Adaptive and maladaptive ruminative self-focus during emotional processing. *Behaviour research and therapy*, 42(9), pp.1037-1052. [Accessed 26 October 2022] Available from: DOI: 10.1016/j.brat.2004.01.009

Weerasinghe, J., Morales, K. & Greenstadt, R. 2019. “Because... I was told... so much”: linguistic indicators of mental health status on Twitter. *Proceedings on Privacy Enhancing Technologies*. [Accessed 1 April 2025] Available from: https://petsymposium.org/popets/2019/popets-2019-0063.pdf

Weetman, K.; Dale, J.; Scott, E. & Schnurr, S. 2020. Adult patient perspectives on receiving hospital discharge letters: a corpus analysis of patient interviews. *BMC health services research*, 20(1), pp.1-537. [Accessed 25 October 2021] Available from: DOI: 10.1186/s12913-020-05250-1

Wen, B., Wang, N., Subbalakshmi, K. and Chandramouli, R., 2023. Revealing the roles of part-of-speech taggers in Alzheimer disease detection: scientific discovery using one-intervention causal explanation. *JMIR Formative Research*, *7*(1), p.e36590. [Accessed 4 June 2025] Available from: DOI: 10.2196/36590

WHO 2022. International Classification of Diseases Eleventh Revision (ICD-11). Geneva: World Health Organization

WHO 1993. *The ICD-10 classification of mental and behavioural disorders: diagnostic criteria for research.* Geneva: World Health Organization.

Wright, B.; Dave, S. & Dogra, N. 2017. *100 Cases in Psychiatry*. Oxford: Taylor & Francis Group

Ye, Z., 2021. Comparing the distributions of attitudinal resources of news discourses on London and Wuhan lockdown during COVID-19 outbreak based on Appraisal Theory. *Proceedings of IConSES*, pp.160-169. [Accessed 2 June 2025] Available from: ISBN: 978-1-952092-23-7

Zimman, L. 2021. Gender diversity and the voice In Eds. Angouri, J. & Baxter, J. (2021*) The Routledge handbook of language, gender and sexuality*. Abingdon: Routledge. pp.69-89.

Appendix A

Pilot study participant information sheet

Research Project Title: Pilot corpus linguistic analysis of the language of people who have symptoms of depression

Name, position and contact details of researcher: Kate Sayers, PhD researcher ks589@le.ac.uk

Name, position and contact details of supervisor: Redacted

You are invited to an interview to talk about the things that are important to you, and then a follow up questionnaire. Before you decide whether or not to take part, it is important for you to understand why the research is being done and what it will involve. Please take the time to read the following information carefully, you are welcome to discuss this project with others before you make your decision.

What is the purpose of the research project?

This project aims to better understand the language used by people experiencing the symptoms of depression when compared to those who are not experiencing these symptoms. This is a pilot study to make sure the interviews run smoothly and to start comparing the words used in the interviews themselves. You have been invited to participate in one of these interviews and then a follow up questionnaire to find out if you are experiencing the symptoms of depression.

Who has been invited to participate?

Anybody studying at the University of Leicester has been invited to participate, since it is important to interview people both with and without the symptoms of depression. In exchange for your time, you will be entered into a prize draw to win £50.

Do I have to take part?

Your participation in this interview and follow up questionnaire is entirely voluntary. Your contribution would help us to be able to help people who may be experiencing the symptoms of depression. If you do decide to take part you can keep this information sheet along with a privacy notice that will explain how your data will be collected and used. You will also be asked to provide your consent to take part. If you do decide to take part, you are able to withdraw from the project at any time, without providing a reason.

What will happen if I take part?

If you decide to take part, it is expected that this project will take up to half an hour of your time. Kate will call you individually over Microsoft teams and ask you about what is important to you.

The audio only of this call will be recorded so that Kate can write it up later. Once you have finished this conversation, you will be asked to fill in the CORE-OM wellbeing categorisation questionnaire right away. This questionnaire will show if you are experiencing the symptoms of depression. Once you have finished the questionnaire, you will receive a debriefing sheet including information on some of the support available for people with the symptoms of depression.

