

Linguistic change in an online support group

Daniel J. McDonald
B.A. (Hons)

6th December 2016

School of Languages and Linguistics &
Melbourne Medical School
The University of Melbourne
Victoria, Australia

Linguistic change in an online support group

Daniel J. McDonald

Declaration

This thesis contains only original work by the writer, except for the references that have been appropriately acknowledged. Sections of this thesis contain work that has been presented in earlier versions in the following presentations, publications and repositories:

1. McDonald, D., & Woodward-Kron, R. (2016). Member roles and identities in online support groups: Perspectives from corpus and systemic functional linguistics. *Discourse & Communication*, 10(2), 157–175.
2. Zinn, J. O. & McDonald, D. (2015). Discourse-semantics of risk in *The New York Times*, 1963–2014: a corpus linguistic approach. Project report, University of Melbourne, Australia. Available at <http://git.io/vZ7yh>.
DOI: <http://doi.org/10.5281/zenodo.228363>.
3. Zinn, J. O. & McDonald, D. (2015). Changing Discourses of Risk and Health Risk: A Corpus Analysis of the Usage of Risk Language in *The New York Times*. In Chamberlain, M. (ed.), *Medicine, Discourse, Power and Risk*. Abingdon: Routledge: 1–30.
4. McDonald, D. (2014). Building web-corpora of patients' online mental health communication for mixed-methods investigation. Presented at the International Association of Applied Linguistics (AILA) World Congress, Brisbane, Australia.
5. McDonald, D. (2014). Contrasting mood and transitivity choices in new and established members' posts to an online support group. Presented at the Australian Systemic Functional Linguistics Association (ASFLA), Sydney, Australia.
6. McDonald, D. (2015). The Influence of medium and culture on language use in an online support group. Interaction Design Lab (IDL) seminar, University of Melbourne. Available at <https://youtu.be/baIelu6sERM>
7. Lam, M., McDonald, D., Raklao, T., Pun, J., Cheung, G. & Slade, D. (2015). Applying functional linguistics in exploring healthcare communication. Roundtable Communication, Medicine & Ethics (COMET), Hong Kong.

8. McDonald, D. (2015). Investigating online healthcare communication using systemic-functional and corpus linguistics. Poster. Communication, Medicine & Ethics (COMET), Hong Kong.
9. McDonald, D. (2015). Using an IPython Notebook to investigate language use in an online support group. Micro-analysis of online Data (MOOD-Z), Zürich, Switzerland.
10. McDonald, D., Woodward-Kron, R. (2015). Lexicogrammatical and discourse-semantic change over the course of membership in an online support group. International Systemic Functional Linguistic Congress, Aachen, Germany.
11. Zinn, J. O., McDonald, D. (2015). Discourse-semantics of RISK in *The New York Times*, 1963–2014: a corpus linguistic approach. International Systemic Functional Congress (ISFC), Aachen, Germany.
12. McDonald, D. (2015). `corpkit`: a new tool for functional linguistics. Presented at the International Research Centre for Communication in Healthcare (IRCCH) Seminar Series. The Hong Kong Polytechnic University, Hong Kong.
13. McDonald, D. (2015). Combining functional and computational linguistics to investigate discourse in an online support group for bipolar disorder. Presented at Mahidol University Linguistics Forum, Bangkok, Thailand.
14. McDonald, D. (2015). Discursive change over the course of membership in an online support group. Presented at BAAL Health and Science Communication SIG Event: Computer-mediated health communication: Perspectives from ethnography and discourse analysis. Queen Mary University, London, UK.
15. McDonald, D. (2015). Language change over the course of membership in an online support group: a corpus linguistic perspective. UCREL Corpus Research Seminar. Lancaster, UK.
16. McDonald, D. (2015). Using corpus and systemic functional linguistics to investigate language change in an online support group. Saarland University PhD Colloquium. Saarbrücken, Germany.
17. McDonald, D. & Zinn, J. O. (2015). Shifting discourse-semantics of risk in US newspapers, 1987–2014. Forum Entwicklung und Anwendung von Sprach-Technologien (FEAST) Seminar. Saarbrücken, Germany.

18. McDonald, D. & Rubino, R. (2016). Classifying Risk in News Articles using Grammatically Weighted Topic Models. Machine Translation Departmental Seminar, Saarland University. Saarbrücken, Germany.
19. McDonald, D. (2016). `corpkit`: a tool for functional analysis of parsed and structured corpora. SFB 833 Departmental Seminar. University of Tübingen, Germany.

The length of this thesis, exclusive of tables, references and appendices, is approximately 80,000 words.

Daniel J. McDonald

6th December 2016

Acknowledgements

Thanks to those who made this work possible, including supervisors, colleagues, family and friends. Funding for this research was provided by the Australian Commonwealth Government, the University of Melbourne, the German Research Foundation, and the Hong Kong Polytechnic University.

Abstract

Online support groups (OSGs) are popular sources of both health information and social support. Though early research into OSGs highlighted a concern that non-expert members may give harmful advice, more recent studies have typically shown that engagement with OSGs can increase consumer satisfaction with the treatment process, enhance wellbeing, and ultimately improve health outcomes. OSGs are well-researched within applied linguistics. Qualitative studies have focussed on member roles within OSGs, as well as the ways in which group members discursively construct their identities, especially with respect to their illness. These approaches have generated rich insights into consumer healthcare discourse that can inform strategies for fostering consumer-centred care. Qualitative approaches, however, are resource-intensive, difficult to reproduce, and limited in terms of generalisability and representativeness. Quantitative and computational approaches are able to overcome these shortcomings, adding transparency, reproducibility, scalability, and reducing the potential for researcher bias. Current computational approaches to consumer healthcare discourse, however, tend to rely on simplified conceptualisations of language, prioritising lexis over grammar, and thus ignoring the central role played by grammar in the meaning-making process.

To address current methodological shortcomings in OSG discourse research, this thesis presents an interdisciplinary, corpus-based investigation of lexicogrammatical and discourse-semantic choices made by members over the course of membership in an online bipolar disorder support community. 8.2 million words in over 66,000 posts from approximately 3,500 members were transformed into a metadata-rich, grammatically annotated corpus and investigated from a systemic-functional linguistic (SFL) perspective using purpose-built corpus/computational linguistic tools. An analysis of MOOD and MODALITY choices made over ten stages of membership highlights differences in the ways members negotiate role-relationships, with changes in Mood Type, Modality and Speech Function reflecting a longitudinal increase in the provision of advice and social support. An analysis of

the TRANSITIVITY system shows longitudinal changes in the kinds of participants and processes construed by Forum members, as well as changes in how these participants and processes behave lexicogrammatically. The *diagnosis* Event, for example, is represented by newcomers as a process and modified temporally; at later stages of membership, it is more often reconstrued as a participant in discourse, framed in terms of veracity. Longitudinal shifts were also observed in the preferred ways of ascribing/attributing bipolar disorder to Forum members: new members use *being* forms (*I'm bipolar*), while veteran members prefer *having* constructions (*I have bipolar*).

The thesis has implications for corpus linguistics, systemic-functional linguistic theory, and healthcare communication research. For corpus linguistics and corpus-assisted discourse studies, the main contribution is **corpkit**, an open-source software tool designed to build and analyse parsed and metadata-rich corpora. It is suggested that the developed tools and methods can circumvent theoretically problematic current practices, and increase the accuracy and automatability of the analytical process. For healthcare communication research, the case study demonstrates the importance of expanding the conceptualisation and analysis of the consumer healthcare journey to include intra-consumer communication that occurs outside of hospitals and clinics. The thesis also advances an argument that the emerging field of clinical natural language processing stands to benefit from increased engagement with functional linguistic theory and insights generated within the qualitative paradigm. I argue that combining the insights from functional linguistics and discourse analysis with automated computational workflows is a step toward an important future goal of improvement of consumer health outcomes through analysis of large, digital collections of spoken and written healthcare discourse.

Contents

1	Introduction	1
1.1	Context of the thesis	1
1.1.1	Language use in online support groups	3
1.1.2	Consumer-centred and computer-mediated healthcare	4
1.1.3	Exploiting computer-mediated discourse	4
1.1.4	Corpus linguistics and discourse	5
1.1.5	Functional linguistics and healthcare discourse	6
1.2	Statement of the problem	7
1.3	Aims of the thesis	9
1.4	Scope of the thesis	10
1.5	Research questions	11
1.6	Research site and approach	12
1.7	Contributions of the thesis	12
1.7.1	Implications for corpus linguistics	13
1.7.2	Implications for systemic functional linguistics	14
1.7.3	Implications for healthcare communication research	15
1.8	Overview of the thesis	16
2	Health discourse online	18
2.1	Computer mediated communication (CMC)	18
2.1.1	The changing face of CMC	21
Revisions of key claims in CMC research	21	
2.1.2	Contemporary CMC and its affordances for research	23
2.2	Healthcare and online communities	23
2.2.1	New members, first contributions and legitimacy	25
Legitimation strategies in newcomer talk	26	
Structure of first posts	28	
Limitations in current understanding of newcomer talk	29	
2.2.2	Veteran membership and legitimacy	30
Advice	33	
The lay-expert/proto-professional	39	
Operationalising ‘veteran membership’	41	
2.2.3	Pathways to sustained membership	42

Socialisation	42
2.2.4 Discourse socialisation in online communities	44
Four challenges to socialisation theory	47
2.2.5 Current limitations in health discourse research	50
2.3 Computational perspectives on online communities	52
2.3.1 Analysing health discourse quantitatively	53
Theoretical issues in computational health discourse research .	57
2.4 Chapter summary	59
3 Methods for investigating online health discourse	60
3.1 Preconditions for useful online healthcare discourse research	60
3.2 Corpus linguistics	61
3.2.1 Types of corpora and corpus research	62
3.2.2 Specialised corpus creation	63
3.2.3 Annotation of corpora	65
Parsing	67
3.2.4 Corpus interrogation practices	68
Keywording	68
Lexicogrammatical querying	68
Concordancing	69
3.2.5 CL and the World Wide Web	70
3.2.6 Corpora and discourse analysis	71
Emergence of the field	72
Contemporary practices	73
Common practices in CADS	74
CADS and computer-mediated communication	75
CADS, healthcare and the internet	77
3.2.7 Key debates in corpus-based approaches to discourse	78
Lack of available datasets	78
Under-utilised and unavailable digital tools and resources . . .	79
Simplified common practices	80
Addressing criticisms of the CL approach	81
3.2.8 Summary: corpus linguistics as an approach	85
3.3 Systemic functional linguistics	85
3.3.1 Cline of instantiation	86
3.3.2 Hierarchical structure of language and context	87
3.3.3 Metafunctions of language	87
3.3.4 Affordances of SFL for researching an OSG	88
Language as constitutive of context	89
Language as a meaning-making resource	90
Interpersonal and experiential functions of language	90
SFL and corpus linguistics	91

3.3.5	Overview of relevant elements of SFG	92
	Register	92
	Interpersonal meanings, MOOD and MODALITY	94
	Experiential meanings and the system of TRANSITIVITY	98
	Context of culture: genre and ideology	101
3.3.6	Criticism of SFL	105
3.4	Chapter summary	107
4	Case study design	109
4.1	Site description and selection	109
	Forum architecture	110
	Forum users	110
4.1.1	Justification for site selection	112
4.2	Ethics	113
4.2.1	Considering CMC/OSG data	113
	Contact with participants	115
	Anonymity and privacy	115
4.2.2	Interpreting the National Statement	116
	Participants with mental health issues	117
4.2.3	Minimising risk	117
4.3	Corpus building	117
4.3.1	Content retrieval	118
4.3.2	Corpus creation	118
4.4	<code>corpkit</code> : tools for corpus building and analysis	124
4.4.1	Rationale for tool development	125
4.4.2	Contents of the toolkit	126
	Key design parameters	128
4.4.3	Interfaces and functionality of the tool	128
	API	129
4.4.4	Alternative interfaces	138
	Graphical interface	139
	Interpreter	140
4.4.5	Open source development	140
4.4.6	Contributions of the toolkit	141
4.5	Approach to data analysis	142
4.5.1	<code>I</code> PYTHON & the <code>Jupyter Notebook</code>	144
4.5.2	Limitations	144
4.6	Situating the thesis methodology	145
4.7	Chapter summary	146
5	An introduction to the data: generic features of the Forum	147
5.1	Genre and first contributions	148

5.1.1	A user's first contribution to the Forum	149
	Generic stages in Jess first post	150
5.1.2	Replies to a first post	152
	Analysis of replies	153
5.2	Shallow lexicogrammatical features	157
5.3	Controlling for self-selection bias	159
5.4	Mapping membership stages to Forum history	160
5.5	Chapter summary	162
6	MOOD and MODALITY choices in the Forum	163
6.1	Mood and Indicative Type	163
6.2	Modalisation	166
6.3	Mood elements	170
6.3.1	Tense	173
6.3.2	Polarity	174
6.4	Summary	177
7	TRANSITIVITY choices in the Forum	178
7.1	Participants	179
7.1.1	Jargonisation	183
7.1.2	Metadiscourse	184
7.1.3	Vague language	186
7.1.4	Construing (in)stability	188
7.1.5	Construing human agency	188
7.2	Key processes	190
7.2.1	Construing diagnosis	193
	Diagnosis and grammatical metaphor	194
7.2.2	Construing the relationship between people and bipolar	197
7.2.3	Process-participant type configurations	201
7.3	Chapter summary	205
8	Discussion: meaning-making in the Bipolar Forum	206
8.1	Addressing Question 1: Lexicogrammatical features at risk	207
8.1.1	MOOD and MODALITY	207
8.1.2	TRANSITIVITY	207
8.2	Addressing Question 2: Discourse-semantics and register	208
8.2.1	Discourse-semantics in the Forum	209
	Social actors and interactants in the healthcare journey	209
	Discursive shifts	212
8.2.2	Register	215
	Tenor	216
	Field	217
	The healthcare journey	219

8.3	Critical reflection on the investigation	219
8.3.1	Unexplored linguistic phenomena	220
8.3.2	Data source issues	223
8.3.3	Theoretical and methodological challenges	225
Parsing	226	
Limitations in available systemic-functional resources	229	
The limits of lexicogrammatical querying	230	
Rank shift and grammatical metaphor	232	
The influence of genre	234	
Multimodality	235	
SFL, the individual, identity and the mind	236	
8.4	Summary	238
9	Implications of the thesis	239
9.1	Corpus linguistics	239
9.1.1	Corpus structure and metadata	240
9.1.2	Natural language processing tool use	241
Incorporating programming	243	
Indigenous reference corpora	245	
Improving normalised frequency calculation	247	
Collapsing the corpus/computational distinction	248	
9.1.3	Corpus assisted discourse studies	249
Exploiting parser output for discourse features	250	
Automating thematic analysis	251	
9.2	Systemic-functional linguistic theory	252
9.2.1	Theoretical contributions	252
Ergative transitivity and corpus methods	252	
Interaction between the metafunctions	253	
9.2.2	Practical contributions	255
Accounting for genre	255	
Ergative approaches	255	
Quantitative register modelling	256	
9.3	Health discourse	256
9.4	Addressing Question 3: Needed tools and methods	258
9.5	Implications: a summary	259
10	Future directions	260
10.1	A computational approach to healthcare discourse	261
10.1.1	Necessary developments	262
Automatic annotation of the semantic stratum	262	
10.2	Summary of the thesis	263
10.2.1	Lexicogrammar at risk over membership	263

10.2.2 Linking lexicogrammar to meaning	264
10.2.3 Linking back to research	264
10.2.4 Needed tools and methods	265
10.3 Conclusion	265
Notes	267
Bibliography	270
Glossary	302
List of abbreviations	306
Appendices	307
A Jess' first post	308
B corpkit documentation	313

List of Tables

3.1 Recent papers in CADS	74
3.2 Rank Scale	87
3.3 Mood Type and Speech Function	94
3.4 Grammatical metaphor as politeness strategy	97
3.5 Summary of Process Types	100
4.1 Membership Stage Structure: subcorpus attributes	121
4.2 Shallow features of the <i>Membership Stage Structure</i>	121
4.3 Shallow features of the <i>Future Veteran Structure</i>	122
4.4 Shallow features in the <i>Longitudinal Structure</i>	122
4.5 Shallow features in the <i>Comparative Structure</i>	124
4.6 Possible object–attribute combinations	131
4.7 Adjectives modifying doctor words	132
4.8 Concordancing adjectives modifying <i>doctor</i>	133
4.9 Example result	135
4.10 Example concordance	136
4.11 Tasks performed in <code>corpkit</code>	138
5.1 Genre stages in Jess's post	151
5.2 Genre stages in short first contributions	152
6.1 Relative frequency of Mood types	164
6.2 Thematic categories of <i>I would</i> + adjunct	168
6.3 Emphasising action in veterans' advice to newcomers	170
6.4 Concordancing <i>I</i> + <i>can</i> Mood Blocks	173
7.1 Relative frequencies of common participant heads	180
7.2 Key and unkey participants in three stages of membership	182
7.3 <i>Stability</i> in veteran posts	188
7.4 Human participants and lexical realisations	189
7.5 Most common modifiers of <i>diagnose</i> and <i>diagnosis</i>	196
7.6 Other Members as Senser in veteran posts	204
9.1 Methods in CADS and the cline of instantiation	251

List of Figures

2.1	Burnett's typology of online behaviour	20
3.1	Strata and metafunctions of language	88
4.1	Stated locations of members	111
4.2	Total members by postcount	111
4.3	A parsed sentence with metadata	119
4.4	Number of posts in the <i>Longitudinal Structure</i>	123
4.5	Relative number of posts in the <i>Longitudinal Structure</i>	123
4.6	Replies to new threads in the <i>Longitudinal Structure</i>	124
4.7	Using the <code>corpkit</code> API	132
4.8	Example investigation code	135
4.9	Example figure	135
4.10	Screenshots from <code>corpkit</code> 's graphical interface	139
5.1	Derived shallow features for the Membership Stage Structure	158
5.2	Derived shallow features for the Future Veteran Structure	160
5.3	Derived shallow features for the Comparative Structure	161
5.4	Lexical density in the Longitudinal Structure	161
6.1	Tregex queries for Major clause Mood Types	164
6.2	Mood features in the Membership Stage Structure	165
6.3	Advice provision via imperatives and modalised declaratives	167
6.4	Relative frequencies of modal lemmata	167
6.5	Functions of ' <i>I would</i> ' in new and veteran posts	169
6.6	New and veteran members' ' <i>I would</i> ' constructions	170
6.7	Pronoun-Subject + Modal-Finite blocks	171
6.8	Tense of tensed clauses over the course of membership	174
6.9	Pronoun-Subject + Past/Present tense blocks	175
6.10	Clause polarity over the course of membership	176
6.11	Constellations of Subject, Modal and Polarity	176
7.1	Trajectory of common participants undergoing change	180
7.2	Keywords on increasing and decreasing trajectories	183
7.3	Jargon term use by postcount	184

7.4	Key metadiscourse words	185
7.5	References to <i>board</i> in new and veteran talk	185
7.6	Two functions of <i>we</i> in veteran users' language	186
7.7	Increasing use of <i>things</i>	187
7.8	Frequencies for four participant types	189
7.9	Proportion of each participant type in Agent role	191
7.10	Key and unkey Events in each subcorpus	192
7.11	Goal of <i>diagnose</i> processes	193
7.12	Circumstances in <i>diagnose</i> processes	193
7.13	Grammatical metaphor in <i>diagnosis</i>	196
7.14	Processes with bipolar as Medium, combined	197
7.15	Processes with bipolar as Medium, separated	198
7.16	A veteran user discussing the being/having distinction	199
7.17	A veteran user employs <i>being</i> forms	200
7.18	A veteran user's characterisation of pdocs as incompetent	201
7.19	Four participant and Process Types	202
7.20	Keyness of processes involving four participants	203
8.1	Socio-semiotic processes	218
8.2	Universal dependency annotation	228
8.3	Dependency relationships	228
9.1	Switching between symbolic corpus structures	241
9.2	Complex query formulation	243
9.3	Recursive corpus investigation	244
9.4	Editing, sorting and visualising results	245

1. Introduction

In this chapter, I provide an outline of the investigation and contemporary literature from relevant research areas. This is followed by a statement of the problem and an explanation of the approach taken to address this problem. The main contributions of the thesis are summarised, and the structure of the thesis outlined.

1.1. Context of the thesis

It is well-established that the Internet is a popular source and vast repository of health information, the provision of which may take place via a diverse set of modes. These modes range from dedicated websites for particular health conditions or health organisations, in which healthcare professionals author content targeting healthcare consumers, to modes oriented toward intra-consumer interaction, such as social networking, web forums or wikis (Sillence, Hardy, & Briggs, 2013). Of particular interest to linguistic research in the past two decades has been interactional sites of health information exchange such as online support groups (OSGs): websites (or parts thereof) where users can post and reply to threads. OSGs, and online forums in general, are a well-known mode of computer mediated communication (CMC). In many cases, Online support groups can be viewed without logging in, and their threads are returned in search engine queries. Posting and replying to threads, however, usually involves creating an account. Forums vary widely in terms of their medium (i.e. technological) and situation (i.e. socially prescribed) affordances (Herring, 2007): some have strict moderation and can enforce bans; some allow users to embed images and videos into posts; some allow users to create profiles and send private messages (Morzy, 2012). Having remained in use since

their beginnings in the 1990s, these forums have a long research history (e.g. Sharf, 1997). That said, there is growing evidence that the popularity of online forums is in decline, with consumers instead receiving information and support from social networking sites (e.g. *Facebook*), link/content aggregation platforms (e.g. *Reddit*) or any of countless dedicated mobile apps (for diet and weight management, exercise, smoking cessation, and so on).

When compared with face-to-face support groups, OSGs may have unique benefits and consequences for users' health. In terms of benefits, forum users generally have round-the-clock access to a global community, facilitating larger groups and constant support (Stommel & Koole, 2010; Stommel & Meijman, 2011). This support has been linked to improved understanding of illness, the development of coping strategies, reduced anxiety and depression, and an increasing confidence in health professionals (Mulveen & Hepworth, 2006; Swan, Lau, & Bromberg, 2010; Manchaiah, Stephens, Andersson, Rönnberg, & Lunner, 2013, 25; Yao, Zheng, & Fan, 2015, 3). Furthermore, the degree of anonymity in such environments has sometimes been found to lead to a disinhibiting effect, encouraging honest discussion (Mo & Coulson, 2013) with qualitative differences from professional-consumer and/or face-to-face interactions (MacLean, Gupta, Lembke, Manning, & Heer, 2015). Of concern to some researchers, however, has been the potential spread of misinformation due to non-expert advice (Ziebland et al., 2004). Regular forum members may gain expert status within OSGs despite a lack of formal medical credentials (Hardey, 1999; Thompson, Bissell, Cooper, Armitage, & Barber, 2012). Such concerns are compounded in situations involving vulnerable participants who may not have the ability to make sound decisions regarding their own health. Finally, forum cultures may have the effect of normalising mental health issues such as suicidal ideation and eating disorders, and may encourage users to exhibit symptoms or obtain diagnoses for the purposes of legitimating themselves socially within the community (Horne & Wiggins, 2009; Vayreda & Antaki, 2009). These issues have been hypothesised to be the result of the dual function of OSGs as sites for both health information and social support exchange (Nambisan, 2011; Attard & Coulson, 2012).

1.1.1. Language use in online support groups

In OSGs, language is the dominant resource through which meanings are made. Accordingly, OSG research almost always involves some level of analysis of the language use of forum members, with or without explicit reference to linguistic theory. Within discourse-analytic OSG research, key areas of interest include (i) member roles, (ii) advice provision, (iii) legitimisation and (iv) socialisation. In terms of member roles and advice, though some early research in this area has been motivated by a concern that non-professional ‘experts’ may provide incorrect or harmful information, more contemporary findings generally suggest that the advice provided by veteran members to newcomers is often in line with mainstream biomedical norms (Vayreda & Antaki, 2009), and commonly mundane (i.e. *Go and consult with your doctor*) in nature (Smithson et al., 2011b). Legitimisation research has for the most part focussed on the ways in which newcomers construct an identity and message that warrants useful responses from others (Galegher, Sproull, & Kiesler, 1998; West, 2010). Depending on the community, legitimisation strategies have been found to vary widely: newcomers may emphasise the severity and uniqueness of their case (Varga & Paulus, 2014) or stress their inexperience and need for guidance. The third main focus within OSG research is socialisation—that is, how users learn through participation in meaningful social interaction with more experienced members of groups (Ochs, 1991). Lee, Park, and Han (2014), for example, approach socialisation from a member-life-cycle perspective, arguing that members transition through a number of roles during their time within the online community, and that each role has accompanying needs and responsibilities. Newcomers have strong needs for both information and social support, but may not contribute due to the potential for loss-of-face if information they provide is judged by experts to be incorrect (Füller, Jawecki, & Mühlbacher, 2007) or at odds with community-specific values (Weber, 2011). At later stages in the member life-cycle, users become less anxious about producing content, but lose the motivation to seek out information or support (Lee et al., 2014). Within this literature, lacking so far have been accounts of the longitudinal evolution of role-relationship negotiation strategies, comparisons of new and veteran users’ linguistic choices, quantitative approaches, and attempts to

map longitudinal change in discourse and semantics to shifts in lexicogrammatical features.

1.1.2. Consumer-centred and computer-mediated healthcare

Over the past few decades, there has been increasing recognition within mainstream healthcare institutions that consumer satisfaction and overall health outcomes can be improved through the practising of *consumer-centred medicine* (Stewart, 1995). Under this paradigm, clinicians foster collaborative exchanges with those they treat: greater attention is paid to consumers' feelings and beliefs; consumers are actively involved in decision-making processes; clinicians establish long-term relationships that can be responsive to consumers' prior journeys through the healthcare system (Woodward-Kron, 2016). The consumer-centred model thus recognises the centrality of communication and interaction to the practice of medicine, necessitating functional linguistic analysis of healthcare communication (HC)—that is, research into the relationship between language use in healthcare and more effective healthcare practice. So far, however, healthcare communication research has typically focussed on clinician-consumer interactions within formal settings, such as hospitals and clinics (Slade et al., 2015a). Despite acknowledgement that the consumer journey extends far beyond his/her interactions with healthcare professionals (Balka, Krueger, Holmes, & Stephen, 2010; Dickerson, Reinhart, Boemhke, & Akhu-Zaheya, 2011), and despite the increasing prominence of the interactional Web in daily life (Hadlington, 2015), little has been done to connect consumers' interactions with clinicians to their use of computer mediated communication (CMC) modes such as OSGs.

1.1.3. Exploiting computer-mediated discourse

Just as online healthcare spaces may have unique effects on health, so too do they provide researchers with a unique window into the consumer healthcare journey (Harvey, 2012). First, as noted above, situation factors present in OSGs (i.e. the informal, intra-consumer Tenor of interactions) lead to texts that construe the consumer journey candidly, in detail, and unadulterated by potential influence of health-

care professionals. At the same time, the medium factors of OSGs (i.e. the way the interactions are produced and archived) make viable the use of state-of-the-art computational linguistic methods: because OSGs are generally anonymous, large, well-structured and metadata-rich, they can be automatically transformed into high-quality corpus linguistic resources, grammatically annotated, and searched using tools and methods from corpus linguistics (CL).

Using these emerging computational methods, it is possible to build reliable, quantitative accounts of healthcare consumers' language choices (see MacLean et al., 2015, for a recent example). In contrast to the collection and manual analysis of face-to-face, clinician-consumer interactions in formal healthcare settings, computational methods are dramatically more scalable, reproducible and time/cost effective (Yao et al., 2015, 3). Moreover, as computational methods improve in speed and accuracy, a number of new applications of computational analysis of consumer language use are beginning to emerge. Large, digitised datasets are being used to predict health outcomes, to identify health risks (Kim, Seok, Oh, Lee, & Kim, 2013; St Louis & Zorlu, 2012; O'Leary, 2015), and to build intervention programs (Chen et al., 2015). Currently, however, computational models of healthcare discourse are often more heavily based on metadata features of CMC texts, such as timestamps and geotags, with the authors' actual linguistic content remaining under-utilised (Yesha & Gangopadhyay, 2015). The main reason for this is that natural language processing (NLP) methods for accurately processing healthcare texts are, in many respects, still in their infancy: despite recent advances in parsing and information extraction, automatic extraction of useful information from large quantities of consumers' health talk is far from a solved task. As such, outside of research environments, clinical applications of NLP to date have generally been limited to non-critical/low-stakes tasks such as supplementary data analysis, speech-to-text systems, and the sorting and filtering of texts (Maddox & Matheny, 2015; Wasfy et al., 2015).

1.1.4. Corpus linguistics and discourse

Corpus linguistics (CL), and more recently, corpus assisted discourse studies (CADS), provide both a potential methodological orientation and a set of practices that may

assist in quantitative investigation of a corpus of contributions to an OSG. As a branch of discourse analysis, corpus assisted discourse studies (CADS) is well-suited to use within consumer-centred healthcare research: both prioritise meaning-making and lived experience; both are sensitive to grammar, narrative and context (Crawford, Gilbert, Gilbert, Gale, & Harvey, 2013; Partington, 2004). A key strength of CADS is its ability to locate what Gee (2007) has called *Big D Discourses* (those concerning social status, power, etc.) within sets of related texts whose topics may be *small d discourses* (about particular things and events in the world; see Baker, 2013). At the same time, the use of CL methods in discourse analysis can be used to limit researcher bias and enhance reproducibility: frequency information can demonstrate that the linguistic patterns being discussed are typical of, rather than ‘cherry picked’ from, the very large samples of language from which corpora are constructed (Baker, 2012). To date, however, CADS has focussed more on corpora of news articles, policy documents and government communications than CMC. Of the few studies of CMC corpora (e.g. Harvey, Brown, Crawford, Macfarlane, & McPherson, 2007; Harvey, 2012; Prentice, 2010), none has dealt with a single, self-contained online community or OSG, despite the relative ease of building large, well-structured corpora from posts to a publicly accessible forum. Finally, CL and CADS practitioners generally do not take advantage of state-of-the-art NLP tools for annotating and extracting features from natural language (Groom, Charles, & John, 2015), limiting the extent to which grammatical patterns can be analysed alongside lexis. This leaves a gap between what is automatically quantifiable (i.e. lexis, and the adjacency of lexical items) and the linguistic strata of interest (meaning, discourse and semantics), limiting the extent to which CADS can be automated, and ultimately, the extent to which CADS research can inform clinical practice or other non-research settings.

1.1.5. Functional linguistics and healthcare discourse

Central to useful analysis of discursive patterns *en masse* are a conceptualisation of language as a functional resource, an understanding of how words and wordings relate to meaning and context, and an awareness of how structure unfolds through text. The majority of quantitatively oriented studies of online health discourse,

however, simplify what language is and how it works: language is often reduced to lexis and collocation of lexis; texts become bags-of-words (e.g. MacLean et al., 2015; Yesha & Gangopadhyay, 2015). Few have explicitly drawn upon functional grammars or theories of language designed to connect linguistic systems and strata in a reliable way.