The written down version of your interview will be analysed using corpus linguistic techniques. This will help us to find out if people experiencing the symptoms of depression use different words or types of words when compared to people who are feeling less depressed.

What are the possible disadvantages and risks of taking part?

The CORE-OM wellbeing questionnaire will include questions about your mental health, which you may find distressing. Contact information for appropriate support groups will be provided in the debrief. If your score from this questionnaire is outside of a comfortable level, you will be confidentially told this and given advice on where to get help. If the researcher fears that you may cause harm to yourself or others then they will contact student support services on your behalf and for your protection.

What are the possible benefits of taking part?

The interview will be an opportunity to talk about what is important to you, with no interruption, judgement or expectation. This information combined with your questionnaire results will go towards a research project that aims to help people who experience the symptoms of depression.

What data will you collect about me?

During the interview, your words will be recorded and alongside the wellbeing questionnaire you will be asked your age, ethnicity and gender. This information will be anonymised, meaning that your personal details will be removed and replaced with a code. Your interview will be recorded and typed up by Kate, and nobody else will be able to see the information that shows who you are. Whilst typing up your information, anything individual to you that could allow other people to work out which interview was yours will be changed.

Will what I say in this research project be kept confidential?

All information we collect about you will be kept confidential (subject to legal limitations). Your name will be removed from the transcription as well as the questionnaire, and a code used instead. The recording of the interview, your personal details and the transcription will be kept on a secure university server, which is password protected.

How will you look after the data you collect about me?

We need to ensure that you understand what will happen to data we collect about you as well as your legal rights. This document is accompanied with a separate Privacy Notice which provides further details.

Your normal rights under the Data Protection Act and the General Data Protection Regulation apply. At all times this research study will comply with the General Data Protection Regulations

(GDPR, 2018) approved by the EU parliament on 14 April 2016 and passing into UK law on 25 May 2018. However, we need to manage your records in specific ways for the research project to be reliable. This means that we won't always be able to let you see or change the data we hold about you.

You can stop being a part of the research project, without giving a reason, at any time, up to and including August 2021. However, we will keep information about you that we already have at the point of withdrawal, and continue to use this for the purposes of the research project as outlined here.

The transcription of the interview and your questionnaire will be securely stored in two separate files using an anonymised code. If you say anything that could be used to identify yourself or others during the recording, then this will also be anonymised in the transcription. The original recording of the meeting will be destroyed 6 years after the end of the project (i.e. September 2030), but the anonymised transcription and background information will continue to be stored securely beyond this time. We will keep a separate secure file with a 'key' that links your personal details to your anonymous number code, in case we need to contact you about anything important related to the safety of you or others. Only Kate will have access to this 'key' file, and it will be destroyed 12 months after the end of the project (i.e. September 2025). After that there will be no possible way to link the data you have given back to you.

All data is kept in accordance with the University's Research Code of Practice. It will be kept electronically on the university's secure servers, with password protection.

What will happen to the results of the research project?

The results will be used as part of a PhD probation review, eventually contributing towards a PhD thesis. All participants are welcome to request a copy of these publications from the researcher. It will not be possible to identify you from the data included in these research papers.

What should I do if I want to take part?

To participate, you will need to complete the Informed Consent From and opt-in to a variety of research options by ticking the Yes/No boxes. This will confirm that you understand how your data will be processed, protected and reviewed for research purposes.

Who is organising and funding the research project?

The organiser of the project is Kate Sayers, School of Arts, College of Social Science, Arts and History. The project is self-funded.

What if something goes wrong?

If you are unable to attend the interview at the scheduled time, please inform Kate using the contact details provided as soon as you are able. If you become uncomfortable at any point during the interview or questionnaire then please feel free to excuse yourself from the project. Contact information for professionals who can help with negative thoughts and feelings will be provided in the debrief.

Who has reviewed the research project?

The project has been reviewed and approved by the University of Leicester's Research Ethics Committee.