Perhaps the most comprehensively articulated of contemporary functional grammars (Eggins & Slade, 2004) is systemic functional linguistics (SFL), a socio-semantic theory of text and context (Halliday, 2004). SFL theorists argue that language is structured in order to achieve three kinds of meaning: **interpersonal meanings**, which construct and negotiate role-relationships between speakers; **experiential meanings**, which communicate doings and happenings in the world; and **textual meanings**, which reflexively organise language into coherent, meaningful sequences. These discourse-semantic functions are realised by different parts of a language's lexis and grammar. In English, interpersonal meanings are made via the MOOD and MODALITY systems. Experiential meanings are made through the TRANSITIVITY system. Textual meanings are made through referential and conjunctive functions, as well as the system of THEME AND RHÈME. In the context of OSGs, the distinction between interpersonal, experiential and textual meanings potentially allows the researcher to separate analyses of social support (via MOOD choices) and health information provision (via TRANSITIVITY choices), while remaining sensitive to how the architecture of the forum itself may affect choices of THEME. Despite this potential, and despite recent use of SFL in healthcare communication (Matthiessen, 2013; Slade et al., 2015a; Woodward-Kron, 2016), CMC (Lander, 2014; Zappavigna, 2013) and CL (Hunston, 2013, 4–5; Thompson & Hunston, 2014) contexts, SFL has yet to be operationalised within a linguistic exploration of an OSG.

1.2. Statement of the problem

Recent developments across the intersecting fields outlined in the sections above have made it possible to use online health discourse, combined with computational and/or discourse-analytic methods, to gain new insights into consumers' healthcare

journeys and experiences. Discourse-analytic research has already demonstrated how OSG members negotiate role-relationships, reinforce community values, and give and receive health information and social support. At the same time, quantitative and computational approaches are beginning to show that largely automated methods can also yield insights into the same kinds of data, many of which may be useful in clinical settings. So far, however, despite their overlapping goals, the qualitative/discourse-analytic and quantitative/computational streams of research have yet to be brought together, and as such, each has been unable to profit from theoretical and methodological advances made within the other.

Because of a lack of dialogue between the two main approaches to OSG research, both have identifiable current limitations. The discourse-analytic stream, while providing rich insights into social support and information provision online, has yet to be able to describe the interaction between these semiotic commodities, or the ways in which their lexicogrammatical realisations may change throughout membership. Due to this stream's reliance on researcher interpretation of data, results are more expensive to generate, and more difficult to reproduce. At the same time, potential approaches for computationally analysing OSGs require further development: computational linguists interested in classification of and prediction from health-oriented talk have not taken advantage of existing accounts of how users interact and construe the world online, nor of functional frameworks for connecting instantiated words and wordings to the more abstract plane of meaning more generally. As a result, computational analyses of OSGs have prioritised lexis at the expense of grammar, and thus failed to account for the central role of grammar in the meaning-making process.

The final key problem addressed by the thesis is the narrow focus of existing healthcare communication (HC) literature. Within this research area, despite widespread adoption of the consumer-centred paradigm, the consumer journey typically only becomes an object of study when the consumer first engages with a health professional, and often ceases to be studied when the consumer leaves the hospital or clinic. This leaves the vast quantities of readily accessible health-oriented CMC uncharted, and thus, unable to be exploited for the purposes of informing practice

or improving health outcomes. Potential tools and methods for extracting insights from intra-consumer health discourse, as well as the viability of the endeavour more generally, remain under-explored. Moreover, focussing on communication in formal healthcare institutions means that the journeys of those who suffer from health problems, but do not seek treatment through formal channels, are undocumented.

1.3. Aims of the thesis

The primary aim of this thesis is to investigate linguistic change over the course of membership in an OSG. To achieve this aim, I conducted a case study of a large online Bipolar Disorder support group (henceforth, the *Forum/Bipolar Forum*). Posts to the Forum were transformed into an annotated, metadata-rich corpus, structured to contain ten subcorpora of members' posts at different stages of membership. These subcorpora were then interrogated for a combination of lexical and grammatical features, in order to locate components of lexicogrammar that are *at risk*—that is, likely to vary in frequency—over the course of membership. MOOD and TRANSITIVITY systems were queried separately, with concordancing used where necessary to build an account of how the located lexicogrammatical phenomena work to make interpersonal and experiential meanings. Because current corpus analysis tools are insufficient for these aims (see Chapter 4), a secondary aim of the research project was the development and application of reusable corpus linguistic tools and methods for extracting functional linguistic information from parsed, structured corpora. Accordingly, Chapter 4 and Appendix B describe an open-source Python module, `corpkit`, which facilitates the construction and functional linguistic analysis of corpora (<https://github.com/interrogator/corpkit>). The code used to generate all findings is also available within interactive *Jupyter Notebooks* (via <https://github.com/interrogator/thesis>), allowing other researchers to manipulate the dataset and reproduce results.

1.4. Scope of the thesis

Necessarily, the scope of the investigation must be constrained in a number of ways. First is in the selection of a forum-based OSG as the focus of the case study, rather than any of a diverse range of other online modes for health information/social support transmission (blogs, chat sites, YouTube videos, etc.). Forums were chosen due to their ubiquity, their sustained history of academic inquiry (Kim et al., 2012; Sillence et al., 2013), and the publicly accessible nature and the comparative ease of harvesting posts and their metadata. As such, given that the case study concerns only one community, I can make no definitive claim that the findings are generalisable to OSGs for other health conditions, OSGs hosted on other domains, or OSGs or online communities more generally. Second, in terms of linguistic areas of interest, the case study analysis centres on MOOD and TRANSITIVITY choices, as defined by the systemic-functional grammar (as in Halliday & Matthiessen, 2004). A notable constraint is that the dimension of Mode—of THEME AND RHÈME, as well as reference and conjunction—is largely ignored due to spatial considerations. It must also be noted that although a number of linguistic theories could be useful in an investigation of an OSG (e.g. Conversation Analysis, Interactional Sociolinguistics, Frame Semantics), and though multiple theories can be usefully applied simultaneously (Eggins & Slade, 2004), SFL is the dominant theory used in the thesis, mainly because of its comprehensive treatment of lexicogrammar, its usefulness for discourse analysis (Widdowson, 2000), and, to a lesser extent, its history of use in computational contexts (see O'Donnell & Bateman, 2005). Computationally, though tools were developed for extracting features from parsed and structured corpora, and for mapping constituency and dependency annotations to systemic-functional constructs, scope did not permit development of tools (or improvement of existing tools) for systemic-functional parsing. As a result of these limitations, systemic notions are simplified at times in order to be operationalisable within a predominantly computational workflow.¹

A further limitation in scope is the ability to perform sustained analyses of individual posts and threads. Forum threads are dialogic, with meanings made cooperatively between speakers. The main unit of analysis, however, must be the

post, in order to track how language use in members' contributions changes over time. The genre analysis presented in Chapter 5, while highlighting generic features of language use in the community, is limited to a very small selection of texts, and cannot be taken as representative of the community as a whole. Rather, it is intended to be illustrative of the kinds of texts in the corpus, and of the difficulty of accounting for genre and context in automated workflows.

The final major limitation concerns the kinds of computational tools and techniques applied to the data. While the methods used for the case study analysis often augment and improve currently dominant Corpus Linguistics methods, the research design does not include any of a number of computational linguistic developments that have demonstrated success in classifying texts and linguistic features therein. Techniques such as language modelling and vector space representations, topic modelling, or machine learning approaches to text classification have been shown to outperform rule-based approaches. For the purposes of this case study, however, methods were drawn from those typical of corpus, rather than computational linguistics. In later chapters, I advance an argument that a rigid distinction between the disciplines is unhelpful, and that computational developments could play an important role in expanding the explanatory power of corpus approaches to healthcare communication.

1.5. Research questions

1. Which components of lexicogrammar are *at risk*/subject to change over the course of membership within an Online support group?
2. How do these changes relate to discourse–semantics and register in the OSG?
3. What implications do the findings generated by this case study have for corpus linguistics, corpus–assisted discourse studies, systemic–functional linguistics, and healthcare communication research?
4. What kinds of tools and methods are needed in order to effectively analyse the data, and, more generally, to perform functionally driven analysis of natural language corpora?

1.6. Research site and approach

The site of the investigation is the Bipolar Disorder discussion board of a popular health-centred website. The board is one of the more popular of hundreds of sub-communities dedicated to individual physical and mental health issues. At the time of data collection in February, 2014, the board had received over 66,000 posts and contained almost 9000 threads since its creation in February, 2001. Posts have been made using 3588 unique usernames. Accounting for the possibility of users creating multiple accounts, well over 3000 unique participants can be safely assumed. *Lurkers*—those who may read posts, but not actively contribute, may comprise as much as 90 per cent of the readership (Preece, Nonnecke, & Andrews, 2004).

Corpus/computational linguistic tools were used to build a grammatically annotated, 8.2 million word corpus of every post to the Bipolar Forum. This corpus contains ten subcorpora, reflecting posts made at different stages of membership. Using the purpose-built tools, the lexicogrammar of the corpus was then interrogated for shallow features (e.g. lexical density, part-of-speech distributions), MOOD and MODALITY features (e.g. Mood and Indicative Type) and TRANSITIVITY features (e.g. key Participants and Processes, and the way these typically behave). Using SFL as a theoretical framework, lexicogrammatical findings are linked to discourse-semantic functions. A brief analysis of generic features of members' first contributions to the Forum is also performed, in order to gain a contextually grounded understanding of Forum texts, and to illustrate issues in automated CL methods. Finally, to ensure that the approach is modelling longitudinal linguistic change, rather than modelling inherent differences between the language use of those who drop out early and those who do not, investigations of alternative corpus structures (where posts from early dropout members are discarded, and where these dropouts' choices are contrasted with those of future veterans) are briefly presented and discussed.

1.7. Contributions of the thesis

The investigation uncovered a number of lexicogrammatical sites of change within the Bipolar Forum. Within the MOOD system, though declaratives and interrog-

atives stayed largely stable in frequency, imperatives and modalised declaratives became increasingly common over the duration of membership, with veteran members explaining to newcomers what course of action they should take. In terms of interpersonal meanings made by these MOOD choices, the thesis demonstrates the ways in which member roles and responsibilities within the community shift over time. Veteran members provide social support and health information to newcomers through advice, blending the speech functions of giving information and demanding goods and services (in this case, actions or general changes in behaviour). This is typified by strategies for advice provision: hypothetical, modalised declaratives (*I would seriously consider changing docs*) become a dominant pattern in veterans' talk with newcomers. These forms foreground lay experience as the source of knowledge underlying claims about what new members should do. In terms of experiential meanings, *metadiscourse*, *vague language* and *jargonisation* emerge as key sites of change over the membership course. Users of the Forum are increasingly construed as Agents. Participants and processes associated with instability and negative emotions are displaced by lexis that emphasises stability and control. Shifts can be observed in the kinds of participants and circumstances that attach to key processes, such as the process of diagnosis. Further, the ways in which forum members represent the relationship between people and bipolar disorder itself change over time: veteran members experientially position the disorder as a possession, rather than an identity—in systemic-functional terms, as an Attribute, rather than a Value. This distinction foregrounds members' agency over their condition and thus their ability to manage its symptoms and their effect on daily life.

1.7.1. Implications for corpus linguistics

The methodology developed for the investigation presents a number of new possibilities for CL/CADS research. First, I demonstrate the viability of using novel kinds of data and data structures as corpora, showing how corpus-based approaches can be used to investigate phenomena such as online communities, role-relationship negotiation and socialisation. At the same time, I demonstrate the utility of automatic parsing for discourse-analytic work, and provide new tools and methods for

traversing these annotations and extracting functionally useful information. Second, I demonstrate the utility of SFL as a means of mapping lexicogrammatical change to discourse-semantic change, and of delineating interpersonal and experiential metafunctions of language in corpora. Also presented are novel approaches to core CL tasks such as relative frequency and keyword calculation, linear-regression based sorting, and corpus visualisation. Together, these contributions significantly extend the ability of CL/CADS research to both describe and explain novel kinds of texts and to make claims about recurrent discourses through the use of quantitative, corpus-based approaches.

1.7.2. Implications for systemic functional linguistics

The case study also has implications for SFL, contributing to both theory and the range of registers that have received sustained treatment within the tradition. At the level of theory, the interrelatedness of metafunctions within the corpus data is described. In SFL, experiential meanings are not considered to play important roles in role-relationship negotiation. In the community, however, experiential choices—such as the distinction between *being* and *having* bipolar disorder—appear to not only communicate propositional information, but also construct and negotiate the newcomer/veteran identities within the board. This tension is also apparent in the way jargon terms are instantiated by veteran members, with jargon appearing to play important roles within both interpersonal and experiential meaning. Interpersonally, jargon demonstrates familiarity with community norms and expectation—solidarity and contact, within the Appraisal system (Martin & White, 2005). Experientially, jargon in this context may also be a useful marker of knowledge about health, as jargon typically achieves more delicate distinctions between important participants in a given Field of discourse: community users distinguish between kinds of bipolar disorder and kinds of anti-depressants, for example, with a great deal more specificity than would be expected outside of mental health oriented communities.

An added contribution of the thesis for SFL is its answering of calls for further research integrating SFL and corpus linguistics (Hunston, 2013, 4–5; Thompson &

Hunston, 2014), and in using SFL for consumer-centred healthcare research (Matthiessen, 2013; Thompson & Collins, 2001). The case study also contributes to Matthiessen's call for *registerial cartography*, which aims to situate described registers within maps drawn along either the axis of instantiation or the axis of stratification (Matthiessen, 2015a). The articulation of methods and findings presents new ways of doing SFL for potential take-up by other researchers, and furthers claims that SFL can be usefully applied in diverse contexts as a means of describing and explaining the role of language in interaction. The freely available software developed for the case study also integrates a number of systemic functional concepts (insofar as is possible with current parsers), including Process Type matching and translation of Universal Dependencies (see Nivre, 2015) to systemic labels.

1.7.3. Implications for healthcare communication research

The final implications of the thesis are for HC research. The shift toward consumer-centred healthcare has resulted in a need to account for language use in the diverse kinds of situations encountered by consumers on their progression through formal healthcare systems (Matthiessen, 2013; Slade et al., 2008). As Jones (2013) reminds us, however, consumer healthcare journeys exist both inside and outside of hospitals and clinics: consumers' knowledge of health problems, their feelings, and their decision making practices are influenced not only by health professionals, but by friends and family, online and offline. Importantly, unlike HC research within clinics and hospitals, interactions within a publicly accessible OSG include the voices and lived experiences of those who may be marginalised, and/or those who have never sought professional treatment (Harvey, 2012; Mautner, 2005). The thesis therefore provides an empirical perspective on health discourse of consumers of formal institutions, highlighting their journeys through an emphasis on longitudinal change over the membership course. At the same time, with recognition within medical institutions that consumer-centred healthcare can lead to better health outcomes (Woodward-Kron, 2016), and that corpus/computational methods can be used to extract meaningful information from health discourse (Mayfield, Laws, Wilson, & Rosé, 2014, e1), the creation of a reproducible, resource-effective methodology for

analysing online healthcare discourse facilitates future research into other online communities, health conditions or consumer demographics. As mentioned above, future work based on the methods presented here could also be used to better link HC research to clinical outcomes.

1.8. Overview of the thesis

Chapter 1 introduces the areas of research, the research questions, and the site of the investigation. It also summarises the contributions of the thesis for CADS, SFL and HC.

Chapter 2 reviews existing linguistic research into online health communities, in order to synthesise and critically reflect on what is currently known about language use in OSGs.

Chapter 3 proposes CL as a set of methods for investigating OSGs, and SFL as a theory of language amenable to corpus linguistic analysis of CMC. Shortcomings in corpus linguistic tools and methods for doing discourse analytic work are highlighted, as are difficulties in implementing systemic-functional concepts into corpus/computational linguistic workflows.

Chapter 4 describes the case study of the thesis. A general description of the Bipolar Forum is provided, as well as ethical considerations, and processes for transforming the Forum into a linguistic corpus are explained. Tools developed for interrogating the data are described and justified. The approach to analysis is also outlined.

Chapter 5 presents an introduction to the investigation. This involves a bottom-up, qualitative, genre-based analysis of a small selection of contributions to the Forum. The genre analysis is followed by an account of shallow linguistic features in the Forum, such as word length, clause and sentence length, and lexical density.

Chapter 6 presents findings from the investigation of MOOD and MODALITY features of the corpus highlighting their relationship to Forum users' interpersonal meaning-making practices.

Chapter 7 presents an analysis of TRANSITIVITY choices, mapping longitudinal change to shifting experiential semantics.

Chapter 8 addresses Research Questions 1 and 2 by providing a semantically organised discussion of lexicogrammatical and discourse-semantic features at risk over the course of membership within the Forum. Findings are related to those of earlier OSG literature reviewed in Chapter 2. A critical reflection on theoretical and methodological issues of the case study design is also presented.

Chapter 9 outlines the implications of the study for corpus linguistics/corpus-assisted discourse studies, SFL and healthcare communication research. It then addresses Research Question 4, concerning the development of new tools and methods for CL.

Chapter 10 provides a brief discussion of possible future research, followed by a summary of the thesis and a conclusion.

2. Health discourse online

In this chapter, I synthesise current knowledge about language use in OSGs. This begins with an overview of computer mediated communication (CMC) and CMC theory. Within this context, I then review contemporary linguistic accounts of OSGs, focussing on member roles, advice provision, legitimisation and socialisation. I highlight limitations in the existing literature, including issues of reproducibility and generalisability. A review of computational linguistic approaches to health discourse is also presented. I argue that such approaches may be usefully complemented by the kinds of insights generated by discourse analysis informed by functional linguistic theory.

2.1. Computer mediated communication (CMC)

From the 1970s to the early 1990s, CMC was largely limited to specialised communities, with technologies such as ARPANET designed with military and research purposes foremost in mind (Thorne, 2008). As the Internet user-base gradually diversified to include university staff and students, as well as select members of the private sector, CMC became progressively more social in nature. By the early 1980s, linguists were researching both synchronous (i.e. chat) and asynchronous (e.g. bulletin boards, email) digital environments (e.g. Carey, 1980; Myers, 1987; Pullinger, 1986).

This early body of research posited enduring representations of computer-mediated and online environments (Postmes, Spears, & Lea, 2000). Often articulated was the *reduced-cues perspective* (Thorne, 2008), which argued that CMC was a ‘lean’ or ‘low-bandwidth mode’, stripped of the kinds of context available in face-to-

face settings. Due to the observation that ‘social cues are filtered out in on-line settings’ (Parks & Floyd, 1996, p. 81), for instance, it was argued that complex corporate information was better suited to face-to-face transmission, due to the comparative ‘richness’ of the face-to-face mode. The main utility of CMC was as a way of transmitting large quantities of quantitative data (Daft & Lengel, 1983). Similarly, in social scenarios, Parks and Floyd contended that the unavailability of ‘information regarding physical appearance’ (1996, p. 84) stymied the potential for users to build meaningful social relationships online.

Given the novel mixture of intimacy and detachment occurring during text-based CMC (King, 1996), research into the affordance of anonymity and its ramifications also proved major themes of early literature (Tanis & Postmes, 2007). It was commonly argued until as late as the mid 1990s that anonymous CMC was inherently genderless and egalitarian, due to the difficulty of verifying the identity and credentials of participants (Herring, 2001). Related was the idea that people online were able to locate communicative partners based on shared interests and values, unbound by geographical location, with the potential consequence of more meaningful exchanges. On the other hand, many theorists argued that participants’ anonymity allowed for reduced inhibitions, causing an increase in hostile discourses, or *flaming* (Collins, 1992). Kiesler, Siegel, and McGuire (1984) contended that flaming arose due to a lack of audience feedback, an inability to control or sanction those who transgress community rules, and ‘depersonalization’ stemming from an absence of non-verbal communication. Kim and Raja (1991, p. 7) concurred, stating that in CMC, participants were more likely ‘to abuse, make offensive comments, or criticize sharply’.

During this period, some scholars aimed to classify different types of CMC modes, or the kinds of behaviours exhibited by interactants therein. Crowston and Williams (2000) argued that modes of CMC can be understood from the perspective of genre theory. They found that for the most part, online genres of interaction may be *reproduced* from pre-existing online antecedents (for example, online books and academic articles) or *adapted* from offline antecedents into new genres (where, for example, hyperlinking creates new ways of browsing collections of texts). *Novel* genres with

no identifiable antecedent, such as website homepages and search engine listings, were also described. Interactive websites such as online discussion forums were conceptualised as *novel*, bearing little resemblance to pre-existing offline genres. Burnett (2000) provided an early typology of types of behaviour within novel interactive genres, noting the potential for non-interactive lurking, and dividing active participation into hostile (flaming, trolling, etc.) and collaborative (humour, announcements, gossip, etc.) types (see Figure 2.1). Though perspectives on early CMC were variously optimistic or pessimistic, common to most was a treatment of CMC under a *deficit model*, where online interactions were considered inherently impoverished when compared to face-to-face modes. Burnett's typology is a good example of the implicitness of the deficit model in early research—indeed, it is difficult to imagine a researcher proposing such a reductive classificatory scheme to account for all offline human behaviour.

Non-interactive behavior		Lurking	
Interactive behavior	Hostile	Flaming	
		Trolling	
		Spamming	
		Cyber-rape	
	Collaborative or positive	Behaviors not specifically oriented toward information	Neutral behaviors: pleasantries and gossip
			Humorous behaviors: language games and other types of play
			Empathic behaviors: emotional support
	Announcements	Behaviors directly related to either inf. seeking or providing inf. to other community members	Announcements
	Queries or specific requests for information		Queries or specific requests for information
	Directed group projects		Directed group projects

Figure 2.1: Burnett's typology of online behaviour (2000)

2.1.1. The changing face of CMC

Between the mid 1990s and the mid 2000s, the nature of CMC—and consequently, CMC research—underwent a period of rapid change. As the result of a number of interwoven factors (the growing affordability of home computers; increasingly digital literacy; the development of early social network sites; etc.) the landscape of the Internet shifted from being primarily static and text-based to dynamic, multimodal and participatory in nature (Herring, 2011; Lindholm, 2012). The three currently most popular websites according to *Alexa* in 2016 (*Google*, *YouTube* and *Facebook*) exemplify this shift, in that each allows a great deal of user-input and provides content in textual, audio and graphic modes. Herring (2011) reimagines the typology of familiar, adapted and new genres proposed by Crowston and Williams (2000) to relate the dominant text types of ‘Web 2.0’ to those that came before. Importantly, she notes that web genres can shift toward the *new/emergent* categories as they mature, gaining features and responding to users’ needs.

Revisions of key claims in CMC research

The profound nature of the shift to ‘Web 2.0’ has meant that much early CMC literature has now lost some of its relevance or explanatory power. Research into email messages is problematised by the fact that social media has overtaken email as the CMC mode of choice for most digitally literate citizens (Thorne, 2008), and by increasingly blurred boundaries between synchronous and asynchronous modes, such as email and instant messenger services. Similarly, the anonymity posited as critical to the discourse of early CMC, while still possible, may no longer be the norm: on social networking sites such as Facebook, users communicate through personal accounts, disclosing their identities as they communicate and often making communication viewable by family and friends (boyd & Ellison, 2007). Furthermore, given that social networking sites are multimodal, and often contain a mixture of synchronous and asynchronous interactions, research informed by a characterisation of the Internet as fundamentally text-based, or dealing exclusively with synchronous or asynchronous communication, is of limited usefulness for researchers of contemporary, often multimodal, CMC.²

In addition to the problem of reduced applicability, the findings of earlier research have since been challenged by research from the previous two decades (Herring, 2001; Postmes et al., 2000). As early as the mid 1990s, the characterisation of CMC as impersonal, ineffective and hostile had come into question: Walther (1996) argued that many early findings were caused by researchers' placing of time restrictions on the observed CMC interactions. If CMC interactions are given unlimited time, Walther contends, they can achieve the same level of depth as face-to-face scenarios, both in terms of task-completion and in the development of social relationships. In fact, Walther notes the potential for *hyperpersonal* interaction in CMC: due to an optimised presentation of the self (now often called *self-curation*; see Van Kleek, Smith, Shadbolt, Murray-Rust, & Guy, 2015), as well as idealisation of the Other, CMC groups were found to be more polite and intimate than a face-to-face counterpart. Walther thus recognises two critical facts about interaction online: first, computer-mediated interaction can be more candid than what could be observed in a comparable face-to-face setting; second, relationships develop in CMC just as they do offline, necessitating longitudinal research.

The notion of a genderless and egalitarian Internet has faced similar scrutiny: CMC research has shown that gender in anonymous, text-only CMC is often encoded by the lexicogrammatical choices of the writer, as well as through the use of gendered discourse strategies such as assertiveness, politeness and aggression (Herring, 2000). Likewise, education level is conveyed through vocabulary and complexity of message structure, and age through the discussion of interests and life experiences (Herring, 2001). These findings echo the systemic-functional notion that *context is in text*, rather than around it, and can thus be reconstructed from linguistic analysis (see Section 3.3, as well as Eggins, 2004). Finally, research from the perspective of critical discourse analysis (CDA) has questioned the notion of an egalitarian Internet at two separate levels. At the level of discourse, newer studies show that complex and rigid power structures exist online, even in low-bandwidth forums and discussion lists (e.g. Stommel & Koole, 2010). More broadly, contemporary theorists acknowledge that the landscape of CMC has 'inherit[ed] power asymmetries from the larger historical and economic context of the Internet',

with a notable over-representation of English speaking, white males positioned as moderators, webmasters, and page creators (Herring, 2001, p. 12).

2.1.2. Contemporary CMC and its affordances for research

CMC has now become a central component of daily life, accounting for an ever-increasing proportion of all human communication. Today, CMC is used to contact those already close to us, rather than like-minded strangers. Instead of well-defined sessions of CMC at a home computer, CMC is now dispersed through work, travel and social occasions, and facilitated by an interconnected ecosystem of smartphones, tablets, laptops and desktop computers. Media convergence has led to new genres of communication, between journalists and readers, or companies and customers. The current Web therefore provides language examples that cannot be obtained through other means (Harvey, 2012), or which have no antecedent in face-to-face discourse (Herring, 2011). As such, analysis of CMC texts becomes necessary for investigation of many longstanding aims of linguistic research, including how language changes or evolves, how language is used to form and attend to social relationships (Canary & Yum, 2015), and/or how language is used to construe the human experience of the world.

2.2. Healthcare and online communities

Health discourse, in some form or another, is a large part of the landscape of contemporary CMC—72 per cent of adult Internet users in the U.S. report having searched for health-related content online (Fox & Duggan, 2013; Fox, 2014). As mentioned in Chapter 1, talk about health covers a spectrum of registers (in the systemic-functional sense—see Halliday & Matthiessen, 2004, as well as Section 3.3). Ideationally, healthcare is a key topic online, with information about every conceivable symptom, illness and treatment strategy readily available through search engine results. Textually, health is represented in every popular mode of CMC, including wikis, blogs, online news, chatrooms, static information pages and mobile apps. Interpersonally, online health information targeting consumers³ may be authored

by governments, non-profit organisations, health professionals, researchers, journalists, or, importantly, by healthcare consumers themselves. The latter of these—personal experiences with healthcare, authored by those living with health problems, and those who care for them, are in significant public demand: a quarter of surveyed U.S. adults report specifically seeking out consumers' subjective accounts of their experiences with illnesses and journeys through healthcare systems (Fox, 2014).

Much consumer-generated talk about health goes on inside dedicated online communities—that is, ‘mediated social spaces in the digital environment that allow groups to form and be sustained primarily through ongoing virtual communication processes’ (Shen & Khalifa, 2013, p. 986). These communities most commonly exist within social networking sites (e.g. Facebook groups, Subreddits) or within bulletin board/forum platforms—text-based modes of CMC where registered users can create and reply to threads. In academic literature and beyond, health-oriented forums have been conceptualised as OSGs—an online permutation of more traditional support networks for people living with health problems, or for those who care for them.

Forum-based online communities and OSGs have long been used as data sources for linguistic research due to their size, ubiquity and the ease with which they can be accessed: access to forums is round-the-clock and global; recording and transcription are unnecessary; and in many cases linguistic data is accompanied by rich, well-structured demographic metadata (Leech, 2006). In general, empirical studies have shown that when contrasted with face-to-face equivalents, online communities have fewer barriers to entry, higher dropout rates and a comparatively high proportion of peripheral membership—that is, a greater rate of inexperienced or new members, compared to longstanding veterans (Sandaunet, 2008; Zhang & Storck, 2001). That said, the efficacy of comparisons and contrasts between face-to-face and computer-mediated communities has recently been questioned, as daily life becomes increasingly mediated by and merged with digital technologies (Wu, 2013). Indeed, a great number of users of OSGs have never attended the offline antecedent of the mode.

Linguists have paid attention to the central role played by language in online communities and OSGs. It has long been argued by researchers that the limited

semiotic resources available in many online environments (chiefly, a lack of physical co-presence) makes language the most suitable resource for identity construction and role-relationship negotiation (Thorne, 2008). As Lam (2008, p. 303) notes, ‘language practices are instrumental in creating the norms of behavior of particular online groups and how these norms function to provide sociability, support, information, and a sense of collective identity’. A similar perspective is provided by Postmes et al. (2000), who argues that social identity issues have a stronger presence in CMC environments, due to the de-individuation that occurs in part due to reduced cues online. Because OSGs facilitate intra-consumer exchange, and because of the centrality of language to this task, the main thrust of linguistic online community and OSG research has been toward language as an interpersonal, rather than an ideational resource—that is, toward the ways in which language is used to enact social relationships, rather than as a means of construing reality. Many studies have focussed on differences in communicative practices over the course of membership—most typically, on how newcomers position themselves as legitimate prospective members whose contributions deserve replies, and how longer-term members represent their expertise and construct/reinforce normative community values. In the sections that follow, I review key themes in linguistic accounts of interpersonal meaning-making in online communities, with preference given to forum-based or health-oriented communities where available.

2.2.1. New members, first contributions and legitimacy

People can sign up and begin contributing to most OSGs at any time. A common practice is for new users to create a new thread, which serves as a self-introduction, often outlining the user’s motivation for having joined. Because drop-out rates are high, such ‘first post threads’ are a constant feature in the landscape of many popular forums, and have thus received attention during studies of OSGs.

There are broad interpersonal similarities between first contributions within many OSGs. Generally speaking, new users attempt to carve out a social position from which it is possible to elicit certain kinds of language from others—that is, to increase the perlocutionary force of one’s own utterances (Austin, 1975; Roberts &

Bavelas, 1996). Most commonly, new users want to be given information and social support after having requested it. This has been referred to as a position of *legitimacy*, arrived at through an ongoing process of *legitimation* through semiotic exchange (Davies, 2005; Smithson, Jones, & Ashurst, 2012; Van Leeuwen, 2007). In comparison to offline support groups, where physical co-presence and extralinguistic communication can serve legitimating functions, new contributors to OSGs instead exchange meaning and communicate legitimacy almost exclusively through the linguistic content of their posts (Galegher et al., 1998). The kinds of language choices that manoeuvre the user into the legitimate position may vary, with individual community cultures shaping what messages contain, how messages are structured, and the kinds of replies they receive (Gallagher & Savage, 2015).

Legitimation strategies in newcomer talk

Because texts are structured in order to make things happen, and because one of the functions of first posts is to legitimate the self, it is possible to look within the structure of these texts to see how legitimisation is realised at the strata of lexicogrammar and discourse-semantics. Commonly, first posts are designed to demonstrate the fulfillment of explicit, implicit or perceived community membership criteria. Varga and Paulus (2014) qualitatively analysed more than 100 first posts to an OSG for grief, unpacking the ways in which new users socially construct grief in a way that elicits useful responses. Three main strategies were identified: presentation of an atypical story, presentation of an uncontrollable emotional state, and through *troubles telling*—posts in which the new user explains a problem but does not explicitly request advice.

Similarly, Smithson et al. (2011a) describes the legitimating function of first posts to an OSG for self-harm: new members were observed setting out their credentials for group membership, giving narrative medical histories, utilising medical jargon and making reference to other related communities in which they have participated. This finding is echoed by Varga and Paulus (2014, p. 5), who find that ‘story formulations often serve particular functions in discourse, such as displaying affiliation with a group and establishing eligibility for group membership’.