Contact Information

For information about the project, please contact Kate Sayers (ks589@le.ac.uk)

If you have any concerns or queries about how the research has been conducted, please contact the Chair of the University Research Ethics Committee

If you require more GDPR data protection information then you can access this via the University's Information Assurance Services (IAS):

Information Assurance Services

University of Leicester
University Road
Leicester
LE1 7RH
+44 (0)116 229 7945
Email: ias@le.ac.uk
Website: www2.le.ac.uk/offices/ias

Thank you for taking the time to read this participant information sheet.

Pilot study privacy notice

Privacy Notice for Research Participants

Pilot corpus linguistic analysis of the language of people who have symptoms of depression

Research Contact: Kate Sayers (ks589@le.ac.uk)

This Privacy Notice provides information about how the University of Leicester collects and uses your personal information when you take part in this research project.

Please also refer to the **Participant Information Sheet** given to you for further details about the research project, what information will be collected about you, and how it will be used.

The University of Leicester will usually be the Data Controller of any data that you supply for this research. This means that we are responsible for looking after your information and using it properly. This means that the University will make the decisions on how your data is used and for what reasons. You can access more information on this via the University's Information Assurance Services:

Information Assurance Services
University of Leicester

University Road
Leicester
LE1 7RH
T: +44 (0)116 229 7945
E: ias@le.ac.uk
W: https://www2.le.ac.uk/offices/ias

Why do we need your data?

The study investigates the language of people experiencing the symptoms of depression when compared to the language of people who are not experiencing the symptoms of depression. Your data – the answer to your categorisation and wellbeing questionnaire as well as your interview - will be used as part of this investigation.

What type of data will the University of Leicester use?

The researcher will use anonymised transcriptions of your interview for their analysis, alongside the details you provide afterwards in the categorisation questionnaire. There will also be the original recording of the interview which will be used to create the transcriptions. The recordings will not be used for any other purpose. There will also be an anonymised personal details document, based on the information you provide when completing the consent form.

Who will the University of Leicester share your data with?

Your data will not be shared with any other contact within or external to the University of Leicester including data sharing outside of the UK.

What rights do I have regarding my data held by the University of Leicester?

Your normal rights under the Data Protection Act and the General Data Protection Regulation apply.

You can stop being part of the research project at any time, without giving a reason, but we will keep information about you that we already have and continue to use this for the purposes of the research project as outlined in the Participant Information Sheet.

Where did the University of Leicester source my data from?

You provided your contact details in order to receive this notice and the rest of the data will be part of the questionnaire you complete after your interview.

Are there any consequences of not providing the requested data?

There are no consequences of not providing data for this research. It is purely voluntary.

Will there be any automated decision making using my data?

There will be no use of automated decision making in scope of UK Data Protection and Privacy legislation.

How long will the University of Leicester keep my data?

In line with the law, we will only keep your data for as long as we need to so that we can fulfil our research objectives.

We will keep your personal data, such as your name and email address until we have completed all the actions that require us to hold them, for example sending you a copy of the results of the study if you have requested this. The recordings and the transcriptions of the interview will be destroyed six years following the PhD completion (September 2030).

Who can I contact if I have concerns?

In the event of any questions about the research project, please contact the researcher in the first instance.

If you have any concerns about the way in which the research project has been conducted, please contact the Chair of the University Research Ethics Committee.

The University of Leicester Data Protection Officer is: Redacted

For further details about information security, please contact the Information Assurance Services team.

Pilot study consent form

Full title of project: A corpus linguistic and stylistic analysis of the language of people who have symptoms of depression, concentrating on temporally prior foci alongside practical diagnostic or therapeutic applications.