Horne and Wiggins (2009, p. 173) use discursive psychology as an underlying theoretical framework to analyse the structural and functional components of first posts to a suicide OSG. They explore the difficulty faced by members who present as suicidal, but whose continued presence in the forum casts doubt upon the legitimacy of their claim. Three major types of messages are identified: *life narratives*, in which a medical history is presented without a specific addressee; *immediate threats*, which are typically short, containing present and future tense; and *requests*, which involved requests for advice, and the use of mainstream medical terminology. The latter category received the fewest replies. The authors hypothesise that this is due to a lack of urgency within the lexicogrammatical choices of *request* posts, leading to an impression that the new member is ‘inauthentically suicidal’ (2009, p. 180). At the same time, explicit requests for advice construct potential responders as equally inauthentic, as it casts them as using the OSG for purposes other than support seeking.

Stommel and Meijman (2011) adopt conversation analysis (CA) as a way of showing how new users in an eating disorder forum engage in self-legitimation by representing themselves as having been formally diagnosed with a disorder. While having a diagnosis is not an explicit community rule, the authors argue that it nonetheless functions as an ‘entry ticket’ for continued participation within the group (2011, p. 6). Though this is an interesting suggestion, so far lacking is a detailed investigation of the syntagmatic behaviour of the diagnosis event as it is construed in OSG texts: if the process of diagnosis functions as an ‘entry ticket’, we might expect that it is construed metaphorically as a participant, so that it can be classified and possessed.

In the same paper, Stommel and Meijman also show how newcomers often express reluctance to participate and insecurity concerning the content of their first post. These hesitations are marked in the lexicogrammar, by modal auxiliaries and adjuncts, and by ellipsis, which may be marked graphologically with ellipses (e.g. *I'm still a bit unsure about what I should write ...*). The authors argue that marking hesitation allows the new user to appear equitable and well-prepared. At the same time, hesitation is argued to indicate a level of respect for the opinions of the rest of the community, communicating an intent to take replies seriously.

Structure of first posts

Some authors have commented on the overall structure of first posts. Varga and Paulus (2014), for example, note that new threads posted to the OSG for grief often conform to common structure:

We found that newcomers opened their initial posts with stories that began at the event of loss and then moved to establish the background of their relationship with the deceased. Emphasizing the unusual circumstances of their loss and the depth of their connection with the deceased provided an account for their grief. Newcomers continued their accounts through descriptions of their uncontrollable emotional and physical symptoms, which worked to display affiliation with members of the group and to make the case for their legitimate entry (2014, p. 5).

Similar structural components—the narration of a personal (medical) history, and its lead-up to a current problem—have been identified in other analyses of OSGs. Weber (2011, p. 4) analyses the role of dispute and conflict in the socialisation process of new members of an online sexual abuse support group. She provides a basic account of the generic structure of newcomers' posts:

Contents typically found in newcomers' messages include: a greeting; a description of the person's contact with the group thus far; a reference to sexual abuse experiences or related problems; and a request.

Using terminology from Goffman (1959) and Brown and Levinson (1987), Weber makes a case that this structure represents an *entrance frame*: during each component of the frame, devices such as humour, insecurity, and unease can be used to both perform identity and to set up an exchange with a lessened potential for loss of face or redress.

Using the framework for narrative stage analysis outlined by Labov and Waletzky (1967), Kouper (2010) provides an account of the structure of initial posts to a *LiveJournal* community. She notes that messages contain sequences of ORIENTATION, PROBLEM DISCLOSURE, REQUESTS (for advice and for information; of varying degrees of explicitness) and JUSTIFICATION for posting. Because there was no hard limit on the size of a contribution (as is typical of OSGs), all stages aside from the ORIENTATION can occur multiple times in a single text. No component was found to be obligatory in every message.

Limitations in current understanding of newcomer talk

A potential methodological issue in the work of Horne and Wiggins (2009) (as well as others, e.g. Stommel & Koole, 2010) is that the authors treat the presence or absence of replies as indicators of a post's 'success'. While this may be a sensible or convenient assumption when doing larger, quantitative-based studies of number of replies, it is perhaps less reliable in small-scale qualitative research: there is no explicit evidence concerning whether or not replies would have been posted if the first post were written differently. Moreover, when using publicly available forum data, there is no reliable way to tell that posts were even seen by those who would be likely to reply. Another key factor in whether or not replies are received is the user's profile: Feng, Li, and Li (2016) find that OSG participants with recoverable first names and portraits as avatars receive friendlier, more personalised replies than do more highly anonymised users. In general, however, this factor has not been taken into account in related literature.

A second key issue in the study of new members and initial contributions is the potential for a user's first experience with the community and his/her first contribution to it to be conflated. Though first posts represent the first time a user actively engages in a discussion, he/she may not in fact be new to the community: a user may have signed up or read through threads for any length of time before choosing to post. Some first contributions, therefore, are made by people who have just discovered the community and are unaware of its normative communicative practices, while others are made by members who are already intimately familiar with the community's discursive norms (Dennen, 2008; Han et al., 2012; Preece et al., 2004; Smithson et al., 2011a). Weber (2011), for example, shows that long-time lurkers often draw upon the normative genre and register appropriately when they first decide to author a post. Such users, she finds, often flag in their posts the fact that they have lurked for extended periods before posting, in order to account for their level of familiarity with the lexicogrammatical, discursive or ideological orientation of the group. The phenomenon of lurking therefore poses a specific challenge to particular theoretical conceptualisations of group dynamics. Socialisation theory, for example, in stressing the notion of learning through active participation, may become an

unsuitable model for interpreting communities where a great deal of learning may take place non-interactively.

A final limitation in existing work on first posts to OSGs is the dominance of bottom-up, CA-informed approaches. While CA is certainly suitable for exploratory, qualitative and discourse-analytic work, it is not intended to uncover generalisable relationships between linguistic system and language instance. For this reason, CA is not typically used in large-scale quantitative investigations, which access discourse through automatic location of lexicogrammatical phenomena, and manual abstraction of the meaning of these phenomena via a grammar. CA, therefore, while useful as a way of identifying legitimisation strategies, has not provided an explicit description of how legitimisation may be realised in words and wordings. This limits the ability to perform automated feature discovery or frequency counting, as well as the ability to uncover register differences based on fluctuations in frequencies of lexicogrammatical phenomena.

2.2.2. Veteran membership and legitimacy

A smaller proportion of OSG legitimacy research has concerned the language use of veteran members, especially as replies to initial contributions (Paulus & Varga, 2015). As noted in Section 1.1.1, these studies address a hypothesised concern from earlier CMC theory that members may take advantage of the lack of social cues in the online community and linguistically convey a level of expertise or legitimacy at odds with their actual level of health literacy or knowledge (Varga & Paulus, 2014). Evidence regarding the existence and ramifications of this phenomenon is conflicting, however (Sillence, 2013). Hoch, Norris, Lester, and Marcus (1999), for example, found that six per cent of all advice provided in an epilepsy forum was objectively wrong. Similarly, Hoffman-Goetz, Donelle, and Thomson (2009) concluded that nine per cent of information in a diabetes community deviated from clinicians' guidelines. On the other hand, Sillence's study of a breast cancer forum found that 'only a very small amount of messages observed in the present study reflected a lay belief or disbelief in [patient-controlled analgesia] treatment' (2012, p. 8). Smithson et al. (2011b) reported that in contrast with researchers' expectations, replies to

new members of a self-harm forum were almost always found to be surprisingly mundane and conservative in nature, with ‘go and visit the GP’ being by far the most common advice dispensed. Research has also yet to show a conclusive link between incorrect or poor quality information online and negative treatment outcomes. As Wang, Kraut, and Levine explains:

It is highly likely that the effectiveness of such groups depends on the communications that members exchange with one another, but surprisingly little systematic research has been devoted to specifying how the quality and quantity of such communications affect groups’ outcomes and members’ health-related outcomes (2012, p. 1).

In general, analyses of OSGs have shown that long-term forum members’ language use differs from that of newcomers in a number of respects. Veterans are more likely to welcome newcomers, speak on behalf of the forum and its other members, dispense advice and instructions, and/or act as gatekeepers by ratifying or rejecting membership bids (Paulus & Varga, 2015; Pederson & Smithson, 2010; Weber, 2011). As such, these members are usually understood as being in a position of power, or higher social standing, than the newcomers they address.

Inspired in part by SFL theory, Van Leeuwen (2007, p. 92) proposes four major discursive strategies for legitimisation used by those in positions of power: *authorisation* (reference to ‘persons in whom institutional authority of some kind is vested’, potentially including both the speaker him/herself and/or those with whom he/she agrees), *moral evaluation* (reference to venerated social values), *rationalisation* (references to hegemonic social action) and *mythopoesis* (narratives casting legitimate action and those who perform them in a favourable light). Informed by SFL and CDA, Reyes (2011) augments Van Leeuwen’s framework for the purposes of accounting more specifically for the ways in which social actors construct legitimate action through the use of emotion, the description of hypothetical future outcomes, and through displays of altruism.

The kinds of legitimisation strategies described by Van Leeuwen (2007) and Reyes (2011) have been empirically observed in the language of those who respond to newcomers’ posts. Particularly relevant is legitimisation via reference to personal authority. First, a number of studies have highlighted the framing of health professionals as authorities. Smithson et al. (2011b) and Vayreda and Antaki (2009) have found that

advice dispensed by veterans often aligns with a biomedical ideology, in which the health professional is the definitive source of knowledge, as well as the agent behind critical points in the consumer journey such as diagnosis. Second is the framing of the self and other veteran members as authoritative. Van Leeuwen distinguishes between three different subtypes of authoritative legitimisation: PERSONAL (in which the authority figure and the target are in a culturally recognised hierarchical relationship, such as student/teacher or child/parent), EXPERT (where the authority figure has relevant experience that is lacking for the target) and ROLE MODEL authority, where actions are justified on the basis that they are also performed by people that the target may respect or admire). Grammatically, authoritative legitimacy often involves verbal and mental processes with the authority positioned as Sayer/Senser. The personal subtype is likely to also involve high-obligation modality (*she said that we should go*—see Section 3.3 for an elaboration of the SFG). Veteran OSG members have been shown to variously fulfill each of the three subtypes. They may explicitly moderate problematic content or ban users who break rules (Weber, 2011). They may provide expertise by providing health information in an objective, impersonal Tenor, as a set of facts (Kaufman & Whitehead, 2016). Alternatively, they may self-present as role models, providing health information and advice alongside personal narratives from their ongoing consumer journey (Koteyko & Hunt, 2015).

Kaufman and Whitehead (2016) use a combination of CA and discursive psychology to analyse replies in an OSG for depression and their usefulness as social support devices. They note that responders to first posts communicate empathy through a two-part structure of explicitly claiming to feel empathy (e.g. *I feel the same way*) followed by a demonstration of a shared trouble, and, potentially, a construal of the speaker as role model. An example from their data demonstrates this structure clearly:

I went through the same things that you are going through right now about 2 years ago. I PROMISE you that things will get better. The way that got me out of being depressed is by taking out my anger and sadness on making music or writing stuff down on some paper (2016, p. 8)

Unlike most OSGs research, the authors problematise the distinction between information and support provision: in cases such as depression, psychoeducation can

relieve symptoms by normalising the experiences of the addressee. As such, even objectively presented health information functions as a gesture of social support. Likewise, the authors point out the reciprocal nature of offers of empathy within the depression community. Those who respond to new threads (most typically, veteran members) are also benefiting from the exchange, in being given a platform and space in which they can engage in a kind of talk therapy. As such, provision of support by veteran members may simultaneously affect the second person (the addressee), the third person (other community members and readers) and the first person (the self). More directly, Pudlinski (1998) has shown that sharing related stories may in fact elicit direct social support from the addressee. In this way, part of the motivation for the responder is to receive the same manner of support that he/she is providing in the response.

Advice

Either explicitly or implicitly, studies of replies and/or veteran member behaviour in OSGs have often centred on the notion of *advice*. Defined here as ‘opinions or counsel given by people who perceive themselves as knowledgeable, and/or who the advice seeker may think are credible, trustworthy and reliable’ (a definition taken from DeCapua & Dunham, 1993, p. 519), advice is interesting due to its inherent ties to legitimacy, and due to its multifunctionality: advice has both interpersonal purposes (to cause the addressee to do something; to negotiate power dynamics) and experiential purposes (to construe facts and/or ideal behaviour) (Heritage & Sefi, 1992). Peer-to-peer advice is also important for clinical research, given its potential influence over users’ healthcare decision making practices (Jones, 2013; Sillence, 2013).

Linguistic perspectives on advice provision in English have been provided by Hudson (1990), DeCapua and Huber (1995) and DeCapua and Dunham (1993). Generally, the focus of these accounts has been on the directness of advice, whether or not it was solicited, and the influence of these factors on how it is received. Common is the recognition that advice is potentially face-threatening, and as such, its realisations are diffuse within the grammar. Advice is often accompanied by hedging strategies,

including incongruent mood choices, modalisation, or politeness markers: ‘even solicited advice-givers can resort to a variety of stylistic and linguistic means to protect themselves from accusations of unfairly taking on the role of ratified experts’ (DeCapua & Huber, 1995, p. 126). A given instance of advice may therefore have agnate realisations as an imperative (*Go and make an appointment*), a declarative (*You should make an appointment*) or an interrogative (*Could you make an appointment?*).

Kouper (2010) investigates an online motherhood community hosted by *LiveJournal*, focussing on both realised forms of advice, and on the structure of the text to which the advice giver is responding (see Section 2.2.1 for studies of first posts). Occurrences of advice in a one-month sample of contributions were coded according to four categories:

1. Direct advice (Any comment that included imperatives or the modal verb *should*)
2. Hedged advice (Any comment that contained explicit hedges, hedging devices, or softeners of various types)
3. Indirect advice (Any comment that had no explicit or hedged advice, but had enough information to act on it)
4. Description of personal experience (Any comment that had no explicit, hedged advice, or indirect advice, but had an account of how the person dealt with the situation an advice seeker had described) (adapted from Kouper, 2010, p. 7)

The categories are primarily designed to distinguish directness of advice, with grammatical distinctions naturally playing an important, but secondary role. Notably, imperative and *should*-modalised declarative are understood as equally direct—a classification at odds with the SFG, where MODALITY is a more delicate grammatical component than Mood Type. Also of interest to the case study of this thesis and its associated methods is the most indirect kind of advice, which has no explicit grammatical criteria. An example is provided:

beautiful becca :) i gave wesley a bath every 2-3 days. but now he is 10 weeks old and i bathe him every night because he loves his bath time so much Kouper (2010, p. 13).

Naturally, such indirect realisations are a complicating factor for CL approaches to advice, which may struggle to automatically locate instances that are not indexed

by any one particular lexicogrammatical feature. This is discussed in more detail in Chapter 8.

Advice research has also highlighted the fact that strategies for advice provision are affected by the roles and relationships of interactants. DeCapua and Huber (1995) show how the use of unhedged imperatives, for example, is more likely in contexts where the difference in status between advice giver and receiver is obvious, explicit and/or institutionally prescribed. Interactions between speakers of similar social standing are more likely to feature more hedging, including modalisation, and declarative Mood. In such contexts, DeCapua and Huber note, advice giving and requesting can foster interpersonal closeness through (i) the determination and exchange of shared values, (ii) flattery, by a positioning of the advice giver as knowledgeable, and (iii) potentially intimate personal revelations.

Advice online has been investigated extensively by Locher (2006; 2010), focussing mostly on an online advice column called *Lucy Answers*. Locher's interest is primarily in what she calls *interpersonal pragmatics* or *relational work*—what may in SFL be understood as *interpersonal discourse-semantics* or *meaning*. Her approach relies on manual annotation of various syntactic and semantic features of the questions asked to the columnist and her replies.

For her analysis of relational work, Locher codes the columnist's responses with seven relatively broad semantic categories: *bonding*, *boosting*, *hedging*, *praising*, *emphasizing*, *criticizing*, and *humor*. Categories may contain subcategories (to reflect, for example, the hedging function of humour), and multiple categories may be applied to a part of a text. She finds that hedging is a very common strategy, related to the overall genre of the column. Lucy the advice giver fosters a sympathetic relationship with readers and contributors, rather than a hierarchical one. Hedging lessens the difference in social status that may stifle the candid kinds of language use that the column's readers enjoy. These categories differ in the extent to which they can be mapped to particular features of the lexicogrammar: Locher's examples of hedging are typically examples of modulation (*maybe*) or modulation via grammatical metaphor (*It is important that you ...*); *humor*, however, mostly relies on the judgement of the analyst, the recognition of wordplay, and the like. Because these kinds of

features differ in the extent to which they are indexed in grammar, they also differ in the extent to which they can be identified automatically using computational tools and methods.

Locher (2006) also employs counting of lexicogrammatical features, which can more reliably be counted using methods from CL. When analysing the ways in which Lucy dispenses advice, Locher codes texts for Mood Type, finding that declaratives, often modalised, are the most common (52%), followed by imperatives (36%), and interrogatives (11%). These results are contrasted with other studies of advice in contexts such as face-to-face or radio, which more commonly feature imperative realisations. She argues that the orientation of the advice column ‘toward facilitating decision processes’ (p. 132) is responsible. In situations where the potential for losing face is less critical than effective prescription of a course of future behaviour, direct forms are likely to be preferred.

There are two major limitations in Locher’s work for investigating advice in online health communities. First are apparent inconsistencies and contradictions in Locher’s lexicogrammatical analyses. In the context of advice, Locher analyses declarative as being ‘suggestions’, interrogatives as ‘inviting future action’, and imperatives as ‘directives’. Because some examples of declaratives appear to be directives (*You need to go ...*), however, Locher reclassifies declarative + *need/should* as imperatives, arguing that ‘in written sources such as advice columns, a preference for imperatives and imperatives with *should* or *need* was found’ (p. 39), despite the fact that *should* cannot grammatically modalise an imperative clause. This is a conflation of grammatical Mood Type (imperative) and the speech function it typically realises (a command). A related issue is the conceptualisation of *hedging* as a phenomenon distinct from Mood Type choice. Advice such as *See your doctor*, when realised as an interrogative (*Would you consider seeing a doctor?*) is perhaps better understood as a hedged form of the congruent command—especially in the advice column genre, where the question cannot possibly receive an answer.

The second issue is that there are a number of critical medium and situational factors that differentiate online (forum-based) communities from the *Lucy Answers* online advice column. These differences can be understood with reference to the

register dimensions of Tenor, Mode and Field from systemic linguistics (Halliday & Hasan, 1989). In terms of Tenor, the advice column is an interaction between a pseudonymous advice seeker living with a health problem and a named, institutionally prescribed expert, who is responsible for selecting questions for publication and response. This power imbalance is likely to manifest in the ways in which advice provision is distributed amongst MOOD and MODALITY systems: following DeCapua and Huber (1995), we could expect that imperative provision of advice, for example, would be more common within the *Lucy Answers* dataset than in most online forums, where contributors share a common identity as sufferers of a condition, differentiated hierarchically only by membership length and number of posts, rather than by codified institutional roles. In terms of Mode, interactions within the advice column are limited to an initial question and a single response—in many respects, a reconfiguration of the advice column genre found in newspapers and magazines (see Herring, 2011). As such, there is no longitudinal development and negotiation of role-relationships between the two interactants, nor any way to track the health journey of the advice seekers after their contribution has been made. Compared with Tenor and Mode, the dimension of Field remains more or less consistent, however, with both *Lucy Answers* and online health forums broaching a broad array of topics related to not just health, but work and social relationships as well. Though *Lucy Answers* is unlike most health forums in not being dedicated to a single health concern, this is likely to affect only the ideational meaning embedded in advice (i.e. the specific health problem and remedial action construed), and not the ways in which advice is interactively exchanged.

Related to Locher's work is that of Harrison and Barlow (2009, 1), who investigate the linguistic negotiation of role-relationships in an online community for arthritis. They categorise 255 advice moves by Mood Type using the same criteria as Locher (2006). Their findings are similar, with declaratives the most common (64%), followed by imperatives (21%) and interrogatives (15%). They argue that this range of possible realisations is linked to the multifunctionality of the act: advice provides suggestions for future action while also reinforcing the veteran's claim to authority, and thus maintaining the hierarchical relationship between newcomer and veterans:

[Instances of advice] embody suggestions, while at the same time evidencing the writer's authority to make these suggestions: the narrative demonstrates the writer's experience of the recipient's problem, and the writer's chosen way of addressing the problem. Thus, through their narratives, the advice givers reflect on and give structure to their own experience, constructing their identities as expert patients (Harrison & Barlow, 2009, 1, p. 107).

While this study addresses the aforementioned concerns regarding the generalisability of Locher's work, and while it acknowledges the concurrent unfolding of interpersonal and ideational meanings in language, it nonetheless still provides only a general impression of the MOOD and MODALITY probabilities of advice in online health forums: as with Locher (2006), Harrison and Barlow do not provide a longitudinal account of interpersonal meaning-making, nor an empirical account of the relationship between interpersonal status and lexicogrammatical choices.

Particularly relevant to this thesis is the study of a bipolar disorder forum undertaken by Vayreda and Antaki, which uses CA to explore 'an apparent contradiction between a new user's first post and forum members' replies with ostensibly unsolicited advice' (2009, p. 931). It was found that new forum users rarely explicitly asked for advice, but were provided with it regardless. The authors contend that advice was thus unsolicited, but complementary nonetheless: new users made a 'low bid' by only vague specification of the kinds of replies they sought; established users took this opportunity to direct the user to shift toward group norms (2009, p. 940). This contrasts with Kouper's finding that most advice was in fact solicited by the user who began the thread. This contradiction is perhaps the result of the fact that Vayreda and Antaki take a very narrow view of what exactly constitutes a request for advice, discounting formulations in which requests may be incongruently realised, as in the following:

'It'd be great if someone could share this first stage of acceptance with me, or tell me how they got through it.' (2009, p. 11)

Here, though the new user chooses a conditional declarative rather than a modalised interrogative as a means of realising a request, given that requesting advice involves deference and a potential loss of face (Brown & Levinson, 1987), such hedging does not seem inappropriate. Aware of the fact that expert members are fluent in the community's discursive norms, a new member could perhaps expect that such heav-

ily modalised constructions could be unpacked and decoded as requests for advice. As Goldsmith (2000) points out, while overt requests for advice open up a dynamic in which advice can be provided without threatening face, general solicitation of opinions on a topic are often understood by the addressee as permission to advise. In a hierarchical community of newcomers and experts, this is especially likely, as the new user understands that the veteran has personal experience with the nature of the problem. As Eggins and Slade (2004) remind us, rather than relying on linguistic intuition alone, the best way to disinter the intentionality of an utterance is often to analyse the utterances that follow. Given that such responses are apparently treated as advice by those who respond, this may suggest an underlying shortcoming in Vayreda and Antaki's classificatory scheme.

The lay-expert/proto-professional

In medical contexts, a well-noted strategy for legitimisation is the invocation of a *lay-expert* or *proto-professional* register, which is realised at both lexicogrammatical and discourse-semantic levels. Lexically, for example, such a register may involve the use or appropriation of medical jargon in lieu of lay terms (Harvey, 2012; Sullivan, 2003); discursively, the lay-expert has been shown to foreground personal experience and display emotional affect when construing medical fields (Wilson, Kendall, & Brooks, 2007). This conceptualisation, of a social role and associated linguistic repertoire that lies between novice and expert, is one useful lens through which we can understand language features of long-term OSG members.

Thompson et al. (2012) examine the discursive features of lay expertise through interviews with members of *patient and public involvement panels*—laypeople consulted by medical institutions in the U.K. in an attempt to involve the public in the medical research process. They noted that those in lay-expert positions may champion their lack of formal medical training or employment, arguing that it affords a unique point of view that is uncorrupted by financial or socio-political factors. Simultaneously, a high degree of deference to the opinions of health professionals was also observed: interviewed participants 'not only supported the dominant techno-scientific discourse around research, but also seemed to defer readily to

it in place of their own experiential expertise' (2012, p. 609). That said, these findings may be strongly influenced by the social context in which the interviews took place: as the participants of the study were afforded privileges and treated as having 'honorary' roles by health professionals, their construction of both their unique perspective and health professionals' ultimate superiority may be influenced by the desire to maintain their current position and role. Indeed, studies of OSGs where normative community values oppose those of mainstream Western medicine have noted that formal healthcare institutions, their participants and processes (e.g. hospitals, health professionals and diagnosis) are often cast in a negative light and treated sceptically by community members (Mulveen & Hepworth, 2006). One of the advantages of OSGs as sources of data, therefore, is access to the lay-expert register uncorrupted by the presence and potential influence of healthcare professionals.

Koteyko and Hunt (2015) operationalise the notion of the lay-expert in an online context, using ethnography and text analysis to qualitatively examine the ways in which Facebook users construct a representation of the self as a lay-expert in communication about diabetes. Noting that social networking sites have received less attention within online health discourse literature than forum-based communities, they argue that such sites provide new, under-researched channels through which semiotic health resources can be exchanged. The convergence of a number of (monomodal and multimodal) types of CMC within Facebook's infrastructure (e.g. link sharing, microblogging, chat, photo albums, etc.), for example, gives rise to a novel practice where users publicly share links to medical literature, accompanied by a 'translation' or summary of its significance into simpler English for a lay audience.⁴ Lay expertise, therefore, is not simply communicated and constructed through talk, but through online activity more generally, with posting, replying, and 'liking' others' content each contributing to the construction of identity. These kinds of practices, of course, are difficult to analyse through automated CL techniques, which tend to strip away multimodal features of texts, and which may struggle to differentiate between individual voices in group discussions.

Notably, unlike most research on patient empowerment, the authors contextualise their findings against a (critical) sociological backdrop, highlighting the fact that

take-up of the notion of the active ‘e-patient’ role within formal healthcare institutions shifts responsibility for the management of risk to healthcare consumers and potentially extends the influence of medical institutions in daily life. From this perspective, the empowered, resilient narratives presented by users seen in their study are also passively reproducing a neoliberal ideology promoted by government and corporate interests alike.

Operationalising ‘veteran membership’

The overarching theme of literature concerning veteran users’ contributions to OSGs is the simultaneous provision of information and support to the addressee and legitimisation of the self and the community more generally. The action of welcoming someone to a forum not only opens up a space for positive interpersonal exchange, but also construes the forum as a community; the action of providing advice not only provides information, but also negotiates the roles and responsibilities of both the advice giver and the addressee.

A major issue that emerges in studying the linguistic choices of longstanding community members is that there is no objective criteria for determining which users qualify. Among other factors, communities differ in terms of their moderation practices, the existence of health professionals as participants, and their overall size and popularity. Therefore, each community may have different criteria for determining which user sits where on the hierarchy of experience and/or expertise. How do we rank users who appear suddenly and post frequently against those who signed up early in the community’s history, but who contribute less? Many of the approaches outlined in the previous sections for the most part avoid this issue by simply analysing replies to the posts of newcomers. The problem with such an approach, however, is that it is possible, and potentially not uncommon, for users’ first posts to in fact be replies to others’ first posts. Such posts may contain ‘me too’ sentiments, or may also engage in advice provision. The most sensible approach is to develop indigenous criteria by manual or statistical observation of the community. Multivariate analysis could also uncover which factors (number of posts, length of membership, replies received, gender, etc.) most influence the register of a post.

This could then inform development a metric that may be applicable across domains. At the same time, it is important to remember that the categories are not discrete: generally speaking, the membership trajectory in online communities is a gradient, with indeterminacy in the definitions between membership stages, if such stages exist explicitly at all.

Having reviewed literature focussed on both newcomers and veteran community members, I now shift attention to work that seeks to account for longitudinal change from one role to the other.

2.2.3. Pathways to sustained membership

As reviewed above, much research into interpersonal meaning-making in OSGs has highlighted differences in the roles and responsibilities of new and veteran members. Other work have focussed on the transition from one membership category to the other. While it is widely accepted that users' language use, as well as their discursive roles and responsibilities, can change over time, this change has yet to have been examined through quantitative analysis of linguistic features of users' posts across the membership lifecycle. Moreover, the underlying causes of this change remain debated, with plausible interpretations offered by psychology, sociology, linguistics, and by computational models of group dynamics (Kouper, 2010; Preece & Maloney-Krichmar, 2005).

Socialisation

One theoretical framework for interpreting longitudinal linguistic change is *socialisation*—that is, the ways in which newcomers or novices learn through participation in meaningful social interaction with more experienced members of groups (Ochs, 1991).⁵ *Groups* is to be interpreted very broadly: socialisation research may concern groups of any size and type, from small families (where children may learn family values from parents), to communities of practice (where a new employee learns terminology from more senior employees), to entire language groups (where non-native speakers of English students visit an Anglophone country) (Schieffelin & Ochs, 1986). This broad scope has made socialisation research an interdisciplinary area,

spanning psychology, anthropology, sociology and applied linguistics (Duff, 2010, p. 172).

Linguistic accounts of socialisation have shown that it may take place at any stratum of language. Socialisation may take place phonologically, as shown in Polat's (2011) study of Kurdish students' acquisition of Turkish accents. Socialisation at the lexical level has been studied predominantly in occupational settings (e.g. Wolf, 1989), where jargon is transmitted from older to newer employees. In terms of language function, parents have been observed socialising their children to linguistically perform indicators of emotional affect (Clancy, 1999). At more abstract strata, theorists have highlighted 'political socialisation' in youth groups (Lee, Shah, & McLeod, 2013) and 'ideological socialisation' in medical students (Harter & Krone, 2001). As Duff (2010, p. 172) explains, even socialisation targeted at micro-levels simultaneously involves a broader cultural dimension: through interaction with others, learners must necessarily be exposed to meanings about 'normative, appropriate uses of the language, and of the worldviews, ideologies, values, and identities of community members'.

The Vygotskian origins of socialisation in child language acquisition theory have led to a strong focus in linguistics on socialisation in (both first and additional) language learning contexts (Ochs, 1991). To date, the bulk of such literature deals with offline contexts—though CMC research is now a well-established trend. Socialisation of students toward academic norms has been another common interest, as discursive norms between everyday life and graduate study are commonly seen as extremely contrastive (Beckett, Amaro-Jimenez, & Beckett, 2010). Such studies have highlighted socialisation toward specific elements of both an academic register (with lexicogrammatical features including nominalisation, passivisation, use of relational verbal groups) and metadiscourse (Mauranen, 2003).

Academic discourse socialisation has also been investigated in online contexts. Beckett et al. (2010) used posts to academic discussion boards alongside interviews and surveys to investigate the socialisation of early graduate students to 'graduate school language and culture by old-timer 'expert' second-year master's and doctoral students and their professors' (2010, p. 319). Though the central concern of the

study was students' perceptions of online learning, a key finding related to how they learned was that first year graduate students tended to relate theory to personal experiences and anecdotes, while more experienced students were more focussed on the content itself—in systemic-functional terms, an orientation toward interpersonal meaning was gradually displaced by an orientation toward the ideational.

Duff's work on discursive socialisation in academic environments (2010, p. 171) may also be potentially applicable to health contexts. She notes, for example, that a common distinction between academic literacy and academic discourse socialisation may be worthy of reconsideration: both concepts, she explains, 'are concerned with learning processes, with macro and micro contexts for language development, forms of knowledge and practice valued, material products or tools involved in literacy, and outcomes'. This links to key debates in healthcare communication research, where researchers are interested in generating testable definitions of *health literacy* (Frisch, Camerini, Diviani, & Schulz, 2012; Jorm et al., 2006): if socialisation and literacy are to some extent interchangeable, online communities, in providing large amounts of consumers' natural language, may prove useful as data sources that inform definitions and measures of health literacy. Duff also reminds us that socialisation is not necessarily a process involving 'mindless, passive conditioning'—it is *bidirectional*, in that novices may also socialise experts during an ongoing exchange. Such a conceptualisation is congruent with findings from linguistic analysis of consumer health discourse: the previously reviewed findings of Pudlinski (1998) and Kouper (2010), for example, showed that interactions between newer and more senior health community members may involve interpersonal reciprocity of potential benefit to both (see Section 2.2.2).

2.2.4. Discourse socialisation in online communities

As both the meaning and scope of *discourse* is broad and often contested (Gee, 2004, 2013), it is common for researchers to characterise, rather than define the term. For this thesis, it suffices to adopt Martin and Rose's simple characterisation of discourse as being 'more than a sequence of clauses' and 'more than an incidental

manifestation of social activity': discourse is 'meaning beyond the clause'; 'the social as it is constructed through texts' (2003, p. 1).

In online communities, socialisation at the stratum of discourse has been investigated from a wide array of theoretical perspectives and methodological approaches. Lee et al. (2014) approach discourse socialisation from a member-lifecycle perspective, arguing that members transition through a number of roles during their time within the online community, and that each role has accompanying needs and responsibilities. Newcomers have strong needs for both information and social support, but may not contribute due to the potential for a loss of face if information they provide is judged by experts to be incorrect (Füller et al., 2007). At later stages in the member lifecycle, users become less anxious about producing content, but lose the motivation to seek out information or support (Lee et al., 2014).

In a similar vein is the work of Danescu-Niculescu-Mizil, West, Jurafsky, Leskovec, and Potts (2013), which focusses on the user lifecycle of an online beer enthusiasts' forum using a corpus-based approach. They find that new members are responsible for both the introduction and a large amount of the uptake of new jargon (lexical innovation). The register of already established members is argued to become more conservative or fossilise over time, due to the perception that their communicative competence is already sufficient, as resistance against inbound norms, or as a means of marking their seniority.

Cassell and Tversky (2005) used quantitative analysis, content analysis and interviews with participants in a global political forum to track linguistic change over time. They identified three key changes. First, the plural first person pronoun *we* became more frequent when compared to singular *I*. Second, members gave more feedback on others' opinions, rather than promoting and explaining their own. Third, it became more common to collaborate to pursue shared goals. Smithson et al. (2011a) also address discursive socialisation in their analysis of a purpose-built self-harm OSG: normative practices emerged in the days immediately following the site's creation, with all active members expected to provide others with social support using appropriate degree of emotional affect (Smithson et al., 2011b).

Weber (2011, c.f. Section 2.2.1) focussed on the role of dispute and conflict in the socialisation process of new members of a UseNet community for sexual assault survivors, showing how disputes provide a context in which norms can be made explicit by experts. After flouting group norms, a new member is chastised by veterans: as Weber explained, ‘since she did not learn by lurking, she has to learn by direct instruction’ (2011, p. 1). After some contestation, rather than leave the forum, the new user eventually ‘apologizes, self-deprecates, claims technical and social ignorance, and highlights her need to learn’ (p. 12). It is argued that the newcomer’s radical shift in orientation was the result of a realisation that future participation in the group was contingent on her adoption of the discursive features typical of newcomers.

Vayreda and Antaki (2009, c.f. Section 2.2.2) highlighted discursive socialisation toward a biomedical account of bipolar disorder: veteran users of the forum construed bipolar disorder in the same terms as well-established International Classification of Diseases (ICD) and Diagnostic and Statistical Manual of Mental Disorders (DSM) guidelines, while stressing to newer members the necessity of diagnosis and treatment through mainstream healthcare institutions. In contrast to Horne and Wiggins (2009), who found that discussion of diagnosis led to being ignored in a suicide forum, Vayreda and Antaki demonstrate that diagnosis can in some cases be a prerequisite for legitimate community membership within an OSG: even posts displaying an alarming sense of urgency were met with terse commands to ‘get an appointment with a psychiatrist’ when the new member did not explicitly mark their status as diagnosed (2009, p. 940). They explain:

That instruction to go straight to the psychiatrist shows, in microcosm, the ideology of the forum. It crystallizes the site’s motivating spirit [...]: that only one account of bipolar disorder will be countenanced, and that is the biomedical. No time is afforded to any consideration of the user’s symptoms or circumstances. For the forum, diagnosis must be in the hands of the psychiatric profession, the ultimate authority on the illness and its treatment; the forum offers support, information and, indeed, advice, only on that basis. ... The structure of open request allows the response to choose its path. And that path is biomedical diagnosis. Once she has that, then she can enter the community of forum users and the site’s resources will be at her disposal (2009, pp. 940–941).

Here, socialisation is toward a set of ideological norms held by the community’s core members that reflect a hegemonic Western biomedical conceptualisation of

health and illness. These ideological values are realised through discourse and semantics—interpersonal and experiential meanings made in interactions within threads. In turn, discourse and semantics are realised through lexicogrammatical choices in users' posts. It becomes possible, therefore, to create an account of meaning-making in an OSG that connects instantiated words and wordings, which have stable representations in writing, to abstract community values, which are abstract and diffuse. Because instantiated language can be annotated and searched using methods from CL, it is therefore possible to automate, to some extent, the identification of discursive and ideological change in OSGs. Limitations in current knowledge, in terms of theory, methods, and tools, however, have so far prevented comprehensive account of both how ideology is represented in the linguistic hierarchy of stratification, and of how the system of language is employed by OSG users at different stages of membership. These limitations are summarised in the following sections.

Four challenges to socialisation theory

Socialisation has proven an attractive theoretical framework for understanding the nature of language change by members in communities, both online and offline. Indeed, much CMC research either implicitly or explicitly assumes that consistent change in the direction of norms practised by veteran members, or outlined in rules and FAQ pages is evidence of socialisation. This assumption can be challenged on four separate grounds, however. Accordingly, in this section, four provocations for socialisation-based approaches to linguistic change are raised, specifically with analysis of CMC in mind. Each provocation potentially challenges the explanatory power of the framework in linguistic contexts in general and in OSGs research more specifically.

The phenomenon of lurking

The first factor posing a direct challenge to socialisation theory (briefly noted in Section 2.2.1) is the fact that many users lurk in online communities extensively before posting, often familiarising themselves with community-specific values in

order to author normative content (Weber, 2011). This phenomenon is at odds with core tenets of the sociocultural hypothesis that underpins contemporary socialisation research, where knowledge development is understood as taking place through active participation with more other (typically more senior) community members. *To what extent, therefore, can we argue that socialisation is the cause of linguistic change if many newcomers have already adapted to local norms without ever having socialised?*

Scope and strata of socialisation

The second challenge for socialisation research lies in determining exactly what it is that people are being socialised *into*. Looking at the landscape of OSGs online, it is clear that most popular groups are parts of larger web architectures, containing similar communities for different conditions. Furthermore, community boundaries are often indistinct: some communities sponsor or promote groups on social networking sites; others have official or unofficial chatrooms, hosted within the same domain, or on a dedicated chat platform. The original contention of sociocultural theory—that cognitive processes are developed through social interaction—does little to account for overlapping, hierarchical or stratified institutions and communities that comprise daily life. This fact makes analysis problematic, especially given potential confounding variables, where an observed participant is also learning about his/her health condition from other sources, such as health professionals or mass media. *Given that institutions may overlap, or be embedded within one another, how can we determine the scope of socialisation, and the degree to which prior or ongoing contact with macro-institutions are the real causes of change?*

Prior knowledge of community norms

A third challenge for socialisation is that other theories of language provide alternative plausible accounts of the cause of linguistic change over the course of community membership. From a systemic-functional stance, for instance, competent language users may recognise particular register dimensions of a community, and use what is known about related registers (i.e. of offline support groups, or of support groups for different health problems), as well as more abstract contextual configurations (i.e. of interactions between new and existing community members) to determine how

language should be put to use in this new situation. That is to say, the difference between newcomer and veteran language may not exclusively be caused by the fact that the former has not been socialised by the latter, but also by the fact that new members enter groups with an understanding of membership roles and hierarchies, and choose to act in a way that is congruent with what they have previously observed. Learning may of course still take place, with identifiable traces to be found in the lexicogrammatical choices of the user as he/she continues to participate. Simultaneously, however, as the user gains experience, he/she is also able to access new kinds of meanings and their associated realisations in words and wordings that have hitherto been reserved for others, such as directing or commanding others to act in a particular way. *How can discourse socialisation research account for new members' pre-existing knowledge of group structures and hierarchies in general?*

Epistemological issues when using CMC data

The final provocation for socialisation research, more specifically in the case of analysis of CMC, is that available data may be unable to yield insights into the underlying cause of linguistic change. Without access to additional information (follow-up interviews, observation of participants outside of the context of the community, etc.) it simply may not be possible to determine *why* observed linguistic change takes place. As explained by Widdowson (2000), though *third person data* such as that extracted from online communities may give us useful information concerning attested behaviour, it cannot elucidate

the facts of what people know, nor what they think they do: they come from the perspective of the observer looking on, not the introspective of the insider. [... Third person data can only be used to] analyse the textual traces of the processes whereby meaning is achieved: it cannot account for the complex interplay of linguistic and contextual factors whereby discourse is enacted (2000, pp. 6–7).

With this in mind, it is useful to raise the question: *Can socialisation be empirically accounted for with third-person data alone?*

2.2.5. Current limitations in health discourse research

Despite an impressive amount of linguistic research into both legitimacy and socialisation, much is of only limited use for the case study of this thesis. Descriptions of new community members' legitimisation strategies have not prioritised a mapping of meaning to lexicogrammatical forms; accounts of legitimacy that map meaning to lexicogrammar (e.g. Van Leeuwen, 2007; Reyes, 2011), meanwhile, have been intended more for analysis of powerful social actors than of lower-status incoming group members. Within socialisation literature, most attention has been paid to language acquisition, offline contexts and academic discourse socialisation, with less interest overall in online health contexts.

A more serious current shortcoming, however, is that a number of potentially useful theories and methods have yet to be used to analyse discourse in OSGs. First, despite the fact that a key interest in OSGs has been their dual function as sites of social support and health information exchange, theories of language that delineate between these two metafunctions and their realisations in grammar have yet to be applied. As will be introduced in the next chapter, SFL provides an exemplary grammar for this task. In SFL, language users are understood as simultaneously attending to interpersonal and experiential meaning-making; these two functions are performed through the discrete, but simultaneously deployed grammatical systems of MOOD and TRANSITIVITY (Halliday, 2004). SFL also includes a framework for genre analysis (e.g. Eggins & Slade, 2004), which has likewise yet to be operationalised within online discourse socialisation research. This is disappointing, given that many researchers have highlighted generic structures in new members' first contributions. Delineation of genre stages and an overall generic structure would potentially illuminate an important element of discursive norms within online communities, providing a context that can inform more delicate analysis of linguistic patterns in forum posts.

A second shortcoming is the lack of quantitative, corpus-based OSG research. Though many online communities are large and well-structured enough to facilitate programmatic and quantitative approaches, there has so far been little engagement with such methods. To date, impressionistic and qualitative perspectives predomin-

ate, limiting the extent to which methods can be automated, scaled, applied to new datasets or even simply reproduced. This is generally an acknowledged limitation—a number of the reviewed papers have noted that purely qualitative research design limits the ability to make generalisations based on findings:

We acknowledge that as a qualitative research study our findings are not intended to be generalizable to the larger phenomenon of online discussion forums. The construction of grief online is only initially explored here, with a full exploration of each forum thread certain to shed additional light on the way that newcomers and established members of the community negotiate entry and membership (Varga & Paulus, 2014, p. 8).

There are various limitations to the study. Being qualitative, it was useful only for the development of a hypothesis and the findings cannot be generalized to represent/could not lead to a valid representation of/the role of self-presentation and community norms in OSGs overall (Stommel & Meijman, 2011, p. 8).

It should be noted that further research (e.g. based on larger and/or more representative samples and focusing on other peer support settings or other online forums) may be necessary in order to evaluate the transferability of these findings (Kaufman & Whitehead, 2016, p. 14).

As seen in the examples above, traditional discourse-analytic research methods may preclude the ability to generalise findings across domains. Though it is sensible to assume that similar communities (relying on similar software, concerning a related health concern, with similar moderation practices, user demographics, etc.) may contain similar kinds of language use, this cannot be demonstrated empirically without follow-up studies. Re-application of the same qualitative methods to new posts, new users or new communities, however, is not feasible *en masse*, due to the resource-intensiveness of qualitative investigation. Even when using publicly available CMC as data, analysis involves multiple close readings of texts, ideally by more than one well-trained annotator.

To test key claims in the literature more definitively, it therefore becomes necessary to use more sophisticated programmatic methods, which, rather than relying on small samples of text, can provide an account of linguistic patterns within all first posts, within all veteran posts, and within the community as a whole. Though CL, and more recently, CADS have shown promise as a means of highlighting the discursive construction of ideologies within large corpora of related texts (e.g. Koteyko,

Jaspal, & Nerlich, 2013; Salama, 2011), few elements of the approach have been operationalised within legitimacy or socialisation research.

Aside from generalisability and representativeness, the lack of quantitative approaches have other consequences. One key issue is reproducibility: due to the small sample size, it is possible that repeating the method on a similar dataset may reveal different, or even contradictory findings. Another risk is the potential for incorrect or partial analyses. Many of the reviewed studies arrive at discourse and meaning ‘from above’: human comprehension of texts leads to an understanding of key meanings, which then informs a search for examples and frequencies of lexicogrammatical and graphological phenomena in the texts. This means that researchers’ subjective reasoning enters very early in the course of the investigation, and constrains what parts of the text end up being analysed in detail. In this way, the predispositions of the researcher may corrupt the integrity of the results. At the same time, important, yet subtle patterns in texts may escape researchers’ attention, either because they may not appear salient at first glance, or because they do not appear frequently enough in the small sample to lead to identification of a pattern within the text type.

A final consequence of the methodological choices of current literature is the limited potential application of results. Qualitative and manual analysis is of little use in statistically driven research, which may attempt to identify correlations between participation in the community and health outcomes, or to predict future behaviour of a participant based on earlier behaviour and the behaviour of others. Looking further ahead, qualitative findings alone are often unable to inform healthcare practice. Though qualitative findings are useful for generating hypotheses, controlled, quantitative research is generally needed for integration into clinical protocols, medical education, and the like (Giacomini & Cook, 2000).

2.3. Computational perspectives on online communities

Work reviewed so far has come predominantly from a qualitative, bottom-up, discourse-analytic tradition. CMC in general, and OSGs more specifically, have also

received attention from within the largely separate traditions of computer science, computational linguistics and natural language processing (NLP). In general, these traditions prioritise quantitative approaches, automated methods and downstream applications (i.e. implications for practice). In the final section of this chapter, I review a small selection of recent, computationally oriented approaches to OSG analysis. An argument is advanced that such methods are a useful complement to what has already been developed within the discourse-analytic tradition; at the same time however, I explain that the computational tradition suffers from theoretical shortcomings regarding the nature of language, language users and context.

2.3.1. Analysing health discourse quantitatively

One of the key benefits of working with CMC is that it is *born-digital*, having been originally mediated by, rather than transferred to, some kind of digital storage or transmission format. This kind of data is amenable to processing by software, or via programming. To give an example, automatic word segmentation in recorded spoken dialogue is an ongoing computational challenge, while relatively accurate segmentation of natural written language into words in English is as simple as splitting strings of text on punctuation, whitespace or newlines. This word-segmentation can then form an initial step in a larger computational workflow, where text is automatically annotated with information regarding part-of-speech (POS), lemma form, and grammatical role. These affordances open up a range of possibilities for automated and semi-automated analysis of language that would otherwise require prohibitively expensive manual processing.

Healthcare figures prominently in the landscape of contemporary computational linguistics (including text mining and information extraction), just as it does in the discourse-analytic tradition reviewed in the first part of this chapter. The central challenge of the area is to exploit large amounts of readily available healthcare data in novel ways. This data may be statistical or linguistic; it may have been captured inside or outside a hospital or clinic; it may have been authored by clinician or consumer. The central hypothesis is that what is learned through computational linguistic analysis can improve healthcare practice. Velupillai, Mowery, South, Kv-

ist, and Dalianis summarise the general needs of healthcare institutions and the potentiality of NLP:

The interest for clinical NLP is spurred by the need for real-time, large-scale, and accurate information extraction from health records to support clinical care, e.g., through automated generation of a patient problem list, to support biomedical and health services research, e.g., through precise cohort identification, and to support public health practice, e.g., through disease surveillance. Clinical NLP can provide clinicians with critical patient case details, which are often locked within unstructured clinical texts and dispersed throughout a patient's health record (2015, p. 183).

One key difference between the qualitative and computational paradigms, therefore, is that computational analysis of health talk is often performed with clinical applications foremost in mind.

While researchers from both paradigms are well-aware of the potentiality of large written corpora for health research, a major current obstacle for both is the development of tools and methods that can extract useful information from this data (Paul et al., 2016; Anthony et al., 2013). From a functional linguistic perspective, these tasks are inextricably linked: key computational linguistic goals, like discourse analysis, result from analysis of meaning beyond the level of the clause, combined with the use of statistical methods to determine prototypicality of features, trends, and relationships between phenomena. Methodological advances can thus have utility within both traditions.

MacLean et al. (2015) provide a recent example of computational linguistic analysis of discursive linguistic phenomena. They investigate *Forum77*, an OSG for prescription drug abuse, automatically identifying significant events in forum users' medical timelines such as USING, WITHDRAWING, RELAPSE and RECOVERING by quantifying both metadata features (number of posts, time between posts, etc.) and clusters of lexical items. In terms of metadata features, the authors find that users in the RECOVERY stage respond to more threads than users in the USING stage. In terms of language use, the USING phase is characterised lexically by negative mental states, such as *hate, addicted, scared* and *tried*; RECOVERY, on the other hand, is more likely to be indexed by positive lexis (*sober, fight, truly, clean, true, worth*). The identified linguistic and extra-linguistic features of each phase of addiction are then used to train a Conditional Random Field model, which can automatically assign arbitrary

texts to the set of phases. This makes it possible to predict users' stage of illness/recovery based on their linguistic choices in a post. At the same time, the study provides a quantified taxonomy of stages of addiction that could potentially be used within clinical encounters. These goals represent potential downstream applications for computational discourse research that are largely unexplored within qualitatively oriented literature. The authors summarise the contributions of the work:

It is possible that data extracted from sites like Forum77 [...] could help medical professionals and policy makers better understand patients' experiences with drug abuse. For example, insight into the day to day difficulties of opioid-assisted withdrawal might inform policy for improving the management of this popular treatment down the road. It is also possible that research like ours could illuminate poorly understood aspects of addiction (2015, p. 12).

Though the paper is aligned with computational linguistic and clinical NLP domains, the main goal is certainly a discourse-oriented one: phases of addiction are determined by analysing features of language at the level of text, beyond individual clause boundaries, and by taking context (in the form of metadata) into account. It is not, in the typical sense, a discourse analysis, however, as individual texts from the corpus are not analysed in detail. It is also superficial as a linguistic analysis, as identified features are purely lexical, rather than lexicogrammatical. In later chapters, I argue that computational approaches to discourse would benefit from engagement with discourse analytic literature, and with functional linguistics more generally, which connects lexis to grammar as lexicogrammar, and connects lexicogrammar to discourse-semantics through the notion of a hierarchy of stratification.

Chancellor, Lin, and De Choudhury (2016) attempt to automatically determine which features of pro-eating disorder communication on Instagram lead to posts being removed, either by users themselves or by moderators. They demonstrate that training classifier models on content can be used to auto-moderate online communities with rules against self-harm. At the same time, the approach can be used for *just-in-time interventions*, where users posting content containing suicidal ideation may be prompted to connect to a friend, support group or hotline. This provides a useful demonstration of the power of computational approaches to OSGs, when compared to methods reliant on manual analysis: automated methods can

have applications that improve the quality of information to which users have access, and therefore, potentially affect health outcomes as well.

Other research has sought to explicitly link computational analysis of OSG contents to measurable clinical outcomes. Yan and Tan (2015) investigate the relationship between levels and types of social support and health outcomes in a weight-loss forum. On the forum being analysed, many users provided weekly updates regarding changes in weight and diet via a dedicated self-monitoring tool, making it possible to map the relationship between post content, replies received and weight-loss outcomes. Using Hidden Markov Models, a classifier is trained to score posts by their likelihood of providing and requesting social support. Two key findings emerge. First, the authors determine that either overprovision or underprovision of support to those requesting it has detrimental effects on weight-loss. Second is the influence of the *helper-therapy principle*, whereby assisting others through the provision of social support or health information can lead to positive feelings, which in turn may boost self-esteem and psychological well-being. More specifically, tangible weight-loss benefits can be observed for those who provide social support to others.

A similar research design is presented by Althoff, Clark, and Leskovec (2016), who attempt to find correlations between text-based mental health counselling sessions and consumers' evaluation of the successfulness of the interaction, as collected in a follow-up survey. The study involved the development of computational measures of discourse-semantic and registerial concepts, such as ambiguity, creativity and adaptability. The authors determine that the most successful counsellors adapt the staging and wording of the interaction when the interaction appears to be going badly, work harder to reduce ambiguous or abstract talk, and use more follow-up questions, hedges, and expressions of empathy. As with the other reviewed computational health discourse studies, these abstract linguistic notions are often simplified in order to become computationally implementable. Ambiguity of messages, for example, is operationalised by simply counting the number of tokens in the message; messages with fewer words are considered more ambiguous. Such a classification has obvious shortcomings: responses to polar interrogatives, for example, may be very short, but unambiguous. Furthermore, none of the approaches

in the study involved restricting analysis to a particular part of the lexicogrammar, instead focussing on broader patterns occurring across all of the tokens within a text.

These studies highlight two emerging methods for connecting OSG content to clinical outcomes. First is the approach of Yan and Tan (2015) and Althoff et al. (2016), where longitudinal linguistic content can be compared to a second dataset containing some kind of clinically relevant health outcome, such as the success of a counselling session or an amount of weight change. From this comparison, the health effects of participation and participation styles in online communities can be robustly quantified. Such approaches may have potential use within formal healthcare institutions, where clinicians' free text notes can be mined and linked to patient records including duration of stay in hospital and stages of treatment (Elkin et al., 2008; Miller et al., 2013).

The method proposed by MacLean et al. (2015), on the other hand, does not involve the use of an extralinguistic dataset of pre-recorded health outcomes, but instead, uses manual classification of a sample of contributions to train a machine learning algorithm that can be used to categorise unseen text. The clear advantage is that no second dataset is necessary; the drawback is that the method relies on time-consuming manual annotation of a training set of text. Striking, however, is the overlap between the task of manual annotation of training data and the kinds of discourse-oriented text analysis reviewed in the first part of this chapter: foreseeably, the manual labour involved in content and thematic analysis methodologies could be re-purposed as training data in computational applications. Such an effort, though requiring a great deal of foresight and interdisciplinary co-operation, could produce high-quality results.

Theoretical issues in computational health discourse research

Though the emerging field of computational health discourse promises added reproducibility, scalability and generalisability, such benefits have obvious associated challenges that remain unresolved. Some of these arise due to limitation in tools for automatically extracting useful features from text. Computational app-

roaches, unlike qualitative approaches, are inextricably bound to the performance of tools, which typically favour widely spoken (European) languages and registers, and formal, well-structured text over the informal kinds of language that may arise in intra-consumer talk. While developed computational tools can be re-applied to new data at virtually no expense, the initial development of the tools may be time- and resource-consuming; qualitative analysis of OSG texts, on the other hand, can be performed on any language in which the trained researcher is competent. In very large datasets, researchers also face the additional problem of not being able to read through the entire corpus manually. There always remains, therefore, a possibility that inappropriate texts are included within the dataset, or that kinds of meaning not indexed with specific lexical or shallow grammatical features such as attitudinal lexis and passive voice may go unnoticed and unanalysed.

Perhaps the most serious limitation in the computational literature, however, is that conceptualisations of language are at times drastically simplified in order to remain computationally applicable. This simplification exists both within conceptualisations of grammar and within the understanding of the relationship between text and context. The approach taken by MacLean et al. (2015), for example, treats each forum text as a bag-of-words: texts are classified by metadata feature and by the frequency of lexical items, without any consideration of how the lexical items are positioned grammatically. Topic modelling, a common approach to automated classification of texts into Fields of discourse, is essentially blind to linguistic phenomena that may assist in distinguishing topics from one another, such as lemma forms and grammatical positions (Delpisheh & An, 2014; Boyd-Graber & Blei, 2009). It also conflates (meta)functions of language. Since topic modelling clusters based on lexical realisation alone, jargonised and non-jargonised variants of a wordform may be modelled as separate topics.

Despite the promise of such approaches, oversimplification of linguistic constructs and treatment of texts as simply lists of tokens limits the ability to (for example) separate the interpersonal and experiential meanings being made by users.

2.4. Chapter summary

This chapter has provided a theoretical context of the investigation, highlighting current knowledge of language use and language change in OSGs. Specific attention was paid to health discourse analysis from the perspectives of legitimacy and socialisation theory, and to emerging computational approaches to consumer healthcare discourse. Overall, a lack of dialogue between the qualitative and quantitative traditions has meant that each has failed to profit from advances in the other.

In the following chapter, I provide a methodological context, focussing on CL and SFL—a methodological orientation and theory of language that can link the theoretical successes of legitimation and socialisation research with the recent innovations in computational health discourse analysis.

3. Methods for investigating online health discourse

The previous chapter established that valuable things can be learned through analysis of text-based CMC. These things can be both theoretical and applicable: in the context of this case study, this includes insights into the lived experience of people with illnesses, as well as emerging computational methods for identifying correlations language use and health outcomes. In this chapter, I shift attention to methods and theories of language that be used to understand and process online healthcare communication.

3.1. Preconditions for useful online healthcare discourse research

Almost any analysis of CMC necessitates, on some level, extraction of information from natural language. As described in the previous chapter, this either takes place within a qualitatively oriented workflow, where CMC data is sampled, and where one or more researchers hand-code and manually analyse the data, or a quantitatively oriented workflow, where larger amounts of data are automatically processed. The former sacrifices speed and breadth for accuracy and depth of insight; the latter, being oriented toward the development of methods that can be automatically applied to unseen data, tends to obscure the role of context in meaning-making. Because both qualitatively and quantitatively oriented approaches offer different kinds of insight into CMC, mixed-methods approaches that combine both paradigms have become steadily more popular in CMC research, just as they have in linguistics more

generally (Bolander & Locher, 2014). Discourse-analytic research, for example, has increasingly leveraged CL methods, in order to make more reliable generalisations about meanings in larger collections of text (Baker & McEnery, 2015).

Fruitful analysis of computer-mediated discourse, therefore, requires two things: first are tools and methods that can be used to transform CMC into analysable data, and then to analyse it; second is an accurate conceptualisation of how language works, so that words and wordings in a dataset can be connected to their meanings and functions in a reliable, systematic way. With this in mind, in the remainder of this chapter, I put forward SFL and CL as ways of addressing theoretical and methodological gaps in current knowledge about language use in OSGs. The major practices of CL, as well as criticism and needed improvements in these practices, are surveyed first. SFL is then introduced as a framework suitable for analysing CMC and health talk more generally. An argument is made that SFL and CL together provide benefits for analysis of OSGs, including greater accuracy, reliability and scalability, and, overall, greater explanatory power.

3.2. Corpus linguistics

Corpus linguists use *corpus* (plural: *corpora*) to refer to collections of multiple texts (McEnery & Wilson, 1996).⁶ Corpora have two main characteristics. First, they are typically large enough to permit quantitative analysis of linguistic features, and large enough that manual analysis of the entire collection is unfeasible. What qualifies as large, however, has varied considerably over time, from one million words in the **Brown Corpus** (see Francis & Kucera, 1979) to tens of billions in web corpora such as **EnTenTen** (see Jakubíček, Kilgarriff, Kovář, Rychlý, & Suchomel, 2013). Second, corpora are almost always stored digitally, so that they may be queried automatically either through the use of CL software tools, or with code (Butler, 2004).

CL approaches to language are concerned with using ‘authentic’ or ‘real’ language as data: Sinclair (1997), for example, argued that wherever possible (such as in lexicography or language pedagogy) we should present real language examples only. *Real*, in turn, generally means *uninvented* and *not elicited by researchers*, rather than

spontaneous—scripted speeches and fiction are common text types in CL investigations. As a result of the use of realised texts, CL can thus be situated within the functionalist and descriptivist traditions. That said, an increasing number of generative linguists may use corpora to investigate what Chomsky (1986) termed ‘e-language’ (Meyer, 2002).

A second universal within CL is that research is ‘always based on the evaluation of some kind of frequencies’ (Gries, 2009, p. 1226). While acknowledging that there may be some disagreement with this position, Gries demonstrates that not only are the major CL practices quantitative (e.g. keywords, collocation, clustering, etc.), but so too is thematic coding of concordance lines, or noting zero- or low-occurrence of a given feature. He also reminds us, however, that the extent to which these frequencies may inform an overall analysis is in no way fixed: researchers are free to move between quantitative and qualitative approaches as per their individual needs and interests. In CADS for example, quantitative evidence drawn from the corpus may form the sole body of evidence for an argument, or may simply provide an empirical backdrop to an otherwise theoretical discussion (see Section 3.2.6).

Much discussion has centred on whether CL is a ‘discipline’ (Tognini-Bonelli, 2001), a ‘new philosophical approach’ (Leech, 1992b), a ‘methodological innovation’ (Larsen-Freeman, 2000; Lee, 2007), a ‘methodology’ (Gries, 2009; McEnery, Xiao, & Tono, 2006) or an ‘approach’ (Lee, 2007; Stubbs, 2004).⁷ For the purposes of this thesis, CL will be considered an *approach*—‘a set of theoretical positions and beliefs about the nature of language and how we can study it’ (Lee, 2007, p. 87)—though the possible linguistic theories that can be used to understand corpus data are potentially limitless.

3.2.1. Types of corpora and corpus research

Broadly speaking, corpora are used to either make claims about language use generally, or to learn about how language is used within a particular collection of related texts.⁸ Early corpora, such as the *Survey of English Usage*, compiled in 1959 by University College London and digitised by the University of Lund (Quirk, 1960), and the one million word *Brown Corpus*, developed in the early 1960s at Brown University

(Meyer, 2002), were designed to represent English generally, with a mixture and weighting of diverse text types included. Since then, dozens of general corpora have been created, (LOB, ICE, COCA, etc.), of ever-increasing size and scope. In common to these corpora is their reliance on theoretical ideals of balanced and representative composition—ideals that are typically problematised in contemporary discourse research (Baker, 2012).

Specialised corpora, on the other hand, are those comprised of texts from a ‘specific register, genre, or variety’ (Sinclair, 2001). These entered mainstream CL in the late 1980s,⁹ chiefly for use in CDA (see Hardt-Mautner, 1995). When researchers are studying language use in a finite collection of material, they are essentially exempted from concerns of balance, representativeness and generalisability that pose challenges for investigations of general corpora (Hoey, 2005). As Baker (2010) notes, a specialised corpus comprised of the complete works of Shakespeare is uncontroversially representative of Shakespeare’s work. Specialised corpora are rarely constructed with the kinds of resources allocated for general corpora. More often, they are built by individual researchers. As such, the Web in general, and CMC in particular form convenient data-sources (see Section 3.2.5). Particularly common today are specialised corpora comprised of newspaper texts, which are most often from either a specific publication or a specific region (e.g. Caldas-Coulthard & Moon, 2010), as well as ‘lay’ online texts from discussion forums (e.g. Lukač, 2011), blogs (e.g Ptaszynski, Rzepka, Araki, & Momouchi, 2012) or article comments (e.g. Prentice, 2010).