Name, position and contact details of researcher: Kate Sayers, PhD researcher Redacted

Name, position and contact details of supervisor: Mel Evans, Lecturer, Redacted

		Please **initial** box
1	I confirm that I have read and understand the participant information sheet (Version 1.0, 01.05.21) for the above study and have had the opportunity to ask questions.	☐
2	I understand that my participation is voluntary and that I am free to withdraw at any time, without giving reason	☐
3	I understand that at all times this research project will comply with the General Data Protection Regulations (GDPR, 2018) approved by the EU parliament on 14 April 2016 and passing into UK law effective from 25 May 2018 and that if I have any concerns how I contact the University of Leicester to raise these.	☐

4 I understand the interview will be audio recorded ☐

Yes No

5 I wish to receive a copy of the results of this research project, and I agree for my contact details to be retained and used for this purpose ☐ ☐

6 I agree to take part in the above research project ☐

Name of Participant	Date	Signature
Name of researcher obtaining consent	Date	Signature

Appendix B

Main study Participant information sheet

Research project title: A linguistic analysis of narratives from people with low mood.

Researcher: Kate Sayers, supervised by Dr Mel Evans

You are invited to an interview to talk about what is important to you, and then to complete a follow up questionnaire. Before you decide whether or not to take part, it is important for you to understand why the research is being done and what it will involve. Please take the time to read the following information carefully. You are welcome to discuss this project with others before you make your decision.

What is the purpose of the project?

This project aims to help us to better understand the language used by people who have low mood, by comparing their language with that of people who are not feeling low. You have been invited to be interviewed and then fill in a follow up questionnaire about you and your health and wellbeing.

Why have I been chosen?

You have been chosen because you expressed an interest in the project over social media, and you are one of fifty people in total who could be helping in this PhD research project.

Do I have to take part?

Your involvement with this project, both the interview and follow up questionnaire, is entirely voluntary. You can keep this information sheet along with a privacy notice that will explain how your answers will be collected and used. You will also be asked to provide your consent to be involved. If you do decide to take part, you can change your mind and withdraw from the project at any time up to 1st December 2022, and you don't need to give a reason for this.

What will happen to me if I take part?

If you decide to take part, it is expected that this project will take around thirty minutes of your time. Kate will call you individually over Microsoft Teams, Zoom or Facebook messenger and ask you about what is important to you. This call will be recorded so that Kate can write it up later. Once you have finished this conversation, you will be asked to fill in a questionnaire asking for more information about you as well as a Mood and Feelings Questionnaire right away. Once you

have finished the questionnaire, you will see some information about support, as well as a thank you.

The written down version of your interview will be analysed with the help of special linguistic computer software. This will help us find out if people experiencing low mood use different words or types of words when compared to people who are not feeling low.

What are the possible disadvantages and risks of taking part?

The Mood Feelings questionnaire will include questions about how you're feeling, which might feel upsetting. Information about emotional support will be available at the end of the questions. If the results of this questionnaire suggest that you are not feeling well, Kate will let you know confidentially and give you more information about people who could help.

What are the possible benefits of taking part?

The interview will be a chance to talk about what is important to you, with no interruption, judgement or expectation. You can help with a research project that could make a difference to people who are suffering with low mood.

Use, dissemination and storage of research data

The information from this project will only be used for the project itself, and only Kate will have access to anything that could show who you are.

What will happen to my personal information?

During the interview, your words will be recorded and alongside the wellbeing questionnaire you will be asked your age, ethnicity, gender and highest level of education. This information will be anonymised, meaning that your personal details will be removed and replaced with a code. Your interview will be recorded and typed up by Kate, and nobody else will be able to see the information that shows who you are. Whilst typing up your information, anything individual to you that could allow other people to work out which interview was yours will be changed.

As with any online related activity, the risk of a data breach is always possible. To the best of our ability your participation in this study will remain confidential, and only anonymised data will be published. We will minimise any risks by saving this questionnaire in a secure location under a randomly assigned number rather than your name. It will be archived once this PhD is complete and destroyed after 10 years.

Will I be recorded, and how will the recorded media be used?

The audio and video recordings of your interview will only be used as a guide for transcription. Nobody other than Kate will see or hear these recordings and they will be securely archived as soon as they have been typed up.

What will happen to the results of the research project?

The transcription of the interview and your questionnaire will be securely stored in two separate files using an anonymised code. The original recording of the meeting will be destroyed 6 years after the end of the project (i.e. September 2029), but the anonymised transcription and background information will continue to be stored securely beyond this time. We will keep a separate secure file with a 'key' that links your personal details to your anonymous number code, in case we need to contact you about anything important related to the safety of you or others. Only Kate will have access to this 'key' file, and it will be destroyed 12 months after the end of the project (i.e. September 2024). After that there will be no possible way to link the data you have given back to you.

The results will be used as part of a PhD thesis which will be completed around September 2023. All participants are welcome to request a copy of these publications from the researcher. It will not be possible to identify you from the data included in these research papers.

What type of information will be sought from me and why is the collection of this information relevant for achieving the research project's objectives?

After you have finished speaking, you'll be asked to fill in a questionnaire straight away. This will ask for your age, gender, race, sexuality, education and about your moods and feelings recently. This will show if you are currently feeling low as well as giving me more information about things that might affect how you speak.

Contact for further information: Kate Sayers Redacted

Finally …

This information sheet is all yours to keep, so you can remind yourself of what you could be agreeing to! The last thing to do is thank you for taking the time to read through this information.

Research participant privacy notice

Purpose of this Notice

This Notice explains how and why the University uses personal data for research; what individual rights are afforded under the Data Protection Act 2018 (DPA) and who to contact with any queries or concerns.

All research projects are different. This information is intended to supplement the specific information you will have been provided with when asked to participate in one of our research projects. The project specific information will provide details on how and why we will process your personal data, who will have access to it, any automated decision-making that affects you and for how long we will retain your personal data.

Why do we process personal data?

As a publicly funded organisation we undertake scientific research which is in the public interest. This includes processing personal data to develop new data analysis techniques and tools, and investigate factors that affect the treatment and trajectories of different diseases. The categories of personal data included are names, addresses, date of birth, mental and physical health data, ethnic origin and sex life, and this is usually analysed at the level of data pertaining to an individual. The types of organisations we receive data from include health related organisations such as the NHS, data held by government departments (with appropriate permission), private organisations and charitable organisations. The recipients of this personal data are researchers. The DPA requires us to have a legal basis for this processing; we rely upon “the performance of a task carried out in the public interest” as our lawful basis for processing personal data, and on “archiving in the public interest, scientific or historical research purposes, or statistical purposes” as our additional lawful basis for processing special category personal data (that which reveals racial or ethnic origin, political opinions, religious or philosophical beliefs, trade union membership, genetic or biometric data, and data concerning health, sex life or sexual orientation).

How do we follow data protection principles?

- We have lawful bases for processing personal and special category data.
- Data are used fairly and transparently; we will make it clear to individuals what their data will be used for, how it will be handled and what their rights are.
- We only collect and use personal data for our research, for research in the public interest, or to support the work of our organisation.
- We only collect the minimum amount of personal data which we need for our purposes.
- We take steps to ensure that the personal data we hold is accurate.
- We keep your personal data in an identifiable format for the minimum time required.
- We take steps to ensure that your data is held securely.

- We keep a record of our processing activities.

What do we do with personal data?

Research data can be a very valuable resource for improving public services and our understanding of the societies we live in. One way we can get the most benefit from this work is to make the data available, usually when the research has finished, to other researchers. In some cases we will share data with other parties during the research, but if this is the case your project specific information will detail why and how this happens. Sometimes these researchers will be based outside the European Union. We will only ever share research data with organisations that can guarantee to store it securely.

We will never sell your personal data, and any data shared cannot be used to contact individuals. We do not conduct automated decision making with the data.

The project specific information will include more detail about how your data will be used.

Your rights as a data subject

Because we use personal data to support scientific research on the public interest, individuals participating in research do not have the same rights regarding their personal data as they would in other situations. This means that the following rights are limited for individuals who participate, or have participated in, a research project:

- The right to access the data we hold about you.
- The right to rectify the data we hold about you.
- The right to have the data we hold about you erased.
- The right to restrict how we process your data.
- The right to data portability.
- The right to object to us processing the data we hold about you.

Data security

We have put in place security measures to prevent your personal data from being accidentally lost, used or accessed in an unauthorised way and will notify you and any applicable regulator of a suspected breach where we are legally required to do so.