3.2.2. Specialised corpus creation

There is no single method for creating specialised corpora, because corpora can come from a variety of sources. Even CMC corpora will generally come to the researcher in a unique kind of markup language, from which plain text must almost always be extracted. Even so, there are three major considerations noted in the literature: **corpus size**, **creation of subcorpora**, and **context retention**. These factors are outlined in the sections below, and referred to again in Chapter 4 when describing the process of building a corpus for the thesis’ case study.

Corpus size

The size of a corpus correlates with the amount of delicacy with which it can be reliably searched: corpora containing only a few thousand words will generally have too little data to locate very specific configurations of particular processes and participants; even very large corpora may not be sufficient if the researcher is interested in the grammatical behaviour and collocates of a single, infrequent word. Assuming there is no decrease in the quality of collected texts, and assuming that hardware availabilities are not an issue, it is difficult to argue with Leech's assertion that large corpus size is a good thing: 'the larger a corpus is, and the more diverse it is in terms of genres and other language varieties, the more balanced and representative it will be' (2006, p. 6). Very large corpora cannot be manually read-through, however. This means that there is always some possibility that inappropriate, poorly analysed or unanonymised text could persist, even after state-of-the-art automatic detection processes have been applied.

Creation of subcorpora

Another key design consideration is the usefulness of subcorpus structures. In contrast to bag-of-words approaches to CL, where a corpus is simply a flat (i.e. unnested) list of characters or words, the use of subcorpora makes new kinds of research questions answerable. Subcorpora may be thematic, allowing investigation of language use in different Fields of discourse; subcorpora may be for certain interactants; subcorpora may be longitudinal. Because collapsing distinctions between subcorpora during the interrogation process is trivial, even if subcorpus distinctions are not helpful, they will generally not pose a barrier to analysis. Currently, however, few corpus tools are set up to allow iteration over subcorpora (see Chapter 4), with most instead oriented toward a single bag-of-words corpus, optionally accompanied by a reference corpus or reference wordlist. This constrains the kinds of insights researchers can gain from their data, obscures heterogeneity within the dataset, and creates a reliance on reference corpora that may not be comparable to the register(s) under investigation.

Context and metadata retention

A final design consideration is the importance of storing the original versions of corpus data from which plain text was extracted. Corpus creation almost invariably involves dislocating lexicogrammar from its multimodal context—a practice for which CL has been repeatedly criticised. In response, CL practitioners made calls for context retention in corpora, arguing that as ‘the impact of discourses depends crucially on their multimodality’, and that text-only corpora ‘excludes many other elements vital to the meaning-making process’ (Hardt-Mautner, 1995, pp. 6–7.).¹⁰ Access to contextualised data on demand largely ameliorates this issue. Furthermore, context and metadata retention can lead to new kinds of research questions: the BNC, for example, contains rich demographic information that has allowed insights into socio-economic and gender-based variation in British English (Baker, 2010) that would not have been possible with a fully decontextualised corpus. An alternative, emerging approach is to create tools that can query corpus texts without ever having extracting it from its multimodal context, and which can therefore switch between monomodal and multimodal representations of linguistic data (Bateman, 2013).

3.2.3. Annotation of corpora

Once a plain text corpus has been created, a variety of pre-processing steps can be performed with the aim of improving the ability to search or count lexicogrammatical features in the text. These tasks range in complexity, and in the extent to which linguistic theory is imposed on the data. Sentence splitting and tokenisation are theoretically fairly uncontroversial (in the case of English), and are more or less solved tasks (Dridan & Oepen, 2012). Additional pre-processing measures are usually understood as annotation—that is, ‘the practice of adding interpretative linguistic information’ such as POS tags, lemma forms or full syntactic parses (discussed separately below) to a corpus (Leech, 1997, p. 2). Annotation can be carried out manually (on smaller corpora), automatically, or through a combination of both (i.e. hand correction of errors and retraining annotator models on corrected data).

Though generally an accepted practice, Sinclair (2004) voiced a notable dissenting opinion regarding annotation, stating that it may compromise the corpus and blind its interrogator to all that cannot be annotated. Sinclair's position is uncommon amongst contemporary practitioners, however (Archer, 2012). Hunston (2006) takes a less skeptical position, arguing that although annotation could potentially lead researchers to (either consciously or unconsciously) shape their research questions around what they understand to be accomplishable by interrogations of annotated data, such danger is likely outweighed by the affordances of annotation (chiefly, the retrieval of more specific data, more systematically). Moreover, processes such as tokenisation also constitute a theoretical imposition on data, but have been embraced uncritically across CL. Ultimately, the usefulness of annotation depends on research questions: development of grammar from corpus examples may find annotations harmful, as might those opting for grounded theory approaches. Annotation has obvious utility for researchers of discourse, however, who need to make links between salient components of lexicogrammar across clauses, sentences, or beyond (see below).

POS tagging, a prototypical annotation task, facilitates more nuanced searching (e.g. distinguishing between nominal and verbal occurrences of a word like *hand*), basic syntactic querying (e.g. search for sequences of DT+JJ+NN) and the ability to gauge certain stylistic features such as lexical density. It is the longest established and most common automated tagging process, and is often viewed as being 'almost indispensable' for syntactic corpus investigation in particular (Giesbrecht & Evert, 2009, p. 23). For many languages today, automatic POS tagging has also been argued to be a solved task, with taggers for dozens of languages achieving 97 per cent accuracy (Giesbrecht & Evert, 2009). That said, these measurements are for well-structured text, such as journalism, books and academic journal articles. CMC corpora, as will be discussed in Section 3.2.5, often contain non-standard language and grammar features, complicating automatic POS tagging. As Giesbrecht and Evert (2009) note, POS tagger accuracy may drop to below 92 per cent on crawled web corpora.

Parsing

Parsing is annotation of text with grammatical structure. The most common kinds of grammars in use are constituency and dependency grammars, with the latter being to some extent derivable from the former (De Marneffe, MacCartney, Manning, et al., 2006). In research environments, parsing is generally done via the command line, but access to parsers is occasionally provided by web-based CL interfaces such as **Sketch Engine**, and, less commonly, by graphical user interfaces (GUIs), such as **UAM Corpus Tool**. Currently, the reality is that many CL practitioners do not have the training necessary to operate these kinds of tools on the command line, or to write code that can traverse the annotation structures and extract the desired information. For this reason, parsing is under-utilised in CL research. CADS in particular has much to gain from working with parsed data—parses provide a means of looking for relationships between lexical items that are more complex than simple adjacency. Parse-related tasks such as coreference resolution make it possible to map pronominal referents to their original lexical form, facilitating analysis of, for instance, how social actors are construed (Feng, 2015).

One other potential reason for the slow uptake of parsing in CADS is the disparity between functional linguistic frameworks in use in discourse analytic research and the grammars with which texts can currently be automatically annotated. Constituency and dependency grammars differ from the SFG, for example, in the extent to which grammatical categories above word level are derived from semantics (Martin, 1992), and therefore, in their interest in phenomena such as Process Type and grammatical metaphor. Conversation analysts may have little use for grammatical annotations in any case, as many practitioners are more concerned with language as a window into social order than language as a system for meaning-making (Ochs, Schegloff, & Thompson, 1996). As will be discussed in Chapter 8, however, differences between constituency, dependency and systemic grammars are at times overstated by their stakeholders: while proponents are likely to stress the differences in the expressed purposes and ideological orientations of the theories, the three grammars have enough similarity (especially at the phrase level and below) to make translational research possible (Costetchi, 2013).

3.2.4. Corpus interrogation practices

CL does not prescribe particular methods for interrogating corpus data. That said, as most corpus interrogation is performed via dedicated CL tools, the most common kinds of interrogation are those that have been programmed into popular existing software. In the sections below, I provide a brief explanation of key methods of corpus interrogation, highlighting shortcomings to be addressed by the case study and tool design presented in the following chapters. Notably absent from the review are techniques such as collocation and n-gram analysis, which do not form major components of the case study analysis.

Keywording

Keywords are those that, according to a chosen statistical measure (e.g. T-score, chi-squared, log-likelihood), are particularly frequent or infrequent in a target corpus, in comparison to a reference corpus (see Rayson, 2012, for an explanation of common measures). As Rayson (2012, p. 1) explains, for CL practitioners, the term *keywords* is generally accepted to denote a set of words that ‘statistically [...] characteriz[e] a document, text, or corpus’. Keywords have formed an important part of most discourse-oriented CL,¹¹ with two main strategies currently in use. In the first, keywords guide collocational analysis and concordancing (as in Harvey et al., 2007). In the second, keywords are thematically categorised in order to elucidate macro-level ‘meaningful clusters of content’ within the corpus (Harvey, 2012; Williams & Weninger, 2013, p. 357). Notably, the notion of keywording has been problematised by Baker (2004), who explains that keywording of plain text without regard to grammatical distinctions may lead to conflation of different word senses, or other functional differences that may be critical to the meaning of a text.

Lexicogrammatical querying

Lexicogrammatical querying involves searches of annotation structures. These kinds of queries exist on a cline from broad to delicate. Broader queries may target a feature or combination of features across all tokens in a corpus (e.g. *What are the most/*

least common nominal group structures?—see Teich, Degaetano-Ortlieb, Fankhauser, Kermes, & Lapshinova-Koltunski, 2015). More delicate queries target a particular word, lemma, or set thereof (*What are the most/least common nominal group structures when the head is ‘risk’?*—see Zinn & McDonald, 2015). Discourse-analytic investigation may progress from the first to the second, finding frequent/salient tokens in general queries and then exploring their grammatical behaviour (Baker, 2013). Though powerful, this approach relies on the ability to write queries to traverse annotation structures in arbitrary ways. This is a method outside of the scope of most existing CL tools; it is therefore a core feature of the tool developed for the case study of this thesis.

Concordancing

Concordancing involves displaying all instances of a search term in a vertical line, with a specified number of words or characters on either side of each token, as well as potential metadata regarding subcorpus name, location of token or speaker ID (see Figure 7.6 for an example). Concordancing is one of the more common practices in corpus-assisted language learning (Baldry, 2008), as well as in discourse-analytic CL, where researchers can use the context surrounding text to identify key themes or broader contextual information (Hardt-Mautner, 1995). Concordancing, notably, has become the de facto symbol of CL interrogation, with software for performing interrogations often referred to as ‘concordancers’. A pragmatic perspective on the subject is provided by Baldry (2008): he suggests that rather than being the result of the virtues of the approach itself, the popularity of token-based concordancing is, for the most part, the result of researchers’ easy access to concordancing software, its low-learning curve, and the fact that tagged corpora are not required. He argues that other possible types of concordancing have been ‘eclipsed’ by the lemma-based variety, and that standard practice deserves scrutiny and technological improvements—namely, multimodal concordancing and concordancing of non-linguistic (contextual) phenomena within corpora.

3.2.5. CL and the World Wide Web

As the Web grows in size and popularity, researchers have begun to apply CL methods to data from online sources (including, but not limited to CMC). Due to its size and constantly renewing nature, the Web proves especially suitable for researchers interested in emergent or rare language features (Fletcher, 2012; Koteyko, 2010). Though the notion of an egalitarian internet has since faced scrutiny (see boyd & Crawford, 2012; Herring, 1996a), there is little doubt that voices marginalised by traditional channels of media and publication can be more easily found online (Chiluwa, 2012; Ryder & Wilson, 1996). While discourse-analytic accounts of online communities are now commonplace, such studies have tended to be from a CMC, rather than CADS perspective. Though Mautner's (2005) request for more web-based CADS seems to have gained traction in the past five years (see Table 3.1 below for some examples), most of the diverse range of potential data-sources for corpora have been ignored in favour of more 'traditional' specialised corpora, such as government documents and newspaper articles (see Section 12), with web-corpora being more commonly produced for the purposes of lexicography, minority language research, pedagogy or NLP. As Mautner (2005, p. 810) notes, this is surprising, as one would expect discourse-oriented researchers to 'seize every opportunity to look at discourse in a medium which is now such a key space for enacting social practice, and for reflecting and shaping social processes and problems'.

An additional motivation for using the Web as a data-source for CL is pragmatism: by using the Web, vast amounts of constantly updating, pre-digitised language from countless registers can be quickly and automatically compiled, saving time and money (Baroni, Bernardini, Ferraresi, & Zanchetta, 2009). Even so, some have voiced skepticism regarding web corpora. Leech (2006), for example, voices a concern that the practicality of web corpora may cause neglect of more traditional kinds of corpora. He also notes that what is found online may not be representative of communication offline. This is certainly a reasonable point—even familiar modes of CMC with obvious offline antecedents will likely differ in some respects from the offline counterpart in some way. That said, debates surround the representativeness of reference corpora more generally (Baker, 2012)—the exclusive use of web corpora

to analyse ‘standard language use’ is likely no less problematic than the exclusive use of offline texts, as both strategies ignore large parts of the current landscape of human communication. It also needs to be pointed out that CL itself has a long history of favouring pragmatism: the focus on written text over spoken text (with even contemporary general corpora such as the BNC containing 90 per cent written text due to the expense associated with transcription—see Burnard, 2016; Leech, 1992a) stands as case in point.

3.2.6. Corpora and discourse analysis

With key practices and debates in mainstream CL defined, the focus of the chapter now shifts toward the burgeoning body of research integrating corpus methods into discourse analytic research. This synergy of approaches (with a number of additions, and the use of purpose-built tools) forms the overarching methodology of the case study introduced in the next chapter.

In the last decade, the ‘methodological synergy’ of CL and discourse analysis as CADS¹² has generated new ways of interpreting increasingly available sets of structured, digital texts. Such approaches have the promise of being able to demonstrate that texts or fragments of texts chosen for qualitative analysis are recurrently, frequently or systematically instantiated within, rather than being simply ‘cherry picked’ from, the dataset under investigation (see Baker et al., 2008). Moreover, corpus interrogation and analysis allows the interpretation of datasets too large to be manually analysed (or even read through) by individual or teams of researchers. Assuming data is uniform in structure, annotation/parsing (if performed) are accurate and search patterns are well-defined, corpus methods also allow certainty that all instances of a given lexeme within a dataset have been counted, and can be located for qualitative interpretation.

To understand the contemporary state of CADS, it is important to first review earlier developments in the field. By situating the area within its historical context, it becomes easier to see that current issues, reviewed below, are in large part a result of tools, methods and epistemological outlooks inherited from earlier studies. Argued in this section is the idea that many of these practices remain unquestioned within

the area, and perhaps deserve re-evaluation in light of more recent technological developments.

Emergence of the field

An influential early publication featuring what would later become CADS is Fox, Hoey, and Sinclair (1993). There, Caldas-Coulthard (1993) conducted a CDA of a small corpus of British newspaper articles, highlighting the under-representation of females as Sayers by collocation analysis of verbal processes. The subject of the attitudinally neutral verb *say* was eight times as likely to be male than female. Furthermore, men were found to *shout* or *groan*, while women and children *scream*, *yell*, *nag* and *complain*, Caldas-Coulthard's corpus interrogation also revealed that women in the news tend to be characterised in terms of their relationship with a male (*his grandmother, Mrs Barbara Wilkinson; Hillary, Mr Clinton's politically attuned wife*), with no instances of the inverse uncovered in the corpus.

In the same book, Fox (1993) outlined the proceedings of a British court case in which CL techniques were used to dispute the authorship of a document purported to be an unaltered witness statement. Corpus-based analysis of the text's repeated use of the construction *I then* revealed that it patterned strongly with 'policespeak', but was almost absent from 'normal' conversation in the COBUILD corpus. Fox's conflation of the COBUILD corpus with 'normalspeak' brings into focus questions concerning whether or not a corpus can ever be trustworthy, balanced and/or large enough to be representative of language outside of the corpus (in this case, the vaguely defined idea of 'normalspeak'). Certainly, in the case of Fox's study, the ideal reference corpus in this case would have been comprised of authentic statements from witnesses with similar demographic information to the subject in question. But is the creation of such a corpus practical? If not, which more accessible texts can we justifiably substitute for them? And even if we had access to such documents, could we ascertain their authenticity? It is of course possible that they too were manipulated by police. Finally, even assuming we could, how many such documents would we need to permit sensible generalisations concerning *I then* and *then I?*

Hardt-Mautner (1995) established a preliminary framework for integrating CL and CDA. Drawing on the previous two studies, as well as her own analysis of representations of the EU in a specialised corpus of British newspapers, she posited that by using CDA, CL can overcome the long-held criticism that its method inherently obscures utterances from their original context (see Section 3.2.7). By the same token, CDA benefits from CL: through quantitative analyses, specific foci of qualitative CDA can be statistically and empirically justified, silencing the commonly voiced concerns regarding cherry picking in critical linguistic work (Baker, 2012). Under this framework, however, corpus methods and findings were considered ancillary to those of CDA: so long as quantitative findings are used simply as a means of guiding researchers toward areas of analysis, she argues, CDA's 'commitment to analysing coherent discourse' can remain intact (1995, p. 3).

Contemporary practices

More recent research has attempted to grapple with some of these shortcomings of earlier work. Baker, Gabrielatos, and McEnery (2012) analysed collocates for *Muslim* in a specialised corpus of British newspapers from 1998–2009, determining that Muslims were often homogenised through phrases like *Muslim world* and *Muslim community*, and treated as oppositional to *the West*. Here, no reference corpus was used, avoiding issues of representativeness of reference corpora encountered by Fox (1993): as the only language under investigation is the content of the corpus, generalisations concerning specific British news sources within specific time-frames can far more soundly be made (Hoey, 2005). Notably, despite the large size of their dataset (143 million words), the corpus was for the most part treated as unstructured: aside from a brief description in longitudinal changes in the frequency of *Muslim world* and *Muslim community*, changes in the discursive construal of Islam and Muslims over the sampling period were not identified.

Partington (2011) looked at humour in a specialised corpus of White House transcripts by concordancing the word *laughter*, which denoted instances of laughter according to the transcribers' annotation scheme. He annotated the plain text with his understanding of the cause of the laughter, then used concordancing to qualitatively

analyse instances of these certain ‘types’ of humour. Though avoiding earlier studies’ problems of generalisability, this methodology foregrounds an element of CL’s annotation debate. Interrogating corpora tagged for POS, semantic role or theme means implicitly trusting the validity of the theory underlying the tagging algorithm. Accuracy of the tagging often goes unchecked. In the case of Partington’s study, though the transcripts were cross-referenced with audio-visual material at some points, it remains unknown the extent to which laughter was correctly identified and uniformly transcribed. Another issue with such methods is that they scale poorly: the application of Partington’s methods to other datasets is limited by the time and resource constraints associated with identifying and annotating laughter.

Common practices in CADS

Year	Authors	(C)DA	Topic	Site
2008	Baker et al	CDA	Refugees to UK	British newspapers
2010	Caldas-Coulthard	CDA	Women’s bodies	British newspapers
2010	Koteyko	DA	Carbon	RSS feeds
2010	Prentice	DA	Scottish independence	Discussion forums
2011	Bachmann	(C)DA	Homosexuality	UK parliament
2011	Lukač	DA	Pro-anorexia	Blogs
2011	Partington	DA	Humour	White House press releases
2011	Salama	CDA	Wahhabi-Saudi Islam	Non-fiction books
2012	Baker et al	CDA	Muslims	British press
2012	Bevitori	DA	Greenness	Newspapers
2012	Bianchi	DA	Chocolate and wine	Web-crawl
2012	Chilعوا	CDA	Biafra	Blogs and forums
2012	Harvey	DA	Depression	emails
2012	Harvey	DA	Self-harm	emails
2012	Hsiao	DA	Restaurant reviews	Newspapers
2012	Jaworska & Krishnamurthy	CDA	Feminism	British/German media
2012	Mulderrig	CDA	Education policy	Policy documents
2013	Koteyko, Jaspal & Nerlich	DA	Climate change	Online reader comments
2014	Koteyko	DA	Russia	Media & Political writing
2015	Schroeter & Storjohann	SFL	Financial crisis	British news
2016	Bartley & Benitez-Castro.	Appraisal	Homosexuality	Irish news
2016	Jaworska	CDA	Sport	British/South African news articles
2016	Salahshour	CDA	Migrants	New Zealand news

Table 3.1: Recent papers in corpus assisted discourse studies

As the research area has matured, common practices concerning analytical methodologies, data-sources, qualitative methodology and themes have begun to emerge. As can be seen in Table 3.1, early studies such as Caldas-Coulthard’s (1993) CDA of British newspaper articles have set enduring standards for CADS: news-texts and government documents form the primary data-sources of most of the recent papers

found during the literature review. That said, the continuing popularity of such sources is also likely a result of their being well-structured, edited and archived, which renders them easily transformed into corpora. Moreover, they are widely consumed (and therefore influential). Similarly, given CDA's interest in language and power, the analysis of discourses present in the language of powerful institutions is ideologically in line with CDA. Even so, as Mautner (2005) notes, research in the area appears to have lagged behind changes in technology and communication practices: the surging popularity of the web, as case in point, has given rise to many new potential sources of corpus data (such as blogs, chat transcripts, forum posts, etc.) that are yet to be given significant treatment within CADS.

CADS and computer-mediated communication

For a number of reasons, CADS practitioners have begun to look toward CMC as a potential data-source. First, the Web is an undeniably practical source for natural language: online data are easily accessed and stored, with embedded metadata containing speaker information, timestamps, number of views, and so on. Second, as Mautner (2005) notes, the Web also provides channels through which the voices of under-represented or marginal groups may be accessed, facilitating critical discourse studies of language and power. A third influence is a broader shift within the social sciences from a deficit model of CMC (whereby online interactions were treated as impoverished versions of face-to-face equivalents) toward an understanding of CMC as a rich resource through which speakers are able to communicate a plethora of non-verbal cues (Dresner & Herring, 2010; Schandorf, 2013). Finally, the ever-increasing presence of CMC in daily life has created an increased impetus and desire to account for its effects upon language production.

Currently, CMC corpora most often fall into one of two categories. The first is very large general or reference corpora (e.g. Minocha, Hyderabad, Reddy, & Kilgarriff, 2013; Fletcher, 2012; Baroni et al., 2009; Kilgarriff & Renau, 2013), created either through web-crawling, downloading and processing (see Baroni et al., 2009), or (more problematically) through using search engine result counts to gauge the popularity of a word or phrase (see Kilgarriff, 2007). The second are specific purpose cor-

pora created through passing in lists of seed words (generally keywords for a given topic) to a search engine application programming interface (API) and downloading and processing the matching results (see Baroni, Kilgarriff, Pomikálek, & Rychlý, 2006). Both types of corpora, though useful as reference corpora or for lexicographic purposes, are of limited usefulness to researchers interested in discourse. Corpora utilising general web crawling or search engine queries generally contain a broad array of text-types that are treated as a homogeneous set. Determining the influence of specific text-types on the lexicogrammar in the corpus, or contrasting sets of texts in meaningful ways, is a very difficult task. Moreover, search-engine ranking algorithms dictate what will be included in the corpus, leading to corpora skewed toward content from more popular pages (Wikipedia articles, company websites, etc.) (Kilgarriff, 2013).

Many CADS studies using CMC avoid problems associated with general analysis by creating specialised corpora. Danescu-Niculescu-Mizil et al. (2013) show how the length of membership in an online beer enthusiast community affected users' willingness to adopt lexical change. Longer-term members of the community resisted an emerging terminological shift (the replacement of *aroma* with *smell*), while users appearing after the lexical innovation were content to embrace it. Koteyko et al. (2013) have used a CMC corpus to elucidate the discursive framing of a current event: their analysis of a corpus of readers' comments on online news articles revealed how the legitimacy of climate change science is challenged through the use of grammatical systems for intensification and mitigation. Courtney Walton and Rice (2013) investigate identity performance in a corpus of tweets, noting that even in very constrained writing environments, female gender as a normative social construct may be performed through tweets indexing positive emotions and through disclosure of backstage personae. Also looking at corpora of tweets, Zappavigna (2011, 2012, 2013) has used SFL to theorise the ways in which role relationships are formed and negotiated through both the indexing of semantic fields as hashtags and through MOOD choices as per SFG (see Halliday & Matthiessen, 2004).

CADS, healthcare and the internet

CL, as a set of practices for building and interrogating large bodies of natural language, is a natural candidate for the analysis of large online communities. First, as the amount of natural language in such communities may be too large to be analysed by individual researchers, CL provides a means of systematically analysing huge bodies of text. Second, the way the language is stored within the HTML content of webpages (i.e. with metadata including speaker names, timestamps, etc.) makes the transformation of the language into structured corpora a largely automatic process. Despite this suitability, as discussed in the previous chapter, OSG research has rarely engaged with practices from CL.

One of the few examples corpus approaches to online health discourse is the *TeenHealthFreak Corpus* of teenagers' anonymous emails to an online doctor, which has been used to discursively analyse representations of self-harm, depression and sexual health (in Harvey, 2012; Harvey & Brown, 2012; Harvey et al., 2007, respectively). The authors use common CL practices (keywording, collocation analysis and thematic categorisation of concordance lines) to show that adolescents are centrally concerned with their perceived deviations from normative physicalities, bodily functions and states of mind. The adolescents' emails foreground a sense of worry over their perceived abnormality, discursively constructing normalcy as desirable and any deviations from normalcy as potential health problems. While implications for healthcare are proposed, Harvey balances the findings with an important caveat: corpus interrogation, he explains, cannot uncover anything about the phenomenology of depression that could not be discovered through other means. Instead, he argues, CL is primarily useful as a means of elucidating sufferers' understandings of their conditions through the linguistic features they employ to describe them 2012. Researchers within medical NLP may be inclined to agree with this position, while nonetheless stressing that automated linguistic analysis can provide useful, cost-effective insights into medical discourse (MacLean et al., 2015).

Harvey's (2012) study of the same corpus focussed on the ways in which the adolescents attributed depression to themselves: *being depressed* or *having depression* are the most common, but the teens also note *feeling depressed* or *suffering from*

depression. Harvey explains the value of detailed analysis of the differences between these forms:

These lexicogrammatical options and distinctions are important, since preference for a certain form encodes a particular version of events which, in turn, will have consequences for how experiences are constructed and understood. [...] They] are indications of how individuals situate themselves in relation to their illness experiences, and as such are aspects of illness discourse that, as well as revealing a self-labelling position, also provide a potential illness explanation (2012, p. 361).

This analysis highlights the fact that delicate lexicogrammatical choices make discursively significant distinctions between meanings. Even very common relational processes such as *be* and *have*—words that may even be excluded from analysis by the use of stopword lists—play important roles in the way people construe their experience as healthcare consumers.

3.2.7. Key debates in corpus-based approaches to discourse

CL and CADS seem well-suited as possible approaches to an investigation of language use in an OSG: quantitative methods make it possible to analyse communities as a whole, rather than small samples of communications that took place therein, while the grounding of CADS within a discourse-analytic tradition provides sensitivity to how the text stored in corpora can be related to grammatical systems, which in turn can be related to meaning, and then to context. Still, CL is not without shortcomings and contradictions, many of which have long been noted by scholars from other traditions within linguistics such as Widdowson (1991, 2000; 2008).

In the first part of this section, I identify key shortcomings in current CL practices, focussing where possible on work oriented toward discourse. I then describe key criticisms of CL approaches, developing an argument that many key criticisms reflect current shortcomings in available data and tools more than they do any inherent limitations of the CL approach itself.

Lack of available datasets

Corpus and discourse linguistics have no explicit preference for texts of any particular genre or register. Nonetheless, some text types are far more commonly analysed

than others. Due to the fact that news media texts are often well-archived and freely available via databases such as *LexisNexis* and *ProQuest*, newspaper content is perhaps the most common kind of data investigated within CADS (e.g. Baker et al., 2008; Caldas-Coulthard & Moon, 2010; Partington, 2010; Baker et al., 2012; Hsiao, 2012; Jaworska & Krishnamurthy, 2012; Zinn & McDonald, 2015). Also common have been transcripts from government communication (e.g. Bachmann, 2011; Mulderrig, 2012; Partington, 2008a, 2008b, 2011, 2013). Non-fiction articles and books have also been analysed (e.g. De Beaugrande, 2001; Salama, 2011). Emerging is the interest in corpora of CMC by the general public, including studies of emails (e.g. Harvey et al., 2007; Harvey, 2012), blogs (Lukač, 2011; Chiluwa, 2012) and reader comments to online news (Koteyko et al., 2013).

The slow turn toward CMC corpora is disappointing. Despite the registerial diversity of CMC texts, and the access CMC provides to speakers who may generally be marginalised in society, Western print journalism remains the de facto source for corpus data. Currently, the main stumbling block seems to be the difficulty of building CMC corpora from scratch, as many tasks involved in building corpora from CMC (crawling/spidering the web, extracting natural language from HTML/XML, normalising, structuring and parsing texts, etc.) are largely tasks for those with familiarity with command-line based processing. Review of the literature uncovered very little engagement with these kinds of techniques, even in large-scale CADS projects. Easy-to-use GUIs capable of performing these tasks are generally unavailable. As discourse-analytic researchers typically also require training in qualitative methods, the skill-set needed to build and investigate CMC corpora for discursive features is often prohibitively large. This issue is amplified, of course, in cases where the researcher requires the ability to engage with multimodal content, rather than simple plain text, or where the researcher requires annotation of semantic/registerial features that cannot reliably be performed by a machine.

Under-utilised and unavailable digital tools and resources

The second major issue is that a number of available and potentially useful technologies for working with natural language are yet to be implemented within CADS. As

with the previous issue, a lack of researcher training in computational methods is perhaps the overarching cause: many advanced tools must be compiled and operated via the command-line, with no GUI available. Critically, automatic tagging and parsing of texts are often tasks that fall within this domain, leading to a dearth of CL/CADS research involving targeted querying of lexicogrammatical patterns of texts. As will be demonstrated by the case study, and discussed later in the thesis, such corpus interrogation methods radically expand the ability to reliably identify functionally significant grammatical patterns in text, and to distinguish between interpersonal, experiential and textual metafunctions of language.

Another issue is that few CADS have made use of popular available resources for sharing data, findings and/or developed code. Many of the corpora listed in Section 3.2.7 appear to have been used only a handful of times at most, and very rarely shared, despite myriad potential uses for each. In terms of findings, free services like *Figshare* provide places where tables, figures and charts too large for publication in hard-copy can be easily stored and shared. The code developed to automate corpus building, annotation or interrogation can also be easily added to public online repositories such as *Github*. Sharing such resources increases research reproducibility, saves time, and substantially lessens the need to develop new corpora.

Simplified common practices

Perhaps the most serious shortcoming within CADS at present is the (often uncritical) use of a simplified set of practices during corpus interrogation, with many researchers relying on techniques that have faced longstanding criticism within applied linguistics more generally. One example is the process of *keywording*, which relies inherently on the composition of the reference corpus. This reference corpus tends to contain a diverse array of text types that are often inappropriate: when analysing a corpus of political speeches, comparisons of this corpus with a dataset partially comprised of instruction manuals and recipes seems absurd. At the same time, keywording as performed by most current tools involves automatic exclusion of a list of function words (*stopwords*), which may dominate keyword lists simply due to their very high frequency in texts of any type. Indiscriminate and arbitrary

removal of such words before having understood what they do in text is poor practice, however. The base forms and inflections of *be* and *have* are often excluded by such lists, for example, due to the fact that these items are often auxiliaries within verbal groups, with little ideational content. As shown by Harvey (2012), however, *be* and *have* do important work in construing the relationship between self and illness. The appearance of stopwords as key may also be a useful warning sign of differences in tokenisation algorithms that segmented the target and reference corpora.