Retention periods

We will only retain your identifiable personal information for as long as necessary to fulfil the purposes we collected it for; we may then retain your data in anonymised or pseudonymised format.

To determine the appropriate retention period for personal data we consider the amount, nature, and sensitivity of the personal data, the potential risk of harm from unauthorised use or disclosure, the purposes for which we process your personal data and whether we can achieve those purposes through other means, and the applicable legal requirements. Some of the organisations we get your data from specify when the data must be deleted in which case we will ensure we comply with their requirements.

Additional notices and guidance/policies

The University has also published separate policies and guidance which may be applicable to you in addition to other privacy notices:

Current staff privacy notice

Current students privacy notice

The Research and Innovation Service website has other relevant policies and guidance.

Communication

In the first instance please contact the researcher who your initial contact was with (Kate Sayers)

You may also contact the Data Protection Officer for further information (see contact details below). The University is the Data Controller for your data.

Please see the Information Commissioner's website for further information on the law.

You have a right to complain to the Information Commissioner's Office (ICO) about the way in which we process your personal data. Please see the ICO's website or contact 0303 123 1113.

Concerns and contact details

If you have any concerns with regard to the way your personal data is being processed or have a query with regard to this Notice, please contact our Data Protection Officer.

Our general postal address is University of Leeds, Leeds LS2 9JT, UK.

Our postal address for data protection issues is University of Leeds Secretariat, Room 11.72 EC Stoner Building, Leeds, LS2 9JT.

Our telephone number is +44 (0)113 2431751.

Our data controller registration number provided by the Information Commissioner's Office is Z553814X.

This notice was last updated on 28 September 2020.

Participant consent form

Consent to take part in A linguistic analysis of narratives from people with low mood	Add your initials next to the statement if you agree
I confirm that I have read and understand the information sheet dated 25.08.22 explaining the above research project and I have had the opportunity to ask questions about the project.	
I understand that my participation is voluntary and that I am free to withdraw at any time until 1st December 2022 without giving any reason and without there being any negative consequences. Please contact Kate Sayers enksa@leeds.ac.uk to do this. Any data already provided will be removed from the study and securely archived.	
I understand that my contribution will be anonymised, and I will not be identified or identifiable in the report or reports that result from the research.	
I understand that at all times this research project will comply with the General Data Protection Regulations (GDPR, 2018) approved by the EU parliament on 14 April 2016 and passing into UK law effective from 25 May 2018.	
I understand that the interview will be video and audio recorded and that these recordings will be securely archived once they have been transcribed	
I agree to take part in the above research project and will inform the lead researcher should my contact details change.	

Name of participant	
Participant’s signature	
Date	
Name of lead researcher	Kate Sayers
Signature	
Date*	

*To be signed and dated in the presence of the participant.

Once this has been signed by all parties the participant should receive a copy of the signed and dated participant consent form, the letter/ pre-written script/ information sheet and any other written information provided to the participants. A copy of the signed and dated consent form should be kept with the project’s main documents which must be kept in a secure location.

Appendix C

Pilot study follow up questionnaire

Age:

Gender:

Race:

CORE-OM

This form has 23 statements about how you have felt over the last week.

Please read each statement and think about how often you have felt that way last week

Then tick the box which is closest to this.

Over the last week…	**Not at all**	**Only occasionally**	**Sometimes**	**Often**	**Most or all of the time**
I have felt terribly alone and isolated	☐	☐	☐	☐	☐
I have felt tense, anxious or nervous	☐	☐	☐	☐	☐
I have felt I have someone to turn to for support when needed	☐	☐	☐	☐	☐
I have felt ok about myself	☐	☐	☐	☐	☐
I have felt totally lacking in energy and enthusiasm	☐	☐	☐	☐	☐
I have been physically violent to others	☐	☐	☐	☐	☐
I have felt able to cope when things go wrong	☐	☐	☐	☐	☐
I have been troubled by aches, pains or other physical problems	☐	☐	☐	☐	☐
I have thought of hurting myself	☐	☐	☐	☐	☐
Talking to people has felt too much for me	☐	☐	☐	☐	☐
Tension and anxiety have prevented me doing important things	☐	☐	☐	☐	☐
I have been happy with the things I have done	☐	☐	☐	☐	☐