Similarly, the uncritical use of *collocation tests* (counting the frequencies with which certain tokens co-occur within a given window) in CADS may be problematic. Though two tokens may indeed collocate absent a specific kind of grammatical dependency relationship, and though this is certainly worth investigation, collocation in CADS is often used to simply source the most common adverbs modifying a verb, or the most common adjectives modifying a noun. Such methods are ultimately imprecise. As an example, *Men are from Mars, and women are from Venus* would contribute to an understanding of *Mars* and *women* as collocates, regardless of the fact that the opposite meaning is intended. Again, this may be the result of a lack of researcher training: though the building of corpora parsed for grammatical structure can often provide far more accurate means of interrogation, the necessity of some command-line driven processing and scripting is prohibitive to many social scientists who seek to use corpora in their discourse-focussed research.

Addressing criticisms of the CL approach

Despite its successes, CL has faced a great deal of criticism within linguistics more generally. Chomsky's well-known argument for a focus on competence rather than performance (1965) is in part addressed to the then-nascent field of CL, under the auspices of which the development of the *Brown Corpus* had taken place a few years earlier. Often, such sentiments can be ascribed to reactions against CL's uneasy position within the broader landscape of linguistic research (its initial conflict with a better-established generativist tradition; its methodological differences from other functional linguistic approaches; the unfamiliar research areas with which it is associated, etc.—see Baker, 2010). Indeed, much criticism has been levelled at CL by

those who are only vaguely aware of its practices, or are informed by CL stereotypes that have long since been ameliorated by both theoretical and methodological developments within the field (Baker, 2010). That said, coherent and justifiable criticism has also been made. In the following sections, I summarise those made by Virtanen (2009) and Widdowson (1991, 2000), Widdowson (2008), as well as a criticism of *Big Data* approaches to social science research more generally noted by boyd and Crawford (2012).

CL as context-obscuring

As mentioned throughout this section, there are concerns from within and outside CL that the decontextualisation of language necessary to create corpora may obscure the realities of discourse. Virtanen (2009) argues that corpus and discourse approaches are opposed in this respect: a major goal of discourse linguistics, she argues, is the analysis of texts as social processes that make meaning in context. Accordingly, for discourse analysis, the preservation of context and *in situ* analysis of text is not negotiable:

The obstacle in the relationship between CL and discourse linguistics is the issue of text-context reflexivity, which does not readily lend itself to static analyses of decontextualized data in the form of the linguistic output of situated discourse events which have been recontextualized as a corpus (2009, p. 60).

This issue of context sensitivity, she explains, is the main problem underlying the synthesis of corpus and discourse linguistics: ‘it is only through due attention to discourse as process and social action that investigations succeed in truly taking into account the bidirectional relation between actual texts and pieces of discourse’ (2009, p. 62).

A similar criticism is voiced by Widdowson (2000, p. 6), who contends that CL can only ‘provide us with the description of text, not discourse’. Ultimately, he argues, ‘[a]lthough textual findings may well alert us to possible discourse significance and send us back to their contextual source, such significance cannot be read off from the data’ (2000, p. 9.). Like Virtanen (2009) and Mautner (2005), Widdowson notes that the transformation (and subsequent decontextualisation) of text into corpora obscures access to unadulterated discourse. While this sentiment is indeed valid for

many (especially early) corpora, which stripped language of any available context in the name of machine-readability, recent technological developments highlight the possibility of linking multimodal representations of documents to plain or annotated text suitable for corpus linguistic analysis using databasing tools such as SQL, or flexible markup formats such as XML (Hiippala, 2016). Moreover, from a theoretical perspective, linguistic traditions such as SFL have demonstrated that much of information stored in the contextual source of a text is encoded within its linguistic properties (see Section 3.3.4).

Conflation of attested and possible utterances

In the terminology of Hymes (1972), CL deals with the textually attested, but neither the encoded possible, nor the contextually appropriate. Widdowson correctly points out that CL is therefore a partial account of language. He then characterises CL practitioners as conflating the attested with the possible, arguing that corpus linguists treat phenomena not found in the corpus as ‘not English—not real English at any rate’ (1991, p. 14). This is perhaps the weakest of his arguments, and a gross misrepresentation of what corpus linguists understand the utility of the approach to be: CL practitioners are well-aware of the fact that what is represented in the corpus is limited by its size and design features, and that a corpus can never disprove the existence of a feature (Baker, 2010). As Stubbs explains, CL is not used to posit that a feature does not or cannot exist; rather, it ‘is concerned with a much deeper notion: what frequently and typically occurs’ (Stubbs, 2001, p. 151).

CL as only accessing attested language use

Virtanen (2009) has argued that CL and discourse analysis are useful for investigating two fundamentally different things: discourse linguistics is centred on investigating the process that connects instantiated texts to meaningful goals and functions; corpora, on the other hand, contain the outcome of the process of text production—the realisation of only one of a range of possible choices (Martin, 1992). In Virtanen’s words:

Corpora are essentially static, consisting of records of spoken or written text that discourse linguists explore in the hope of being able to reconstruct the processes

through which these products were shaped to serve particular communicative goals and to function as situated social action for interlocutors, readers and writers (2009, p. 50).

This is a sentiment shared by Widdowson (2000), who stresses that a key limitation of CL generally is its focus on ‘third person’ attested language, rather than the appropriate or the possible. ‘Since what is revealed [by CL] is contrary to intuition’, he explains, ‘then it cannot represent the reality of first person awareness’, and instead can ‘only analyse the textual traces of the processes whereby meaning is achieved’ (2000, p. 6).

While Widdowson’s criticism is a healthy reminder of where CL has explanatory power and where it does not, this does nothing to diminish its usefulness as an approach for analysing realised lexicogrammatical choices, nor to distinguish it from the vast majority of functional linguistic approaches that also do not access speakers’ intuitions about language. Moreover, a mismatch between speaker intuition and a finding of research does nothing to invalidate the usefulness of the research method. In many computational linguistic domains, for instance, texts are reduced to vast numerical feature vectors, which model probability distributions for every word’s likelihood of following every other. Though it is unlikely that such vector spaces reflect speaker intuition, this does not preclude the method from being useful in a wide array of tasks, such as word similarity, parsing, natural language generation and machine translation. In fact, in computational domains, machine learning methods with little resemblance to speaker intuitions about language tend to outperform the traditional rule-based approaches grounded in human intuition.

CL falsely claims objectivity

Wider critiques of ‘Big Data’ approaches to social-scientific investigation—of which CL is no doubt part—have pointed out that quantification is often courted by social scientists as a means of legitimating the research area as scientific and giving the appearance of objectivity (boyd & Crawford, 2012; Latour & Canea, 2010). This is said to be at odds with the reality of motivated, situated researchers (Gitelman, 2013). This is an important point to keep in mind, but does not invalidate the CL approach as a whole. Indeed, the notion of objective quantification and analysis may be fiction,

and that number-crunching on unprecedented scales does not in and of itself provide *insights* that cannot be garnered through other means. That said, quantitative and/or computational methods do lend themselves to *applications* that may otherwise not be possible, including automation, or use of results in non-research settings. The task then becomes recognising the difference between objectivity and the mere appearance of it.

3.2.8. Summary: corpus linguistics as an approach

In this section, I have reviewed key practices of CL, highlighting in particular applications of corpus methods for discourse research. Because there has been notable debate about the appropriateness of mixing corpus and discourse linguistics, key criticisms of the fields were described and addressed.

A noted shortcoming within CL was the lack of take-up of tools for annotation and parsing of corpus text. A consequence of this, I argued, is that CADS practitioners have been limited to simpler corpus linguistic methods, such as keywording, collocation and concordancing, which can be performed in readily available graphical tools. It is through grammatical annotation, however, that corpus researchers can operationalise powerful functional grammatical notions that can aid in the process of linking realised language to meaning-making and context.

In the next section, I review SFL and its associated SFG as the theory of language and grammar most suitable for corpus linguistic analysis of an OSG.

3.3. Systemic functional linguistics

Systemic functional linguistics (SFL) is a functional-semantic theory of language and context, used throughout this thesis as a theoretical framework for conceptualising and understanding *texts*. The central focus of the theory is the SFG, which guides the investigation presented in Chapters 5–7, and which is used to link lexico-grammatical choices found in Forum contributions to discourse-semantic meanings in Chapter 8.

As a theory of language, SFL can be understood as centrally concerned with three axes:

1. A **cline of instantiation**, with language as a system at one end, and individual texts (i.e. realisations) of the system at the other
2. A **hierarchy of stratification**, where contexts have probabilistic effects on the meaning, content and expression in a text
3. An **array of metafunctions** that language is structured to accomplish (role-relationship negotiation, the construal of experience, and self-organisation of text into coherent units)

The theory has evolved from the early work of Michael Halliday (e.g. 1966a, 1978), with substantial further contributions from e.g., Hasan (1985), Martin (e.g. 1992) and Matthiessen (e.g. 1995). That said, much of the theory's conceptualisation of language can be traced to developments in linguistics in the first half of the 20th century. In the sections below, I elaborate on the three axes identified above, situating each within a brief historical context.

3.3.1. Cline of instantiation

The cline of instantiation is the space between the linguistic system and an actual sample of real-language use. The cline invokes the Saussurean distinction between language and speech (1916), but rather than treating each as conceptually different, positions each as occupying a pole on an axis of *instantiation*. SFL also expands upon the Saussurean position, by considering language not as a study of signs, but of *sign-systems*—‘not sets of individual things, but rather networks of relationships’ (Halliday & Hasan, 1989, p. 4). In between the two poles are lexicogrammatical systems of increasing delicacy, which extend from broad grammatical choices (between indicative and non-indicative Mood, for example) toward and up to lexis. In this way, SFL treats lexis and grammar as two ends of the same stratum. SFL presents this axis in the form of *system networks*, which represent speakers' possible choices and their constraints on more delicate selections. Doing text analysis in SFL therefore chiefly involves relating text (an instance) to the overall meaning potential of the system.

3.3.2. Hierarchical structure of language and context

In SFL, language and context are organised hierarchically as *strata*, with every stratum realised through the stratum below. Following the anthropological work of Malinowski, context is divided into two strata: a *cultural dimension*, which shapes all interactions taking place within the culture, and a *situational dimension*, which concerns the specific environment in which a given text is produced (Halliday & Hasan, 1989, p. 6). Following Hjelmslev (e.g. 1946), language is also stratified into register, semantics, lexicogrammar and phonology/graphology (depending on whether or not the text is spoken/written).

Within the lexicogrammatical stratum, we can observe a more detailed kind of hierarchical order, the *rank scale* (Table 3.2). The nature of the hierarchy is different, however: in the stratification of language and context, each stratum *realises* the stratum above; in the rank scale, each level is comprised by one or more instances of the rank(s) below.

Clause complex
Clause
Group/phrase
Word
Morpheme

Table 3.2: Rank Scale (Halliday, 1966)

Incongruent realisation, or *grammatical metaphor*, expands the meaning potential of language (Heyvaert, 2003): through this device, realisations of one rank through another (for example) may be chosen by speakers to satisfy experiential, interpersonal or textual goals. These realisations are typically *agnate* (that is, closely related) in terms of experiential semantics (Matthiessen, Teruya, & Lam, 2010).

3.3.3. Metafunctions of language

The final major dimension of SFL is the division between metafunctions of language. In SFL, language users are seen as simultaneously attending to three metafunctions, each of which is responsible for making a different kind of meaning. Each of the

metafunctions is realised concurrently by distinct lexicogrammatical systems. The *interpersonal metafunction*, realised by the MOOD and MODALITY systems, conveys and negotiates role-relationships between interlocutors. The *experiential metafunction*,¹³ realised by the system of TRANSITIVITY, is used by speakers to construe doings and happenings in the real world, inner states of consciousness, and relationships between things and events. The *textual metafunction*,¹⁴ realised by the system of THEME creates coherence within and between texts (Eggins, 2004).

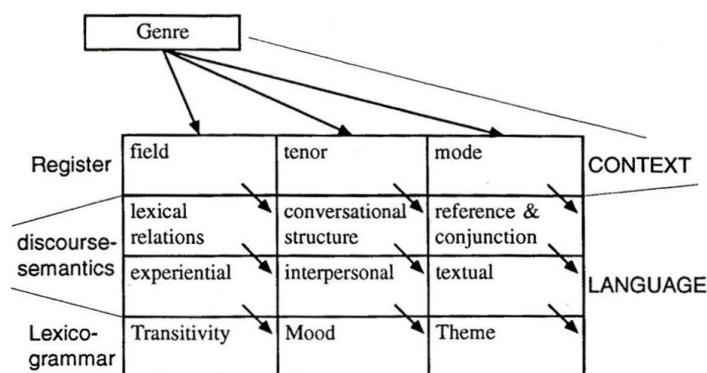


Figure 3.1: Strata and metafunctions of language (adapted from Eggins, 2004)

3.3.4. Affordances of SFL for researching an OSG

SFL has been applied to a diverse array of research areas including first and second language acquisition (Halliday, 1993a; Hasan & Perrett, 1994), language pedagogy (Halliday, 1993b), historical linguistics (Cummings, 2010; Martin & Wodak, 2003) and (critical) discourse analysis (Hunston, 2013, 4–5; Le & Wang, 2009; Martin & Rose, 2003). For the latter of these, SFL is particularly useful, and has for this reason become one of the most popular grammars and theories of language for discursive research: as Halliday explains, at the most general level, SFL is used ‘to understand the quality of texts: why a text means what it does, and why it is valued as it is’ (2004, p. xxx). In fact, from a systemic-functional perspective, the use of a grammar is a prerequisite for all discourse analytic research. In Halliday’s words,

it is sometimes assumed that discourse analysis, or ‘text linguistics’ can be carried on without grammar—or even that it is somehow an alternative to grammar. But this is an illusion. A discourse analysis that is not based on grammar is not an analysis at all, but simply a running commentary on a text (2004, p. xvii)).

As a result of this conviction, the organisation and metalanguage of SFL has in part been designed with text analysis in mind.

The overall utility of SFL for analysis of an OSG may be divided into three main factors: the treatment of language as constitutive of text, the notion of language as a meaning-making resource, and the division of interpersonal and experiential metafunctions within a grammar. These three factors are outlined below.

Language as constitutive of context

Though context is an increasingly central concern within many branches of linguistics, SFL is notable for the extent to which its theory of context has been articulated and empirically applied (Widdowson, 2008). Halliday explains that in SFL, context and text are in fact seen as ‘aspects of the same process’, and that any sample of language use thus in fact ‘goes beyond what is said and written: it includes [...] the total environment in which a text unfolds’ (1989, p. 5). Therefore, SFL treats language as constitutive of, rather than simply bound to, its context. As context can often be accurately deduced from text alone, and as context can be used to predict appropriate kinds of texts, some SFL theorists contend that *context is in text* (e.g. Eggins, 2004, p. 7). Context and language thus together construct both a reality and roles for people within it (Veel, 1997).

This notion of language as simultaneously constructing and responding to the demands of contexts of situation has often formed an implicit theoretical assumption in online community and OSG research: many works reviewed in the previous chapter tacitly take the perspective that language *must* be responsible for the development of distinct cultures in online groups, as language is often the sole semiotic system available to group members (Thorne, 2008).

Language as a meaning-making resource

The systemic-functional approach foremost involves an understanding of language as a semiotic system strategically drawn upon by language users as a meaning-making resource—‘how people use language with each other in accomplishing everyday social life’ (Eggins, 2004, p. 2). The orientation of the theory toward function and meaning-making may be contrasted with phrase structure grammars, which typically do not attempt to account for meaning or pragmatics, and which are not grounded in analysis of realised linguistic patterns, but instead seek to provide rules that conform to speakers’ intuitions regarding what is grammatically possible to say (Martin, 1992). SFL, in contrast, understands language as *purposive*: its use is motivated by purposes that may or may not be transparent or tangible. This is an appropriate theoretical stance for investigations of linguistic choices in an online community, because all communities are inherently social, and because in text-based CMC language is the main resource upon which members can draw to achieve their respective goals.

Interpersonal and experiential functions of language

From a systemic perspective, the simultaneous provision of health information and social support within OSGs is realised by users attending simultaneously to two metafunctions of language: an interpersonal dimension, responsible for negotiating role relationships, and an experiential dimension, responsible for communicating propositions about the world. This is the most useful affordance of SFL for investigation of OSGs, because the division of types of meaning is central to the reason for the community’s existence: if the social element were not desired, users would likely be content to simply read from static pages; if experiential content were not desired, health communities and the threads within them would not need to be organised by topic and subtopic. The interest in the phenomenon of *advice* noted in Section 2.2.2 would likewise stand to benefit from clearer delineation of the relationship between social roles and ideational content, each of which occupies a distinct, but overlapping space within acts of advising or directing others to act. Such a delineation would likely not be unwelcome. Early conceptualisations of the process

of social support bear striking resemblance to the systemic model of interpersonal meaning: Shumaker and Brownell (1984, p. 11) remind us, for instance, that social support is foremost ‘an exchange of resources between two individuals’.

SFL and corpus linguistics

In systemic-functional terms, the corpus is a large collection of instances of language use. These instances are typically organised by register, or by register dimensions. Automated counting of patterns in these instances is one possible way to uncover the grammar—that is, the patterns that generalise across the instances. A particular benefit of CL approaches is that they make it possible to sketch out the probabilities of certain choices in certain contexts. At the same time, CL approaches can test key tenets of SFL theory (Honnibal & Curran, 2007). Clarke (2012), for example, uses a corpus-based approach to test the context metafunction hookup hypothesis—that is, the connection between language and context.

CL and SFL also share a number of underlying similarities. Most obviously, both are concerned with analysis of natural language, and both share a conceptualisation of *register* as playing a significant role in shaping the lexicogrammatical patterns to be found within texts (Hunston, 2013, 4–5). Another key point of convergence between SFL and CL is in the treatment (explicit in SFL, generally implicit in CL) of context as being *contained within* instantiated texts—‘context is in text’, rather than around it (Eggins, 2004). Systemicists evidence this assertion through the fact that we can often accurately deduce the overall functions, purposes and genres of highly decontextualised fragments of texts: *Submissions must contain 8–10 references* can be quickly identified as part of a set of instructions for the submission of academic work, based purely on its lexical (submissions, references) and grammatical (nominalisation, modalisation, etc.) properties. In the same way, Halliday conceptualises lexicogrammatical features of texts as probabilistically determined by their context. That is to say, a given constellation of interpersonal, experiential and textual variables (e.g. the writing of a professor to undergraduates in a written course overview) will likely contain the kinds of lexicogrammatical features described in the example above (Halliday, 1991). This conceptualisation is of great benefit to corpus

linguists interested in discourse analysis: given that CL methods inherently involve the stripping of contextual information from natural language, and that analysis of contextualised language is preferred within the discourse analytic tradition, the recognition of context within lexicogrammar allows CL practitioners to address the common claim (e.g. Virtanen, 2009) that corpus and discourse linguistic approaches, having very different orientations toward context, are difficult to reconcile.

3.3.5. Overview of relevant elements of SFG

The core of SFL is its grammar, known as the SFG (Systemic Functional Grammar), which is organised along the three axes discussed above. In this section, I describe the parts of the SFG that are most relevant to the case study. The major references for this section include Halliday and Matthiessen 2004, Matthiessen 1995, and Eggins 2004.

Register

Register occupies the level of abstraction above discourse-semantics, where language interfaces with a context of situation. The total content of each metafunction (interpersonal, experiential and textual) respectively corresponds to the three register variables of *Tenor*, *Field* and *Mode* (Halliday & Hasan, 1989). Halliday provides a minimal definition of each component:

1. ‘The FIELD OF DISCOURSE refers to *what is happening*’
2. ‘The TENOR OF DISCOURSE refers to *who is taking part*’
3. ‘The MODE OF DISCOURSE refers to *what part the language is playing*’ (1989, p. 12)

In SFL, doing register analysis can refer either to creating qualitative descriptions of the Field, Tenor and Mode of a text (or collection of related texts), or to the process of generating quantitative, corpus-based models of these features (Lukin, Moore, Herke, Wegener, & Wu, 2011; Matthiessen, 2015a). Qualitative descriptions are intended to provide ‘[an interpretation of] the social context of a text, the environment in which meanings are being exchanged’ (Halliday & Hasan, 1989, p. 12).

Quantitatively, the internal dimensions of a single register can be modelled by counting relevant lexicogrammatical features. The distribution of Process Types and the participants engaged in them, for instance, can be used to determine how speakers construe the world around them; MOOD features can be analysed in order to see how speakers position themselves with respect to others (see below).

Using a cartographic metaphor, Matthiessen (2015a, 2015b) calls for further work within functional linguistics that models the internal dimensions of registers, and then situates these models within a cluster of related registers (presented later as Figure 8.1). By comparing these distributions to those found in other registers, registers can be arranged as a landscape, either along the axis of stratification or instantiation. The ultimate, perhaps distant goal, is to synthesise information about individual registers, in order to form a description of the meaning potential of a language, realising the notion of language as ‘an assembly or assemblage of registers’ (2015a, p. 44).

Matthiessen (2013) has also applied register analysis to the domain of healthcare. Within the consumer-centred paradigm, this means focussing on individual health-care consumers’ health journeys, rather than on, for example, how consumers move through a given hospital or clinic in general. These healthcare journeys, Matthiessen explains, consist of sequence of encounters with formal and informal healthcare institutions. Each encounter presents a registerial configuration of topic, interactants and media: we may speak with a doctor in a busy clinic about our symptoms, or to a mental health practitioner about our personal relationships in an hour-long consultation; we may phone an insurance carrier to clarify an issue about coverage, and write to a family member on Facebook about what we have learned. The case study of this thesis, centred on the registerial environment of an OSG, has yet to be described within SFL or HC literature. The registerial description provided in Chapter 8 is therefore intended to facilitate the addition of this register to the current body of knowledge and descriptions of the kinds of registers that may be encountered within consumers’ journeys beyond the clinic.

Interpersonal meanings, MOOD and MODALITY

The Tenor variable of register is realised by interpersonal meanings. In turn, these meanings are made via the MOOD and MODALITY systems of the lexicogrammar. The interpersonal metafunction is responsible for negotiating role-relationships with other speakers. It thus facilitates a constant *exchange* of material and semiotic commodities; it is used to *enact*, and to *interact*.

Speech Function and Mood/Indicative Type

In the SFG, at the broadest level, utterances involve two potential *speech roles*: *giving* and *demanding*. Within either speech role, two types of commodities may potentially be given or demanded: *information*, or *goods and services*. This leads to four main *speech functions*, each of which is congruently realised by a different MOOD TYPE. Information is given via statements, which are realised with declarative Mood. Information can be demanded via questions, congruently realised with the interrogative Mood. Goods and services are given via modalised declaratives. Demands for these non-semiotic commodities are made via commands, which are realised with (non-indicative) imperative Mood. These intrastratal relationships are summarised in Table 3.3.

	Information	Goods and services
Giving	statement → declarative	offer → modalised interrogative
Demanding	question → interrogative	command → imperative

Table 3.3: Speech roles, commodities, speech functions and congruent MOOD TYPES in SFL

Grammar of the MOOD system

Each indicative clause contains a *Mood Block*, comprised of a *Subject* and *Finite*. Optionally, clauses may also contain a *Residue Block*, which can include a *Predicator*, a *Complement*, and/or one or more *Adjuncts*. As non-indicatives, imperatives are comprised entirely of the Residue Block. In the case of declaratives, to determine which element is the Mood Block, a reverse-polarity tag can be added to the end of the declarative:

Geoff had difficulty with the task, *didn't he?*

Whatever is referenced by the pronoun in the tag is the Subject component of the Mood Block. The Finite is the first verb or modal in the verbal group that follows. It too is duplicated in a tag question in the case of *be*, *have* or a modal Finite; it is replaced by *do* in other cases. The *Polarity*, which is often unmarked when positive, is the opposite of the polarity in the tag. These elements together form the Mood Block. What remains in the clause is the Residue.

Residue also has component parts: the *Predicator* is the part of the verbal group that carries a sense of the actual process undertaken. When a clause has only a single-word verbal group, it functions as both Finite and Predicator. If there is secondary tense information such as a progressive element, it is realised in the Predicator (Halliday & Matthiessen, 2004). Residue may also contain a *Complement*. The Complement is what becomes Subject if the clause is passivised. Finally, *Adjuncts* are elements of the Residue that contribute non-essential information. Generally realised by adverbial or prepositional groups, Adjuncts cannot be turned into Subjects through passivisation.

Adjuncts may be *circumstantial* (experiential), *modal* (interpersonal), or *textual* (thematic). Circumstantial adjuncts are those that add information concerning cause, time, matter or agent (discussed in the next section). Modal adjuncts¹⁵ provide information concerning mood or polarity, or comment (the speaker's assessment of the clause) or vocative (naming speakers for addressal, turn-taking, etc.). Finally, thematic adjuncts may have the role of facilitating conjunction or continuity with other clauses.

Importantly, some Mood Adjuncts are realised through *grammatical metaphor*—that is, when what is congruently expressed through one rank or component is expressed through another.

GRAMMATICAL METAPHOR

I think he came back.

CONGRUENT FORM

= *Perhaps* he came back.

I think, for example, involves a grammatical metaphor—in this case, upward rank shift, where meaning typically realised by a word (in this case, a modal operator or

Mood adjunct) is realised by an entire clause (Halliday, 1966b; Taverniers, 2002). The fact that *I think* is functioning as a constituent, rather than as an independent clause, can be ascertained by the tag question test: in the examples below, we can see that *he* and 's are typically the Subject and Finite within the Mood adjunct:

I think he's OK, *isn't he?*

*? I think he's OK, don't I?

Because *I think* is not functioning as an independent clause, it 'play[s] no part in the structure of the interaction' (Halliday & Matthiessen, 2004, p. 162).

Differences in Indicative/Mood Type are realised by different configurations of the Subject, Finite and Predicator. For polar interrogatives, the Finite and Subject simply switch places. Aside from copula constructions, whenever the Finite and Predicator are realised with the same word, the operator *do* must be inserted as the Finite for interrogatives:

Have you gone this week?

Do you know the way?

For WH-Interrogatives, 'the WH-element is always conflated (mapped onto, fused with) another element of clause structure' (Eggins, 2004, p. 175). It may potentially be mapped onto Subjects, Complements or circumstantial Adjuncts. Whichever element it is mapped onto is the same element needed in a minimally satisfactory reply:

Who moved the table?

Mahsa.

When did he do it?

An hour or two ago.

The Mood structure of basic imperatives involves removing the Mood element altogether:

INTERROGATIVE/QUESTION

IMPERATIVE/COMMAND

*Have you been taking your
meds?*

Take your meds.

Metaphorical realisations of Speech Function

The correspondence between interpersonal semantics and MOOD choices is not always one-to-one. In situations where interactants are of a relatively equal social

status, incongruence between Mood Type and Speech Function emerges as a politeness or hedging strategy. In particular, commands may be realised through alternative Mood Type choices (Table 3.4). The most appropriate and/or likely choice is dependent on the overall Tenor of an interaction, and on who is speaking. In unequal relationships, the higher status participant tends toward congruent, imperative realisations of commands, while the lower status participant opts for incongruence (*Would you be so kind as to ...*).

Mood type	Realisation
Command	<i>Go back on the meds.</i>
Declarative	<i>I think you should go back on the meds</i>
Interrogative	<i>Why don't you go back on the meds?</i>
Mod. declarative	<i>I would go back on the meds (if I were you).</i>

Table 3.4: Grammatical metaphor as politeness strategy

As reviewed in the previous chapter, the relationship between congruence of MOOD TYPE selection and the role relationships of speakers in an interaction has already been noted in the context of advice provision by DeCapua and Huber (1995, see Section 2.2.2). Accordingly, in an OSG, we can expect that the role-relationship disparity between long and short term users may manifest linguistically in imperative commands being issued by veterans, with the proportion of advice issued via imperatives increasing with the membership length of the veteran member. Notably, this kind of incongruence poses a challenge for corpus-based methods.

MODALITY

MODALITY is a key resource within the interpersonal exchange. It is realised within Mood structure in two ways: *modalisation* and *modulation* (Eggins, 2004, p. 179). Functionally, modalisation is when a speaker makes meaning related to the *probability* or the *usuality* of an event: as Eggins explains, ‘modalisation is the expression of the speaker’s attitude towards what s/he’s saying’ (2004, p. 180). Grammatically, modalisation may be realised through Mood adjuncts (as explained earlier), a modal operator, or both at the same time. In every case, the modal elements are gradable.

MOOD ADJUNCT	MODAL OPERATOR	BOTH TOGETHER
She <i>certainly</i> knew the text well.	Power <i>may</i> have been restored.	I <i>could sometimes</i> win.

Modals may therefore be expected to appear in a number of situations within an OSG, including marking of hesitation by newcomers (Vayreda & Antaki, 2009; Weber, 2011), and in veterans' polite sharing of information and potential actions.

POLARITY

POLARITY is the binary opposition between positive and negative. In SFL, POLARITY is primarily considered an interpersonal feature, since in contracted forms it becomes attached to the Finite (Halliday & Matthiessen, 2004, p. 143). It represents the two poles of certainty, with modalisation and modulation construing various levels of certainty and uncertainty in between. Halliday and Matthiessen (2004) provide corpus evidence that clausal POLARITY is positive in 90 per cent of cases; negative polarity, they claim, is always the marked variant. Its meaning, however, is highly dependent on the content of the clauses to which it responds, functioning as affirmation or denial of the validity of the previous proposition or proposal. As such, it is difficult to track using existing corpus methods. This difficulty is exacerbated in the case study, which treats individual posts, rather than threads, as the unit of analysis. Though the proportion of responses that affirm or deny the content of the previous clause would no doubt shed light on the extent to which exchanges are harmonious or contentious, this is unassessable without shifting the unit of analysis to the thread, which would in turn make difficult the analysis of how language use changes over the course of membership. POLARITY is therefore accounted for in the case study mostly for the sake of completeness of the overall analysis of the MOOD system, and the interpersonal meanings the system is responsible for realising.

Experiential meanings and the system of TRANSITIVITY

As explained earlier, clauses in SFL are multifunctional units. At the same time as interpersonal relationships are enacted and negotiated through choices of MOOD and MODALITY, experiential meanings are being expressed via the TRANSITIVITY system.

Through this metafunction, speakers make *representations* of inner and outer states, and of Things in the world and Events that they are involved in. In each clause, language users are *construing* reality as change, which may affect or be affected by participants. Coherent clause complexes therefore represent sequences of change.

In SFL, the TRANSITIVITY system can be analysed from two complementary perspectives: the *transitive model*, in which processes are grouped according to types, and the *ergative model*, where processes are bound to Mediums, through which they are made possible and therefore, without whom they cannot exist.

Transitive model

In the transitive model, the clause is structured around a *process*, indexed mostly by the Event—that is, the rightmost verb in the verbal group (and prepositions, for phrasal verbs):

<i>i consider myself blessed having you guys</i>	<i>Oh ...and her ex bf glen broke up with his gf</i>
--	--

This process must belong to a *Process Type*. Process Types are distinguishable based foremost on lexis, but also on the grammar of the clause they head. Various syntactic and/or semantic tests can be used to disambiguate more difficult cases, though accurate Process Type identification by both human and machine has often proven challenging (O'Donnell, Zappavigna, & Whitelaw, 2009). Each Process Type constrains the available configurations of participants (nominal groups) and circumstances (adverbial or prepositional phrases) in the clause. A summary of Process Types, identification tests and associated participant roles is given in Table 3.5.

#	Process Type	Definition	Typical verbs	Identification test	Participants
1.	Material	Doing tangible things	kick, draw	give, <i>What did x do?</i>	Actor (goal) (range) (beneficiary)
2.	Mental	Thinking or feeling	think, believe	What do you think, feel or know about y?	Senser, phenomenon
3.	Verbal	Saying	shout, yell, tell	What did x say?	Sayer, (receiver) (verbiage)
4.	Behavioural	Psychological and physiological behaviour	cough, smile, dream	Unmarked present tense has continuous sense	Behaver (behaviour) (phenomenon)
5.	Existential	'There' clauses	be	Presence of non-locative there	Existent
6a.	Relational (ident.)	Ways of being	equal, mean, symbolise	x is a member of the class a.	Attribute, carrier
6b.	Relational (attr.)	Defining	own	x serves to define the identity of y.	Cannot be passivised

Table 3.5: Summary of Process Types

Ergative model

The ergative model of TRANSITIVITY conceptualises each clause as a quantum of change in the world, centred on a Process being carried out through a Medium. When the cause of the change is not mentioned, the result is a *middle clause*—a construal of a *happening*. In cases where the Medium does not bring about the process, an Agent may also be added (resulting in an *effective clause*, or a *doing*).

MEDIUM + PROCESS: HAPPENING

*my dd is hitting manic mode and i
think it 's time*

AGENT + PROCESS + MEDIUM: DO-
ING

*you hit a nerve and i felt the need to
stand up for myself*

A critical difference between transitive and ergative models is that core participant roles can be disambiguated in the ergative model simply by checking for the exist-

ence of multiple nominal group arguments of a process: a two-participant process will likely contain a *Medium* and an *Agent*, while a single-participant process will be a Process-Medium configuration. For corpus/computational linguistics, the ergative model therefore represents an ability to more easily identify agency, which can be expected to play an important role in the way participants in a Field of discourse are construed. That said, the transitive model is equally useful, highlighting experience of certain types of change by social actors. The two models are complementary; each is drawn upon in the analysis of TRANSITIVITY (Chapter 7) according to its strengths.

Context of culture: genre and ideology

While the case study of the thesis for the most part involves methods drawn from CL, the analysis begins with a generic (i.e. genre-based) interpretation of contributions to the Forum. For this reason, in this section, I provide a basic overview of a conceptualisation of genre that has emerged from the systemic-functional school, and an associated method for performing genre analysis.¹⁶

In contrast to the context of situation, which concerns the environment in which a given text is produced, the context of culture refers to the broader conditions common to texts. Culture is the highest level of abstraction in SFL—all other meaning systems exist within and belong to cultures (Halliday & Hasan, 1989). Thus, the context of culture is the semiotic potential of the totality of sign systems.

The best-articulated component of the context of culture is the notion of *genre*—that is, ‘recurrent configuration[s] of meaning, phased in discourse as a staged, goal-oriented social process’ (Martin, 2013, p. 9). Genres ultimately derive their meaning from their instantiation within a given culture (Halliday & Hasan, 1989, p. 99). Though genre and register alike are realised by the contextual variables of Field, Tenor and Mode, genre theory emphasises the social purpose of the activity being undertaken in the text (Christie & Martin, 2005; Martin, 1992). Thus, genres may be characterised as constellations of Field, Tenor and Mode that are culturally recognised as performing social functions (Eggins, 2004). It is also important to bear in mind when considering genre that individual genres are not necessarily distinct:

macro-genres may contain *micro-genres*: the macro-genre of the *essay*, for instance may draw on micro-genres of *expositions*, *discussions* and *evaluative accounts*.

Eggins and Slade (2004) provide simplified, actionable parameters for performing systemic-functional genre analysis. The six steps are outlined below.

Recognising a generic text

Given that genres are culturally recognised by their very nature, simple reading of texts by those fluent in the relevant culture may be enough to identify the existence of a genre. For analysts and interactants alike, genres are recognised when texts appear to ‘move through predictable stages’ (2004, p. 213). Eggins (2004) argues that the existence of a word for the kind of behaviour seen in the text (i.e. *purchasing*, *commentating*, *gossiping*) may at the very least be a clue that a potentially definable genre exists.

Defining the social purpose of the generic text

This step involves clarifying the overall function(s) of a genre with as much specificity as possible—rather than ‘telling a story’, sub-categorisations such as ‘anecdote’ or ‘exemplum’ are more useful, given the existence of micro- and macro-genres. More theoretically, this task also involves developing an understanding of how the genre constructs a social reality: by virtue of their existence alone, recognisable genre stages may provide insight into social practices and culturally accepted attitudes and values.

Identifying and differentiating stages within a genre

After breaking down the text into clauses as with lexicogrammatical analysis, groups of these clauses must be divided by their role within the text. These roles should be functionally, rather than formally defined: *Abstract* or *Resolution* is superior to *Beginning* or *Chapter Three* because the latter are not genre specific. By convention, each of these functions should then in turn be described in prose.

Specifying obligatory, optional and recursive stages

Stages may be obligatory, optional or recursive. Obligatory elements are considered to be defining features, and in some cases may be unique to the genre under investigation. Optional stages, on the other hand, are likely to be present in other genres. Halliday and Hasan (1989) remind us that optional stages do not occur randomly: in the *buying and selling genre*, the number of customers in the store or the size of the line may affect whether or not a *greeting* or *sale initiation* takes place. Both optional and obligatory stages may be recursive, as in the case of the buying and selling genre, in which *sale request*, *sale enquiry*, *sale compliance* and *sale* may go through limitless iterations (Halliday & Hasan, 1989, p. 61).

Devising a structural formula

The next task is to represent the genre structure. By convention, each genre stage is delineated by a caret (^). Optional stages are bracketed. Recursive stages are square-bracketed, with brackets followed by ⁿ.

A number of notion schemes for representing the relationship between genre stages have been proposed, and many have undergone revisions in order to be easier to render on a computer (Eggins, 2004). Problematic is that many (e.g. those in Halliday & Hasan, 1989) are esoteric, and lagging behind the representational schemes of other grammars in terms of readability (Hovy, 1996). In fact, the development and use of unique means of expressing optionality and recursion is perhaps superfluous and ultimately unhelpful, given that comparatively well-known systems such as *regular expressions* could potentially represent generic structure in a format familiar to at least some non-SFL practitioners.

A particularly important addition to structural representation of genre is Hasan's (1985) notion of *generic structure potential*: that is, a maximally expanded representation of genre staging that exhausts all possibilities for additional optional stages and recursion.¹⁷

Analysing the semantic and lexicogrammatical features for each stage of a genre

As Hasan explains,

a text has many modes of existence and so it can be analysed at many different levels, with each contributing to our understanding of the phenomena involved (1989, p. 116).

Thus, relevant parts of SFG may be operationalised in order to investigate the phenomena of interest: a researcher interested in power dynamics within a text, for example, would likely perform an analysis of features of the MOOD system, as these are responsible for the management of role-relationships between interactants. Simultaneously, this analysis provides a justification of the treatment of the text as instantiating a genre and of the clause complexes as instantiating generic stages. Eggins and Slade (2004) note that since lexicogrammar can provide hints as to genre staging, this step of the analysis may render it necessary to reconsider the previous delineation of stages.

Genre may influence the lexicogrammar of texts in two ways. First, as genres are configurations of register variables, texts within genres must necessarily conform at the level of lexicogrammar. In this way, a genre such as *sports commentary* is likely to bring about language which experientially positions players, teams, umpires and coaches as the main participants. Second, different genre *stages* may influence lexicographical decisions. The *evaluation* stage within the *storytelling* macro-genre, for example, is likely to opt for a declarative Mood, while experientially, it can be assumed that mental and relational Process Types may occur.

Ideology in SFL

Ideology has a complex history within SFL, originally conceptualised as the most abstract analysable stratum of text/context, as the overarching determinant of the context of culture (Eggins, 2004). Later work, however, often refrains from accounting for ideology (e.g. Matthiessen et al., 2010), or disavows its existence as a distinct stratum within the hierarchy of stratification (Martin, 2006, e.g.). Regardless of its exact status within SFL, ideology is commonly discussed in the contexts of both OSGs and socialisation, and, accordingly, is in need of a brief description here.

One way of conceptualising ideology is ‘from above’. As Banks (2009) explains, two texts with very similar Field, Tenor and Mode can nonetheless mean very different things: politicians with opposing views, for instance, may give speeches

within more or less identical registers, while ultimately expressing different sets of values. Ideology can also be conceptualised ‘from below’: ideological values may be found within lexicogrammatical and semantic patterns that are common within a text, but that are rarely subject to negotiation, controversy or debate by the interactants. Experientially, while ‘Field of discourse’ refers to what is being directly spoken about in a text, ideology refers more to the taken-for-granted assumptions about these things and events. For example, while there is a great deal of argument about the efficacy of some kinds of alternative medicine in many OSGs, there is generally much less debate about the value of treatment itself: treatment can lead toward health, and is almost always worth undertaking. Some ideological values in these communities, therefore, may be that *that illness requires treatment*, because *treatment can make you healthy again*. The controversy surrounding pro-anorexia OSGs (see e.g. Chancellor, Mitra, & De Choudhury, 2016) stems from the fact that these communities advocate challenging widely held ideological values about health, weight, illness and treatment.

In this thesis, without taking a particular stance on the role of ideology within SFL, the term can still be operationalised in the case study as a way of referring to the more or less unargued or inarguable components of discourse within the Forum. It is important to bear in mind, therefore, that ideology may or may not be substantively different from discourse-semantics or register; rather, here, it is simply a way of thinking about what the language users in texts take for granted or consider common-sense.

3.3.6. Criticism of SFL

Both the overall orientation of SFL and its grammar have received a number of criticisms that deserve to be addressed. Van Dijk (2004), for example, has taken issue with three main dimensions of SFL as a linguistic theory in general. First and most broadly, he notes that its sheer density creates difficulty when a researcher seeks to use only relevant sections of the theory: ‘not only are the terms (Field, Tenor, Mode) hardly transparent, as to their intended meanings, but also the usual—informal—descriptions of their meanings are barely enlightening’ (2004, p. 341).

Second, he characterises SFL as lacking sufficient engagement with potentially useful interdisciplinary perspectives: ‘there is very little inspiration from the many other approaches to context in linguistics and especially in anthropology, sociology or social psychology, at least in the analysis of the context’ (2004, p. 342). Finally, he points out that SFL has avoided engaging with cognitive accounts of language, and therefore can have little to say about the reality of the meaning-making process for speakers themselves.

Indeed, each of the three points is in some sense valid. SFL is terminologically dense—a necessary evil, according to Halliday’s preface to his *Introduction to Functional Grammar*, when formulating a theoretical account of something as complex as a human language. Van Dijk’s second point, regarding a lack of dialogue between SFL is also accurate. At a terminological level, for example, key terms in SFL such as *Tenor* have been coined in order to avoid ambiguity—a more natural term might be *tone*, but this term also has a phonological meaning. The claim that SFL engages little with related traditions at both the level of terminological and beyond is not a baseless one, however: as a second example, the 650+ page *Introduction to Functional Grammar* does not once mention *pragmatics*, though what is meant by pragmatics in related areas overlaps to a very significant extent with the systemic conceptualisation of interpersonal discourse-semantics.¹⁸

The third criticism, related to the lack of cognitive accounts, is the only one which has some bearing on the overall explanatory power of the theory. Here, Van Dijk is not alone in his criticism: Rohdenburg (1996) has argued that the systemic account of preposition/object ordering in clauses with phrasal verbs (*She put the fire out* vs. *She put out the fire*) as being motivated by textual thematic demands (i.e. information structure, or givenness/newness) is only a partial account of the underlying motivations for speaker choices. Cognitive demands on addressees play an important role in choices between grammatical alternatives, with the likelihood of clause-final preposition decreasing steadily as weight and length of the object grows (c.f. Hawkins, 1992). While a lack of engagement with cognition may have benefits for certain kinds of text analysis, as well as some computational tasks (O’Donnell, 2014), it is prudent to bear in mind that even if the systemic account may not in

and of itself be incorrect, there is compelling evidence for its instead being in some respects only a partial account of how language works.

Others also critiqued the SFG's textual metafunction: Huddleston (1988) and Wid-dowson (2008) have problematised the notion of the Theme as simply the first element in a clause, in cases where there are dummy subjects, for example. More generally speaking, Van Dijk has also further criticised what he sees as an attempt to force lexicogrammatical *conjunction* to line up with the register variable of *Mode*. Perhaps this criticism is tacitly embraced by Eggins and Slade (2004), who tend to advocate the use of terminology and concepts from CA, rather than SFL, for investigations of turn-taking, overlap and the like. For these reasons, compounded simply by issues of scope, I have opted not to consider textual meanings in this thesis in any serious detail. This is perhaps unfortunate, as much could foreseeably be learned about OSGs through analysis of the ways in which the take-up of advice (for example) is realised in forum threads (see Section 8.2.2 for a brief demonstration of the possible value of analysis of choices of Theme).

3.4. Summary: theories and methods for investigating online healthcare discourse

The synthesis of literature presented in this and the previous chapter has involved four research domains:

1. Computer mediated communication (CMC) as a Medium, and forums as a Mode
2. Healthcare communication (HC) as a Field
3. Corpus linguistics (CL) as an approach for analysing digital linguistic data
4. Systemic functional linguistics (SFL) as a framework for understanding language

The argument made throughout the literature review is that by combining these areas, we can learn new things about online intra-consumer health discourse, as well as uncover new ways to learn things. Key claims from qualitative literature can be tested quantitatively with largely automatic CL methods, and evidenced by translating observed discourses into their lexicogrammatical realisations.

The next part of the thesis, which introduces the case study of a bipolar disorder OSG, demonstrates the potentiality of combining CL methods and SFL theory for the analysis of CMC. The systemic grammar and sophisticated corpus methods make possible nuanced kinds of corpus interrogation that differentiate interpersonal from experiential meaning-making.

4. Case study design

In the previous chapter, I reviewed relevant theory and methods for functional analysis of language use in OSGs, centring on CL as a potential approach and SFL as a theory of language. This chapter introduces the case study of the thesis—a corpus-based analysis of an online bipolar disorder support group. In the sections that follow, I describe the site selection and ethical considerations, data collection, and corpus building processes. I then describe the analytical methods used in the investigation, and the development of `corpkit`, a Python module for building and interrogating corpora, and for editing and visualising interrogation results. Its core functionality and user interfaces are summarised. Justifications are provided for key choices throughout the chapter.

4.1. Site description and selection

The case study of this thesis is a corpus comprised of 13 years of posts to a once-popular OSG for those living with bipolar disorder—a mental disorder characterised by oscillation between elevated and depressed moods (Anderson, Haddad, & Scott, 2012). The forum has been online since 2001, with very little change to the interface or account system. As such, the Mode dimension of the Forum has stayed largely static, and many users have participated using the same account details for a number of years. The Forum is a part of a larger architecture of health-related boards, many of which were created at different points in time. Users' accounts work site-wide; as such, many of the contributors to the Bipolar Forum have also created posts elsewhere. At the time of data collection (early 2014), the Bipolar Forum had over 5700 users, with a handful of 'veteran' users recording over a thousand posts (see

Figure 4.2). By the time the Forum content was harvested, it was in a state of decline: 2013 saw a total of only 119 new threads, compared with over 1,699 during its peak in 2007 (see Figure 4.4). The average number of replies to threads has also dropped dramatically since its peak (from 7.59 in 2009 to 1.40 in 2013—see Figure 4.6). Observation of the Forum between 2014–2016 suggests that these trends have continued.¹⁹

Forum architecture

The community is a prototypical example of a *vBulletin*-powered message board, with a main page showing the most recent active threads, and with each thread containing a set of chronologically ordered posts. Users are able to send private messages, search Forum contents, and show posts related to specific medications or symptoms. The community has explicit rules (in the form of *sticky posts* at the top of the message board) stating that new users should not provide offline contact information, or request diagnoses from other members. Posts breaking rules could be edited or deleted by moderators, who can also comment on their moderation decisions. If provided, a user's gender and/or location are displayed beside his/her post, alongside the username and current postcount.

Forum users

Based on location metadata, members are overwhelmingly from Anglophone countries (see Figure 4.1). Most members believe themselves to have bipolar disorder, or have at some stage been diagnosed with bipolar. Less common are family, friends and partners of somebody believed to have the condition. This distribution is similar to the forum investigated by Vayreda and Antaki (2009). Notably, there is a related forum for ‘Family & Friends of Mentally Ill’ whose use is encouraged in an administrator’s sticky post. Health professionals are either not present in the community or do not identify themselves as such. The Forum’s rules explain that the site is for consumers, rather than professionals:

Do not register or post your past, current or future health topic profession in any manner. The boards are to be used for PATIENT opinions, only. Professional titles



Figure 4.1: Stated locations of members

lend undue weight to what is to be only your opinion. Health profession titles are not allowed.

Concordancing the terms used by Forum members for health professionals (*doctor*, *doc*, *pdoc*, *tdoc*, *gp*, *shrink*, *psychologist*, etc.) revealed no instances of users claiming to have formal medical training.

Unlike in Maclean et al's analysis of *Forum77* (2015), users who were active at the time of data collection were not excluded from the dataset. A potential effect of this is that the ratio of veteran to non-veteran users is slightly skewed toward non-veteran users, as some currently active members may eventually progress to become veteran members in the future.

One final consideration regarding the user base is that nothing prevents users from creating and posting from multiple accounts. As such, following De Choudhury, Kiciman, Dredze, Coppersmith, and Kumar (2016), the terms *user*, *member* and *con-*

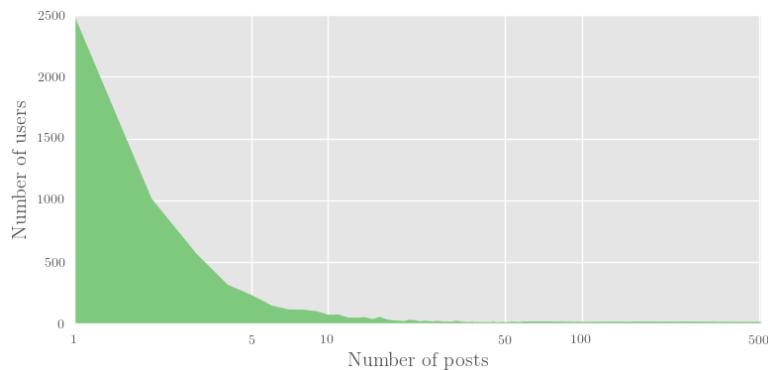


Figure 4.2: Total members by postcount

tributor refer more specifically to users' accounts than to actual people who use the board. That said, there is no evidence to suggest, and no reason to believe, that users create and post under multiple accounts.

4.1.1. Justification for site selection

The Forum is a typical example of a large online message board. As discussed in Chapter 2, these kinds of communities are well-studied, allowing more reliable connection of findings to earlier research into OSGs than would be the case if the selected site was from an emerging mode of CMC. The fact that the Forum has undergone few structural changes since 2001 is particularly useful, as it eliminates a potential confounding variable, where longitudinal changes in language use may be affected by changes in the ways users create and transmit messages. While it is possible that hardware changes (faster web access, increased mobile usage, etc.) could have effects on language use, such effects also occur in other non-researcher-constructed communities. Compared to social networking sites and mobile apps, both of which typically undergo constant interface redesign, the Bipolar Forum can be seen as relatively stable over time.

An important reason for selecting the Forum is its size: with over eight million words of public posts, it is large enough to make quantitative generalisations about language use not only in the corpus, but also in its individual subcorpora. A large dataset increases the level of delicacy that can be quantitatively discussed, as delicate features are by definition rarer than broad features.

Another motivation for choosing the Forum is its general Tenor, which is formal enough to be parseable with off-the-shelf parser models trained on formal English: correct capitalisation, punctuation and paragraph breaks are, for most users, the norm. Texts authored for other varieties of CMC, such as Twitter, may be much more difficult to parse, due to novel linguistic features such as hashtags, emoticons and embedded links.

The fact that the Forum is text-heavy and light on multimodal features is also useful, as this means that the CL approach is quantifying a very large amount of the meaning-making that occurs within the community. In systemic terms, the *division*

of labour in the community is almost entirely semiotic, rather than material; as such, text analysis takes us further in describing and explaining the community than it would in a domain where much of the meaning-making is material. This is the case even when comparing to other forms of CMC (*YouTube, Instagram, Snapchat, etc.*), where text comments play a secondary role to images, audio and/or video. Even so, because discourse analysis ideally involves analysis of texts in something resembling their original contexts, the HTML content of all pages was retained, so that selected examples could be considered with sensitivity to extralinguistic meanings made by multimodal site features.

4.2. Ethics

The ethical parameters of the investigation were informed by consideration of:

1. International research ethics literature that deals with CMC and OSGs
2. The medium and situation factors unique to the Bipolar Forum
3. The *Australian National Statement on Ethical Conduct in Human Research 2007 (updated May 2015)*

The guidelines in the National Statement take precedence over the broader international literature. That said, the National Statement provides little guidance on CMC, which often challenges existing notions of privacy, anonymity and informed consent. As such, consideration of ethical concerns raised in the relevant literature is prudent. Strategies employed to minimise risk to participants are also described.

4.2.1. Considering CMC/OSG data

Much has been written about the ethics of researching CMC (boyd, 2007; Ess, 2007; Eysenbach, 2000; Hewson, 2015; Landert & Jucker, 2011; Stevens, O'Donnell, & Williams, 2015; Walther, 2002). Common is a recognition that different modes require different ethical considerations: forms of CMC that can be easily connected to real-world identities (due to the use of real names, profile photos, etc.) necessitate different protocols from forms that are highly anonymised. Similarly, the level of

privacy and anonymity is not necessarily the same for all modes: on YouTube, video bloggers show their faces and often give their real names, while commenters on the clips are generally highly anonymised.

Another factor influencing ethical decisions is the size of the dataset. In the case of very large quantities of user-generated natural language, it may become impossible for researchers to check that data has been fully anonymised. For these reasons, literature concerned with the ethics of using Facebook/chat data (e.g. Hudson & Bruckman, 2004; Zimmer, 2010) or small-scale, qualitative datasets (e.g. Eysenbach & Till, 2001; Roberts, 2015) are not particularly relevant here.

The use of online forums as data sources has been a contentious issue. Though most acknowledge that forums are ‘public’, the fact that contributions are often authored in private spaces such as bedrooms may influence the candidness of texts (Hewson, 2015). Others have questioned the usefulness of the ‘public/private’ binary in online contexts more generally (Lange, 2007). When compared to face-to-face data, researchers also know less in general about the participants, making it difficult to assess potential harm on an individual basis. In quantitative studies of thousands of users, spanning over a decade of contributions, this becomes an impossibility.

Another issue is the potential for quotations, even when stripped of contextual information, to be linked back to their source: since texts in online forums are indexed in search engines, simply entering quotes into a search engine will often quickly uncover the original thread. There, one can access the user’s profile, and in some cases even send the user a private message. One strategy for avoiding this issue has been to paraphrase or translate quotes (Stommel & Meijman, 2011; Vayreda & Antaki, 2009). In an investigation of subtle changes in linguistic choices, however, this sacrifices the authenticity of the data.

Hewson (2015) and Markham and Buchanan (2012) argue that flexible, bottom-up, contextually sensitive parameters should be created for any study of CMC. Accordingly, below, I summarise issues of privacy, anonymity and informed consent with respect to the specific medium and situation factors of the forum.

Contact with participants

Participants were not contacted at any stage before, during or after the study. Therefore, none of the data was researcher-elicited, and no non-publicly available data were generated for the project. The National Statement includes ‘damage to social networks’ (p. 13) as a kind of harm that research can do to participants. Not contacting Forum users therefore minimises harm by preserving an existing social structure as-is. At the same time, the decision not to contact participants made it impossible to obtain informed consent. The highly anonymised nature of the board, as well as the longitudinal focus of the case study, however makes obtaining such consent from all contributors, or even a plurality thereof, impossible (Kaufman & Whitehead, 2016; Stommel & Meijman, 2011).

Anonymity and privacy

Online spaces vary in the extent to which users’ contributions are publicly accessible, and in how much users can reveal about their identity. In the Bipolar Forum, users protect their anonymity: they use nicknames, and do not contribute information that could identify them offline, such as real names, addresses, social security numbers, or photographs. In both the account creation and posting guidelines, users are reminded to keep Forum contents anonymous, and to assist in the anonymisation of others:

Do not register your surname or put identifying info in your profile or signature:
Your username, profile, signature and messages must not identify you to readers
or contain any part of your email address, blog or website.

Please only use first names of members. To protect anonymous use of the site do
not include surnames. Use the listed first name or the username.

Moderators also have the power to censor content that may de-anonymise the user. Ultimately, users appear to adhere to these rules: searching the corpus for addresses, email addresses and phone numbers did not turn up a single instance that needed anonymisation, nor did any such information emerge throughout the course of the investigation.

Finally, all analysed data is publicly available. Though some researchers have problematised academic use of user-generated CMC (e.g. Eysenbach, 2000; Zimmer,

2010), such critiques centre on the notion that participants are unaware of the publicly available nature of their contributions. This is not a reasonable argument in the case of the Bipolar Forum,²⁰ for a number of reasons:

1. Users can see how many others are currently online, and how many people have viewed each thread.
2. The main page invites users to **Subscribe** for an automatic bulletin of popular posts.
3. Every page contains a **Search** button, showing that all posts are archived and indexed.
4. Users' profiles are explicitly called **Public Profile Pages**. Help pages for Public Profile creation remind users that what they add is 'publicly available'.

To argue that Forum users are not aware of the public nature of their interactions is to argue that users fundamentally misunderstand the design and function of the community. No evidence supporting the idea that users have such a misunderstanding was found over the course of the investigation.

4.2.2. Interpreting the National Statement

Under the National Statement, application for review and clearance from a Human Research Ethics Committee is required when the research carries more than a low risk of harm, distress or, at minimum, discomfort, to participants.²¹ Exempted from the review process, however, is research that:

1. is negligible risk research; and
2. involves the use of existing collections of data or records that contain only non-identifiable data about human beings (p. 70).

The case study qualifies as negligible risk, as any potential risks for Forum contributors do not rise to the level of discomfort. The data qualifies as non-identifiable, as they 'have never been labelled with individual identifiers' (p. 27). More specifically, the Forum data constitutes a subset of the non-identifiable data class, which

are those that can be linked with other data so it can be known that they are about the same data subject, although the person's identity remains unknown (p. 27).

It is indeed possible to connect different contributions to a single author due to the username (and potentially, the linguistic content of the contribution). This, however, cannot be connected to individual identifiers.

Because the case study is exempted from the process of ethics review, and due to the consideration of broader literature and site-specific factors, an application for review was not made.

Participants with mental health issues

The majority of users of the Forum may be classified as having a mental illness. The National Statement mandates that the ‘distinctive vulnerabilities’ of such people must be taken into account, while also protecting the entitlement of such people to participate in research. Researching those with mental health issues may also involve differing guidelines for obtaining consent. According to the Statement, however, in cases where ‘research uses collections of non-identifiable data and involves negligible risk’, the need for informed consent is waived.

4.2.3. Minimising risk

Under the National Statement, ‘researchers have an obligation to minimise the risks to participants’ (p. 14). To minimise any potential risk of privacy invasion, users’ public profiles were excluded from data collection, and usernames have been paraphrased throughout the thesis. Any part of the Forum restricted to registered members was not accessed or considered in the analysis. The corpus therefore includes only text that is freely accessible by navigating the Bipolar Forum.

To ensure that no social structure is disturbed, and to ensure contributors’ privacy, all texts chosen for sustained, qualitative analysis are authored by users who are now inactive within the community, having not posted in the past year.

4.3. Corpus building

In the following section, I describe the process of turning the Forum’s contents into structured, annotated corpora (the *Bipolar Forum Corpus*).

4.3.1. Content retrieval

All threads of the Bipolar Forum were downloaded as HTML to a local machine using a purpose-built tool, based on **GNU Wget**, on December 3rd, 2013. This 1.8GB collection of almost 19,000 files, including metadata (usernames, timestamps, users' locations, etc.) comprises the total dataset for the thesis.

4.3.2. Corpus creation

Python's *lxml* module was used to extract the text and relevant metadata (e.g. speaker, timestamp, etc.) of posts within each thread. Regular expressions were used to search for details in text requiring anonymisation, such as addresses, email addresses and phone numbers. This did not turn up any personal details—the only matches were for emergency services, health-related hotlines, and, more rarely, specific hospitals. For the sake of parser accuracy, some basic spelling normalisation was then performed. Apostrophes were reinserted: *im* became *i'm*, and *whod* became *who'd*. Ambiguous cases (*shell*, *hell*, *wed*, etc.) were left uncorrected. Also changed were common contractions such as *gonna* and *wanna* into *going to* and *want to*. Alternative spellings of *bipolar* (*bi-polar*, *bi polar*) were also normalised. All other misspellings were left uncorrected. The results—the cleaned text of each post in the thread, and some its metadata features (post count at time of posting, date of post, username, gender, location, etc.)—were then saved into text files representing each thread. Below is an example of a single post and its associated metadata in XML form.

```
1 I hope everyone is hanging in with this blasted heat. As we all know
  ↵ being hot, sticky, stressed and irritated can bring on a mood
  ↵ swing super fast. So please make sure your all takeing your meds
  ↵ and try to stay out of the heat. <metadata username="Emz45"
  ↵ totalposts="5063" currentposts="4051" date="2011-07-13"
  ↵ postnum="0" threadlength="1">
```

The tools developed for the investigation (see Section 4.4) were then used to extract the metadata, parse each cleaned text with the *Stanford CoreNLP 3.6.0* pipeline, and to reintroduce the metadata to the parser output. The result of this process was a large collection of files in *CONLL-U* format (Nivre, 2015)—an example of the parsed rep-

resentation of the post above is presented in Figure 4.3. In this representation, the columns are *token index*, *token*, *lemma*, POS, *named-entity tag*, *governor*, *dependency type*, *dependent(s)*, and *coreferences*. The two rightmost columns are not a part of the CONLL-U specification, but have been added by `corplkit` during post-processing to speed up interrogations at runtime. Note that the constituency parse is treated as a metadata feature, so as to not violate the format specifications.

1	# sent_id 1
2	# parse=(ROOT (S (NP (PRP I)) (VP (VBP hope) (SBAR (S (NP (NN ↳ everyone)) (VP (VBZ is) (VP (VBG hanging) (PP (IN in) (IN with) ↳ (NP (DT this) (VBN blasted) (NN heat)))))))) (. .)))
3	# speaker=Emz45
4	# totalposts=5063
5	# threadlength=1
6	# currentposts=4051
7	# stage=10
8	# date=2011-07-13
9	# year=2011
10	# postnum=0
11	I I PRP 0 2 nsubj 0 1
12	hope hope VBP 0 0 ROOT 1,5,11 –
13	everyone everyone NN 0 5 nsubj 0 –
14	is be VBZ 0 5 aux 0 –
15	hanging hang VBG 0 2 ccomp 3,4,10 –
16	in in IN 0 10 case 0 –
17	with with IN 0 10 case 0 –
18	this this DT 0 10 det 0 2
19	blasted blast VBN 0 10 amod 0 2
20	heat heat NN 0 5 nmod:with 6,7,8,9 2*
21	.