I have been disturbed by unwanted thoughts and feelings	☐	☐	☐	☐	☐
I have felt like crying	☐	☐	☐	☐	☐
I have felt panic or terror	☐	☐	☐	☐	☐
I made plans to end my life	☐	☐	☐	☐	☐
I have felt overwhelmed by my problems	☐	☐	☐	☐	☐
I have had difficulty getting to sleep or staying asleep	☐	☐	☐	☐	☐
I have felt warmth or affection for someone	☐	☐	☐	☐	☐
My problems have been impossible to put to one side	☐	☐	☐	☐	☐
I have been able to do most things I needed to	☐	☐	☐	☐	☐
I have threatened or intimidated another person	☐	☐	☐	☐	☐
I have felt despairing or hopeless	☐	☐	☐	☐	☐
I have thought it would be better if I were dead	☐	☐	☐	☐	☐
I have felt criticized by other people	☐	☐	☐	☐	☐
I have thought I have no friends	☐	☐	☐	☐	☐
I have felt unhappy	☐	☐	☐	☐	☐
Unwanted images or memories have been distressing me	☐	☐	☐	☐	☐
I have been irritable when with other people	☐	☐	☐	☐	☐
I have thought I am to blame for my problems and difficulties	☐	☐	☐	☐	☐
I have felt optimistic about my future	☐	☐	☐	☐	☐
I have achieved the things I wanted to	☐	☐	☐	☐	☐

I have felt humiliated or shamed by other people	☐	☐	☐	☐	☐
I have hurt myself physically or taken dangerous risks with my health	☐	☐	☐	☐	☐

Scoring CORE-OM

The scoring for CORE-OM is numerical, as follows:

Not at all: 0

Only occasionally: 1

Sometimes: 2

Often: 3

Most or all of the time: 4

The questions were broken down into sections, as follows:

- Subjective well-being deficits (questions 4, 14, 17, 31)
- Problems/symptoms (questions 2, 5, 8, 11, 13, 15, 18, 20, 23, 27, 28, 30)
- Life functioning difficulties (1, 3, 7, 12, 10, 19, 21, 25, 26, 29, 32, 33)
- Risk/harm (9, 6, 16, 22, 24, 34)

Scores for each section are averaged out, and informants were classified as depressed if they had an average of over 2 for any section, and/or 4 as a response for any individual question

Standard debrief email:

That's everything, your participation today has been very appreciated.

If anything you have spoken about, read or considered as part of this research has caused you any concern or distress, it's really important to talk about this. You can call the Samaritans for free in the UK on 115 123, or alternatively LAMP is a Leicester based charity who are there to help with any mental health questions and support, they are available at https://www.lampadvocacy.co.uk/

Kate will be transcribing your response, then anonymising and securely storing. This data will be analysed for linguistic markers indicative of the symptoms associated with depression. If your questionnaire doesn't reflect depression symptoms, your data will be used as part of a control set.

If you have any questions, would like a copy of the final report or if you decide to withdraw your consent, please contact Kate Sayers on ks589@leicester.ac.uk

Thank you for helping!

Debrief email for informants who registered as depressed

Dear *name*

Thank you again for taking the time to help with the pilot study for my thesis. This email is to let you know that the results of your wellbeing categorisation suggest that you may be distressed. If you feel that you are struggling with your mental health, you are never alone.

The university of Leicester has a wellbeing service specifically for students such as yourself. This service includes counselling, self-help guides and any extra support that you may need. Their information centre can be found here https://www2.le.ac.uk/offices/mental-wellbeing/information-for-students/current-students/what-support-is-available or alternatively you can send an email to wellbeing@le.ac.uk

If you would prefer to talk to somebody outside of the university, the Samaritans are here for you any time you may want them. You can call them for free in the UK on 116 123 and there will always be a person there who is ready to listen. If you don’t want to talk but would like support from outside of the university, LAMP is a Leicester based charity. Their website is https://www.lampadvocacy.co.uk/ and they will be happy to help with questions and guidance.

I hope that you feel better soon, and that the rest of your time at university runs smoothly.

Best wishes,

Kate

Appendix D

Main study follow up questionnaire

Follow up questionnaire

Thank you for taking the time to speak with Kate about what is important to you. The purpose of the questionnaire part of this research study is to find more information about aspects of who you are that could affect how you talk, and it will take around 10 minutes to complete.

Age:

Race:

Gender:

Sexuality:

Highest level of education?:

Would you consider yourself to be depressed?:

Over the past few days, how have you felt?:

(Responses are not true, sometimes and true)

1. I felt miserable or unhappy.
2. I didn’t enjoy anything at all.
3. I was less hungry than usual.
4. I ate more than usual.
5. I felt so tired I just sat around and did nothing.
6. I was moving and walking more slowly than usual.
7. I was very restless.
8. I felt I was no good anymore.
9. I blamed myself for things that weren’t my fault.
10. It was hard for me to make up my mind.
11. I felt grumpy and cross with other people.
12. I felt like talking less than usual.
13. I was talking more slowly than usual.
14. I cried a lot.
15. I thought there was nothing good for me in the future.
16. I thought that life wasn’t worth living.

17. I thought about death or dying.
18. I thought my family would be better off without me.
19. I thought about killing myself.
20. I didn’t want to see my friends.
21. I found it hard to think properly or concentrate.
22. I thought bad things would happen to me.
23. I hated myself.
24. I felt I was a bad person.
25. I thought I looked ugly.
26. I worried about aches and pains.
27. I felt lonely.
28. I thought nobody really loved me.
29. I didn’t have any fun in any of my activities.
30. I thought I could never be as good as other people.
31. I did everything wrong.
32. I didn’t sleep as well as I usually sleep.
33. I slept a lot more than usual.

*

The above is the Mood and Feelings Questionnaire (MFQ), developed by Adrian Angold and Elizabeth J. Costello in 1987. Intended for children and young adults so is very simplistic, but this fits with the potential low education level of some of my participants.

"not true" = 0 points

"sometimes true" = 1 point

"true" = 2 points

27 or higher means low mood.

Debrief for a score of over 27 or answered positive to 19:

Thank you again for taking the time to help with my doctoral research. This is to let you know that the results of your questionnaire suggest that you may be distressed. If things are tough right now, please know that you are never alone.

The Samaritans are here for you any time you may want them. You can call them for free in the UK on 116 123 and there will always be a person there who is ready to listen.

If you would prefer to type rather than talk, you could join an online community of people who are going through a difficult time at https://sidebyside.mind.org.uk/ where a safe and supportive environment is waiting for you.

When everything gets too much, there are also practical tools to help straight away at https://www.mind.org.uk/need-urgent-help/

I hope that you feel better soon

Appendix E

Transcription protocol

Taken from Collins & Hardie 2022:

Punctuation use should be minimal. Question and exclamation marks may be left to the transcriber's intuition regarding whether intonation cues require them (rather than explicit interrogative/exclamative forms).

(word?) Round brackets/parentheses should indicate a transcriber's best guess at unclear words on the recording; an optional question mark can indicate especially uncertain guesses; and empty brackets ( ) a totally uninterpretable word(s).

[name] Square brackets with no flag should indicate content that has been redacted (with a label such as name, place, dateofbirth, as appropriate)

[#comment] Square brackets flagged with hash should indicate transcriber comments: any observations, interpretations or descriptions that are not actual recorded speech.

[@laughs] Square brackets flagged with @ should indicate vocalisations (with a label such as laughs, coughs, groans …) that occur within an utterance. KS note- duration of this is a factor, and overlapping with speech.

… Short pause: less than three seconds.

[pause] Medium pause: 3-10 seconds.

[silence] Long pause: anything more than 10 seconds.

Added by KS:

Aagh- scream sound

Bleugh- tongue twist sound