Figure 4.3: A sentence from the Forum, parsed and stored alongside metadata in CONLL-U format

The texts were put into ten subfolders, representing each of the ten stages of membership. This is the default subcorpus format recognised by the interrogation tool. To investigate a different structure, such as language use in the Forum by year, the tool can simply be told to treat the `year` metadata values as the subcorpora. In this way, it is possible to use the same corpus to answer address a number of possibl research questions. Setting `speaker` as subcorpora would create subcorpora for each of the 5,818 unique members (See Table 4.1), allowing, for example, ranking of users according to some linguistic criterion. Setting `gender` as subcorpora would allow investigation of differences in the language use of those who identify as male,

as female, and those who choose not to provide their gender. As the main concern of this thesis is linguistic change over the membership course, however, almost all querying of the corpus targeted the membership stage variable (the `stage` metadata). That said, other structures are also briefly utilised, in order to control for self-selection bias. An overview of each of the four symbolic subcorpus structures is given in the sections below.

Membership Stage Structure

The main subcorpus structure used in the thesis is the *Membership Stage Structure*, which consists of ten subcorpora of almost equal size, approximating ten ‘stages of membership’. The first subcorpus contains all first posts to the Forum. The number of posts in this sub-folder (5,818) dictated the sample size for the next subcorpus. The second subcorpus contained each user’s second and third posts; the third subcorpus contained posts 4–7, and so forth. This structure makes it possible to learn *how language changes over the course of membership in the community*. Table 4.1 provides an overview of the composition of each subcorpus. Table 4.2 provides absolute frequencies for shallow linguistic features. These features are among those used to generate shallow findings in Chapter 5, and to aid in the calculation of relative frequencies in Chapters 6 and 7.

It is important to note that this structure operationalises *veteran membership* entirely according to a user’s number of posts. This unidimensional approach has the advantage of simplicity, making it possible to examine the question of *how post count affects language use*. At the same time, using a single feature to segment the data means that segmentation is not based on algorithm whose efficacy has not yet been proven in other contexts. It is theoretically simplistic, however—as mentioned in Section 2.2.2, ideally, veteran membership would be determined based on a number of factors including not just number of posts, but duration of membership, the average number of replies received, and any explicit privileges the user may have within the community. Pfeil, Svangstu, Ang, and Zaphiris (2011), for example, provide an alternative method of differentiating between forum members, clustering users by similarity in contributing behaviour (who users reply to) and

the effects of this behaviour (who responds to the contribution). In descriptions and analysis that follow, unless otherwise noted, *Veteran membership* refers to Subcorpus 10, and *New members* refer to Subcorpus 1. It should be borne in mind, however, that a veteran post or contribution is nothing more than a contribution made by a user who has already posted at least 559 times before.

Subcorpus	01	02	03	04	05	06	07	08	09	10
Post range	1	2–3	4–7	8–15	16–30	31–58	59–115	116–219	220–559	560+
Texts	5,818	5,689	5,607	5,937	5,790	5,875	5,848	5,757	5,789	5,570
Users	5,818	3,348	1,777	1,004	529	284	148	76	38	8

Table 4.1: Bipolar Forum Corpus, Membership Stage Structure: key attributes of each subcorpus

	Characters	Tokens	Words	Closed class	Open class	Clauses	Sentences
01	4,380,658	1,258,606	1,092,113	643,779	614,827	277,103	68,267
02	3,185,042	922,243	800,046	471,883	450,360	209,448	51,575
03	3,157,277	917,822	795,517	471,578	446,244	209,990	51,860
04	3,261,922	948,272	820,193	486,065	462,207	216,739	53,995
05	3,164,919	921,098	796,430	473,446	447,652	210,165	52,227
06	3,187,420	928,350	797,652	480,843	447,507	209,895	52,171
07	3,080,956	900,110	771,319	466,254	433,856	202,868	50,071
08	3,356,241	972,652	833,135	502,913	469,739	218,382	52,637
09	2,908,221	840,803	725,108	434,839	405,964	191,851	47,050
10	2,868,652	815,101	708,918	421,403	393,698	185,677	43,474

Table 4.2: Shallow features of the *Membership Stage Structure*

Future Veteran Structure

The second structure, the *Future Veteran Structure*, is a subset of the Membership Stage Structure, with any contribution from users with fewer than 30 total posts removed. This structure makes it possible to *isolate veteran members' language change*, accounting for a potential self-selection bias, where those who go on to become veterans have different linguistic patterns even during their initial contributions. The removal of non-future-veteran contributions means that there are very few posts in the first four subcorpora. For this reason, during analysis, the first four

subcorpora are conflated, in order to keep subcorpora quantitatively reliable in terms of word count.

	Characters	Tokens	Words	Closed class	Open class	Clauses	Sentences
01–04	2,520,668	729,694	631,438	375,581	354,113	165,348	40,332
05	2,360,495	687,311	593,178	354,671	332,640	156,941	38,776
06	3,187,420	928,350	797,652	480,843	447,507	209,895	52,171
07	3,080,956	900,110	771,319	466,254	433,856	202,868	50,071
08	3,356,241	972,652	833,135	502,913	469,739	218,382	52,637
09	2,908,221	840,803	725,108	434,839	405,964	191,851	47,050
10	2,868,652	815,101	708,918	421,403	393,698	185,677	43,474

Table 4.3: Shallow features of the *Future Veteran Structure*, with subcorpora 1–4 collapsed

Longitudinal Structure

The third structure, the *Longitudinal Structure* is chronological, with 13 annual subcorpora. In this structure, threads, rather than posts, can become the unit of analysis. This structure makes it possible to observe phylogenesis—that is, *longitudinal change in the norms of the Forum itself* that result from the constant stream of incoming and outgoing members.

	Characters	Tokens	Words	Closed class	Open class	Clauses	Sentences
2001	21,743	6,061	5,197	2,938	3,123	1,272	333
2002	193,771	55,979	47,920	28,291	27,688	12,752	2,807
2003	2,283,828	656,838	568,429	332,839	323,999	146,266	29,394
2004	2,484,517	708,587	613,078	358,589	349,998	149,217	30,810
2005	4,710,146	1,366,425	1,176,398	699,091	667,334	304,403	60,756
2006	6,512,854	1,851,707	1,627,163	945,150	906,557	429,597	85,598
2007	8,827,854	2,525,622	2,221,659	1,292,264	233,358	590,495	114,341
2008	2,634,440	762,596	662,350	388,076	374,520	172,333	37,527
2009	4,653,461	1,328,212	1,159,234	674,634	653,578	303,264	61,381
2010	931,807	264,611	232,084	133,755	130,856	58,713	12,806
2011	890,052	253,953	222,752	128,321	125,632	55,440	11,442
2012	392,618	111,959	97,118	56,020	55,939	24,149	5,597
2013	202,393	56,803	49,623	28,634	28,169	12,711	2,859

Table 4.4: Shallow features in the *Longitudinal Structure*

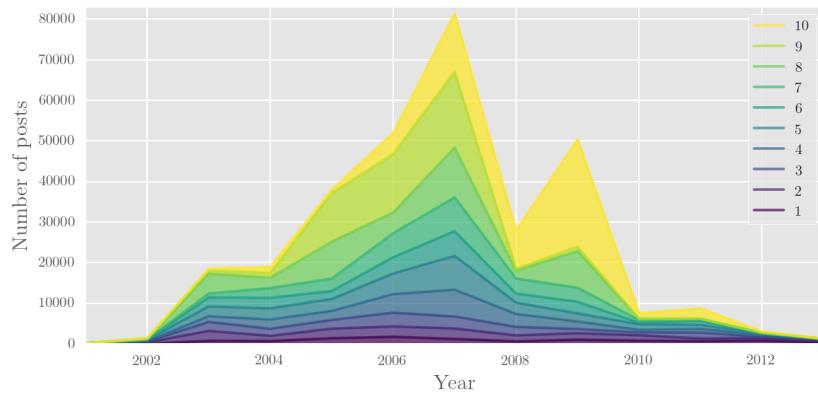


Figure 4.4: Number of posts in the *Longitudinal Structure* by membership stage

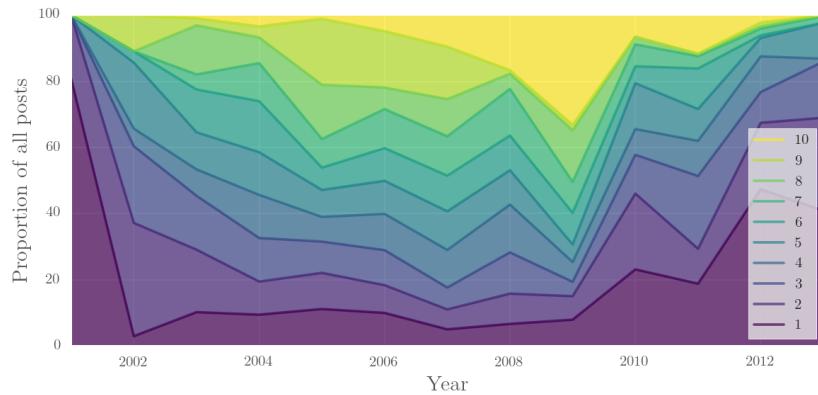


Figure 4.5: Relative number of posts in the *Longitudinal Structure* by membership stage

Figure 4.4 uses the Longitudinal Structure to provide an overview of the amount of activity within the Forum by year. The number of posts steadily increases until 2009, at which point there is a similarly steady decline. Figure 4.5 shows the extent to which different membership stages are represented in the Forum in a given year. Having information about the composition of membership stages at different points in time makes it possible to speculate as to what a low or high concentration of veteran members does to language use in the community as a whole. This is not, however, a focus of the thesis.

Figure 4.5 shows that in the busiest years of the Forum's history, veteran members made proportionally more posts.²² These members then gradually stop posting; in the latest years sampled, many newcomers post a question, but receive no reply (see Figure 4.6). Combined, the smaller number of overall posts, the departure of

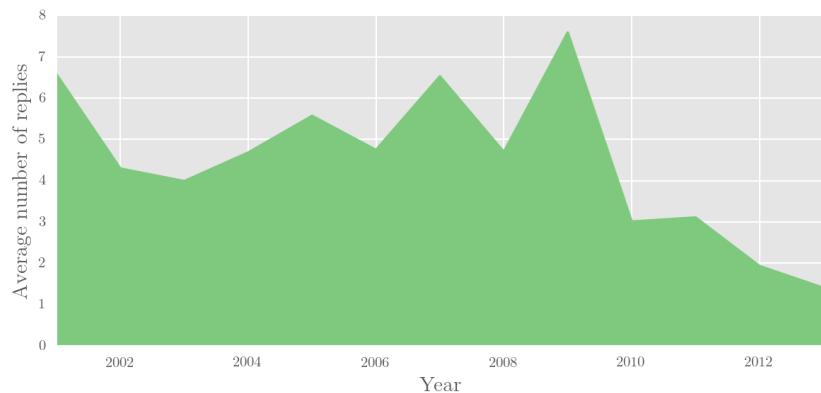


Figure 4.6: Average number of replies to new threads in the *Longitudinal Structure*

veteran members, and lack of follow-up to new threads are taken as indicators that the Forum is moribund.

Comparative Structure

The final structure, the *Comparative Structure* consists of two subcorpora. In the first are the first 30 posts of users who did not progress beyond 30 posts ('Dropouts'). In the second are the first 30 posts of those users who did ('Future veterans'). This structure makes it possible to *check for differences in the early-stage language of future-veterans and early dropouts*. It is designed to address the possibility that there is a pre-existing difference in the linguistic features of members who drop out early and members who become veterans.

	Characters	Tokens	Words	Closed class	Open class	Clauses	Sentences
Dropout	15,480,624	4,478,616	3,883,247	2,293,005	2,185,611	1,011,519	252,142
Fut. vet.	17,070,684	4,946,441	4,257,184	2,559,998	2,386,443	1,120,599	271,185

Table 4.5: Shallow features in the *Comparative Structure*

4.4. corpkit: tools for corpus building and analysis

Data analysis involved the use of a purpose-built Python-based toolkit, **corpkit**. In this section, I explain the need for the tool, its development, core functionality, and

significance within the research area. Complete documentation of its programmatic, graphical and interpreter interfaces is available via GitHub.

4.4.1. Rationale for tool development

Before analysing the corpus data, a survey of existing tools for corpus analysis was conducted, with a number of criteria in mind:

1. Feature-richness: ability to perform lemmatisation, lexical and grammatical searches, handling of plain text, POS tags, constituency and dependency parses
2. Automatability: ability to loop through subcorpora, or through lists of queries, or cycling through options such as lemmatisation, removal of closed-class words, etc.
3. Flexibility: ability to add to wordlists, edit interrogation results (merging search results, spanning subcorpora, etc.)
4. Sensitivity to functional grammar(s): ability to operationalise functional linguistic notions when interrogating corpora and editing results (i.e. locate Events within Processes)
5. Visualisation: ability to generate useful visualisations of language use without the need to export data from the tool

In addition to these selection criteria, preference was given to resources that were:

1. Open-source, freely available, non-proprietary
2. Command-line based, rather than graphical (to facilitate automation, replicability, etc.)

With these criteria in mind, the following tools were tested for suitability:

- | | | |
|---------------|--------------------|---------------------|
| 1. NLTK | 5. Wordsmith Tools | 9. Corpus Workbench |
| 2. WMatrix | 6. Sketch Engine | (CWB)/CQP |
| 3. AntConc | 7. UAM Corpus Tool | |
| 4. CasualConc | 8. Tregex | |

Though individual features from each tool were useful, none satisfied all selection criteria. Shifting between tools during different stages of the investigation was also

unfeasible, as much interrogation was exploratory, iterative, or cyclical. **UAM Corpus Tool**, while able to handle multiple subcorpora, and while providing systemic-functional conversions of **Stanford CoreNLP** parses, was also not suitable, as it:

1. Struggled to cope with the size of the corpus²³
2. Had stability issues (for example, when exporting interrogation results)
3. Is GUI-based, and cannot be scripted
4. Does not grant the user direct access to the systemic-functional parses or the parser itself, making it difficult to verify the accuracy of converted parses²⁴
5. Requires exporting data in order to sort or visualise results

Command-line tools such as the **Open Corpus Workbench** (Evert & Hardie, 2011) provided the necessary ability to perform iterative and recursive searches, but lacked the ability to extract complex features from full syntactic parser output, and to easily interface with state-of-the-art tools for manipulating and visualising data. **Tregex** (Levy & Andrew, 2006), a search query language for constituency trees, is able to perform complex queries, but does not handle lemmatisation, keyword calculation or advanced result manipulation. **NLTK** (Bird, Klein, & Loper, 2009), while containing routines for many linguistic tasks, did not provide an interface to parsing or interrogating **CoreNLP** dependency parser output,²⁵ and did not provide a holistic interface for more general workflows. Given these considerations, the development of purpose-built tools was the best solution, providing increased transparency and reproducibility of this investigation, while also being useful for other researchers interested in functional CL.

4.4.2. Contents of the toolkit

corpkit is a **Python** module designed to create and interrogate parsed and structured corpora, edit interrogation results, and to display/visualise edited output. It is available as open-source software:

1. GitHub: <https://www.github.com/interrogator/corpkit>
2. PyPI: <https://pypi.python.org/pypi/corpkit>
3. Documentation: <http://corpkit.readthedocs.org/>

4. Standalone app: <http://interrogator.github.io/corpkit/>

The toolkit is *object-oriented*. Users instantiate corpora as objects, which have methods for parsing, interrogating and concordancing. Interrogations are also objects, which have methods for editing, the calculation of statistics, saving and visualising. A basic workflow using the Python API involves the following steps:

1. Create a project to house corpus/corpora, saved data, images, wordlists
`(new_project() function)`
2. Instantiate plaintext corpus
`(Corpus class)`
3. Parse plaintext corpus using **Stanford CoreNLP**
`(Corpus.parse() method)`
4. Interrogate/concordance parsed corpus for lexicogrammatical phenomenon
`(Corpus.interrogate() method)`
 - Constituency parses via `Tregex/nltk_tgrep`
 - Dependency parses via `pandas`
5. Edit results
`(Interrogation.edit() method)`
 - Keeping, removing, merging entries or subcorpora
 - Calculating relative frequencies
 - Sorting, generating statistics, doing linear regression
 - Keywording
6. Tabulate, export or visualise edited results
`(Interrogation.visualise() method)`
7. Save data to project
`(Interrogation.save() method)`

Interrogation objects store a dictionary of the parameters that created them, so that they can easily be reproduced. Saved interrogation results are loaded as corpus attributes at the beginning of each session, simplifying the process of managing projects involving multiple corpora.

Key design parameters

The tool was designed in response to the needs of the case study, to noted shortcomings of existing tools, and to a small body of literature describing needs in the future generation of corpus tools (e.g. Anthony et al., 2013; Anthony, 2006; Gries, 2013). Anthony et al. (2013) provide a succinct summary of needed tool development:

[Linguistic research] will rely increasingly on large corpora, advanced functionality, and sophisticated statistical methods. ... Corpus tool development should be an open source initiative with tool components being developed in a modular fashion. By dividing tool components in this way, it becomes easier for tool functions and features to be extended, modified, or simplified depending on the need' (2013, pp. 155–156).

`corpkit` responds to each of these parameters: it is built to work with corpora of any size, and to allow multiprocessing to speed up queries over very large datasets; it has the most advanced functionality of any corpus interrogation tool to date, with support for parsing, interrogating parser output, and distinguishing between subcorpora and metadata tags or values; it integrates with `scipy` (Jones, Oliphant, & Peterson, 2001) and `pandas` (McKinney, 2010) in order to allow complex mathematical operations; it is free and open-source; it is modular, interfacing with dedicated modules for editing, storing and visualising; graphical and command-line interfaces tailored to programming and non-programming linguists.

4.4.3. Interfaces and functionality of the tool

`corpkit` includes three different interfaces. The API itself, used for all data analysis in the thesis, is the most powerful, but requires knowledge of Python. The other two interfaces—a graphical application, and a natural language interpreter—call the API as a backend. These interfaces were developed with the aim of increasing the potential users of the tool to those without a background in computer programming. At the same time, because users can shift freely between interfaces, it was hoped that the tool could facilitate increasingly programmatic workflows in CL. In the sections below, I explain the functionality of the tool, giving examples from the API. The graphical and interpreter interfaces are introduced later.

API

The most complex interface is the API, implemented in Python. Through this interface, users can access the full range of methods for a given object, and take advantage of common programming constructs, such as the writing of loops or conditional statements. It is the interface used for the case study itself. The API is also the interface with the most potential use for computationally intensive downstream applications in the area of medical/clinical NLP: it could easily be scripted to automatically parse, categorise and search new data as it becomes available.

Corpus

The `Corpus` class models a directory (and optionally, subdirectories) of data files, which may be plain text files or `CONLL-U` data. Though the toolkit is oriented toward parsed and structured data, basic functionality for interrogating and concordancing is available for plain text corpora as well. `Corpus` objects have a `parse` method, which is essentially a wrapper around `Stanford CoreNLP`, with keyword arguments for annotators, memory allocation and so on. In the example below, a plaintext `Corpus` object is created and parsed with `Stanford CoreNLP` using default parameters.

```
1 from corpkit import Corpus  
2 unparsed = Corpus('forum')  
3 corpus = unparsed.parse()
```

This method returns another `Corpus` object representing the parsed data, which can then be interrogated and concordanced in complex ways. Each subcorpus and file is represented as `Subcorpus` and `File` objects respectively, which can also be interrogated and concordanced:

```
1 parsed['01']  
2 # <corpkit.corpus.Subcorpus instance: 01>  
3 parsed['01'].files[245]  
4 # <corpkit.corpus.File instance: thread-4450244.txt.conll>
```

Speaker segmentation

One innovation in `corpkit`'s `parse` method is the addition of speaker segmentation. The `parse` method has a `speaker_segmentation` argument, which will add speaker

names to `CoreNLP` output, provided they are delineated with a colon at the start of a line, or by HTML/XML tags in the text. The process of speaker segmentation involves:

1. Creating a duplicate corpus with names removed
2. Parsing the duplicated corpus
3. Using character offset metadata in the parser output to find the original line in the duplicated text
4. Lifting the speaker name from the original corpus
5. Adding the speaker name to the parser output

The Bipolar Forum Corpus also has speaker names included in the annotations, so that interrogations can be restricted to a particular user or set of users. It is therefore also possible to look for differences between how newcomer and veteran users use language in one or more threads. At a broader level, this method makes it possible to computationally model register features of dialogic text, facilitating context-responsive parsing (see Chapter 9). Such tasks are beyond the scope of the current investigation, however.

Corpus.interrogate() method

As explained in the previous chapter, CL is centrally concerned with extracting frequency information from texts. Generally speaking, researchers are interested in either counting the occurrences of a particular linguistic feature in each subcorpus (e.g. counting a particular word or grammatical feature), or in counting the possible realisations of a linguistic feature (counting the frequencies of every word that is a noun, or every subject in a passive construction). Perhaps surprisingly, few tools provide an interface for doing this kind of searching or counting. `corpkit` focusses on iterating over subcorpora, extracting complex lexicogrammatical features, and tabulating the results.

The `Corpus.interrogate()` method, more precisely, centres on a three-step process of *searching, excluding and showing* (Figure 4.7). Searching and excluding involve the specification of one or more combinations of *search objects* and regular expression or wordlist-based patterns to match. Search objects are broken down into a token of

interest (i.e. a token, its governor, one of its collocates, etc.) and its attributes (i.e. its word form, its lemma form, its POS, etc.—see Table 4.6). The `GL` search object, for example, searches for any lemma form matching a pattern, returning the ID of its dependent(s). After each search criterion has been processed, indices of tokens are removed if they appear fewer times than the total number of search criteria—that is, tokens must match all search criteria by default. Then, any explicit excluding is performed, using the same syntax as searching: if a token matches the exclusion criteria, it is filtered from the set of matches. After exclusion is complete, a function is called that determines how to represent the search matches. If the user inputs `show=[I, P, L, GL]`, the program will output the index, POS, lemma form and governor’s lemma form of a match (e.g. ‘`2/NNS/user/be`’). Results are returned as a two-dimensional array of counts for each shown object in each subcorpus. Figure 4.7 provide an examples of this method for constituency and dependency parses. In both cases, lemmatised adjectives modifying terms for doctors are returned. This produces the results shown in Table 4.7.

	Match	Gov.	Dep.	Coref Head	N-gram	Collocate	1L	1R
Word	W	GW	DW	HW	NW	BW	-1W	-1W
Lemma	L	GL	DL	HL	NL	BL	-1L	-1L
Function	F	GF	DF	HF	NF	BF	-1F	-1F
POS tag	P	GP	DP	HP	NP	BP	-1P	-1P
Wordclass	X	GX	DX	HX	NX	BX	-1X	-1X
Index	I	GI	DI	HI	NI	BI	-1I	-1I
Sent. index	S	GS	DS	HS	NS	BS	-1S	-1S
Named entity	E	GE	DE	HE	NE	BE	-1E	-1E

Table 4.6: Possible object–attribute combinations to search, exclude or show

```

1 # define tree-based (tregex) query
2 tquery = {T: r'/_JJ.?/_ > (NP <<# (/NN.?/_ < /^(doctor|dr\.|p*doc)s*/))' }
3 # search the corpus, output matching lemmata
4 t_doc = corpus.interrogate(tquery, show=L)
5 # via dependencies: search func, gov. lemma, gov. pos
6 dquery = { F: 'amod',
7             GL: r'^^(doctor|dr.|p*doc)s*',
8             GP: r'^N' }
9 # search the corpus, output matching lemmata
10 d_doc = corpus.interrogate(dquery, show=L, conc=True)
11 # show table
12 print(d_doc.results.to_latex(columns=range(4)))

```

Figure 4.7: Using `corpkit` find Epithets/Classifiers of health professionals in the Bipolar Forum Corpus

Subcorpus	<i>new</i>	<i>good</i>	<i>different</i>	<i>regular</i>
01	58	16	29	12
02	53	25	19	15
03	45	17	11	13
04	60	21	16	19
05	57	30	8	12
06	57	21	15	18
07	65	25	13	21
08	90	29	8	22
09	51	14	15	11
10	84	40	20	6

Table 4.7: Adjectives modifying doctor words

Concordancing

Concordancing is the process of bringing up a vertically aligned set of query matches, with a window of characters or words on either side (see Section 11). Concordancing has typically been lexically oriented: users search for words via simple or regular expression queries. Some concordancers allow POS tagged datasets (e.g. *AntConc*); command-line tools may allow searching of information such as lemma forms. `corpkit`, however, extends concordancing to the same range of features as can be searched and shown via lexicogrammatical querying; in fact, the tool treats concordancing and interrogating as two variants of the same task of interrogation, where concordances are the (monomodal) full text, optionally alongside grammatical

ical information, and lexicogrammatical interrogations are the frequency counts for searches, reduced to their smallest meaningful form. For this reason, both practices are performed by the same method. In the previous code example, the `conc=True` keyword argument is responsible for generating the concordance shown in Table 4.8.

ASHEA	i mention he has not seen this	actual	doctor for at least 2 years .
ASchwagz	i have seen 3	different	family doctors , 2 different psychi
ASchwagz	3 different family doctors , 2	different	psychiatrists and 1 endocrinologist
Allibo	of course , when i called my	primary	care dr. to prove the urologist wro
AllieTr	find a	good	doctor .
Althea	thyroid meds , psychiatrist or	other	doc ???
AmySue902	through out the last	few	years i have been diagnosed bd by a
AmySue902	i have been diagnosed bd by a	few	other doctors but have always refus
AmySue902	have been diagnosed bd by a few	other	doctors but have always refused the
Angie	i have a	great	doctor that only charges me \$ 50 a

Table 4.8: Concordancing adjectives modifying *doctor*

Interrogation.visualise() method

Visualisation makes it possible to compress large amounts of complex numerical information into forms that aid in the recognition of patterns. Visualisation of linguistic data has two main purposes. First, it can be used to help the researcher to understand and interpret data. Second, it can be used to present results to readers. Most corpus tools require data to be exported and visualised in dedicated tools. This stifles the ability to use visualisation as a way to generate insights, prioritising only the use of visualisation to display data for readers. `corpkit`, on the other hand, contains an interface tailored specifically to the visualisation of linguistic data, via `matplotlib` (Hunter, 2007). Users can thus produce visualisations without leaving `corpkit`, making it possible to generate visualisations for both purposes.

The method allows bar charts, pie charts, line charts, area charts, and so on. Data can be stacked or shown cumulatively. The interface simplifies titling, axis labelling, legend placement, figure sizing, as well as the use of dozens of colours and styles. As used in Chapter 7, heatmaps can also be created. These can be constant or diverging, and work with any kind of frequency (absolute, relative, keyness, etc.) that can be calculated with the tool.

Automating workflows

The tools were developed to facilitate automatic processing of the dataset, with as little need for human intervention and manual analysis as possible. This means that the tool must automate the kinds of things that corpus linguists typically perform by hand. To give an example, a researcher may determine that *wonder* is a frequent or apparently interesting process in a corpus. In response, he/she may:

1. Count frequency and calculate keyness of *wonder*
2. Concordance *wonder*
3. Curate results:
 - Remove instances where *wonder* is nominal
 - Remove instances where *wonder* is not a process
4. Identify a phenomenon of interest in clauses with *wonder* as process, e.g.
 - Subjects of *wonder*
 - Modification of *wonder* through modals, adverbs and negation
5. Exclude non-relevant examples
6. Read through concordance lines, count results, describe patterns, contrast with other features ...

In this kind of workflow, concordancer output is used both to get a sense of how a lexicogrammatical feature behaves, and to do content analysis or thematic categorisation. There are limits, however, to the number of examples human coders can process. Using parsed corpora, however, it is possible to automate a much larger portion of the method. In `corpkit`, the user can search for any participant or modifier that has a governing lemma *wonder* that fills the role of *process*. The `edit` method can then be used to create relative frequencies, and the `visualise` method to plot the top results. Concordancing, in this kind of workflow, can be done to ensure that the query is matching the expected kinds of patterns in text, or to provide a selection of contextualised examples for the reader. Finally, it is important to note the potential power of iteration that comes with the programmatic approach: the developed code (Figure 4.8) could be placed inside a loop, so that processes other than *wonder* can be investigated with little additional effort (e.g. `for verb in ['wonder', 'appreciate', 'care', 'need'] ...`).

```

1  from corpkit.dictionaries import processes, roles
2  # a list of verbal processes
3  verbal = processes.verbal.lemmata
4  # match nsubj of these processes as root
5  criteria = {GL: 'wonder',      # governor lemma
6            GF: roles.event,    # governor function
7            GP: r'^V.*',       # governor pos
8            F: roles.modifier + roles.participant1} # function
9  # show function and lemma form
10 data = corpus.interrogate(search=criteria, show=[F, L])
11 # make relative frequencies
12 rel_data = data.edit('%', SELF)
13 # create 2*2 lineplot with subplots
14 rel_data.visualise(subplots=True, layout=(2,2))
15 # make table
16 print(rel_data.results[:4,:4].to_latex())

```

Figure 4.8: Example investigation code: searching dependency parses for participants and circumstances in *wonder* as process

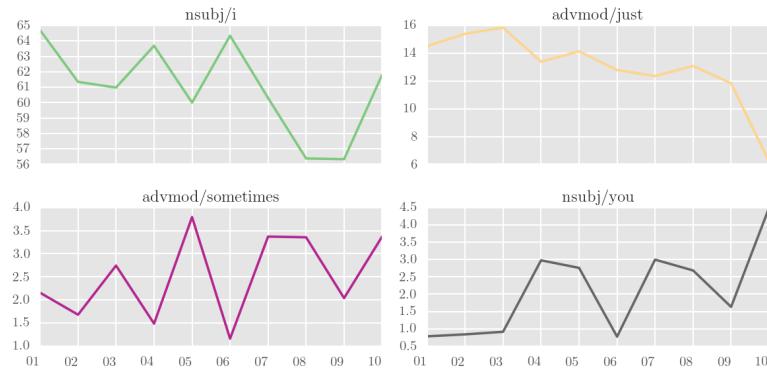


Figure 4.9: Example figure: participants and circumstances for *wonder* as process

	nsubj i	advmod just	advmod sometimes	nsubj you
01	64.705	14.509	2.156	0.784
02	61.344	15.406	1.680	0.840
03	60.975	15.853	2.743	0.914
04	63.690	13.392	1.488	2.976
05	60.000	14.137	3.793	2.758

Table 4.9: Example result as a multi-indexed two dimensional array

0		nsubj/i	aux/be root/wonder ./, dobj/what de
1		nsubj/i	root/wonder det/the amod/same dobj/
2		nsubj/i	aux/be advmod/just root/wonder det/
3	nsubj/I aux/be	advmod/just	root/wonder det/what nsubj/person a
4		nsubj/i	root/see poss/my dobj/doc ./on prep
5	nsubj/he root/say	nsubj/he	aux/be ccomp/wonder mark/if nsubj/h
6		nsubj/i	aux/be root/wonder mark/if nsubj/st
7		nsubj/i	advmod/sometimes root/wonder mark/i
8	nsubj/I	advmod/sometimes	root/wonder mark/if nsubj/i aux/hav
9		nsubj/i	root/wonder mark/if nsubj/anyone ad

Table 4.10: Example concordance: participants and circumstances in *wonder* process

Keywording

Keywording is a common means of determining which words in a corpus are unusually frequent, based on a particular statistical measure such as log-likelihood, mutual information or percentage difference (see Section 3.2.4). Traditional tools treat keywording as a kind of corpus search or interrogation. Keywording is better understood, however, as a statistical operation performed on absolute frequency interrogation results. The new method facilitates using any of these statistical measures in two novel ways. First, keywording is opened up to the full power of the `search`, `exclude` and `show` pipeline. This allows keywording of grammatical participants, or of POS-lemma pairs. Keywords for n-grams, groups and phrases can also be calculated. It becomes possible, therefore, to target particular kinds of constructions, and avoid the use of arbitrary stopword lists. Second, keywords can be calculated in subcorpora using the entire corpus as the reference material. This makes it possible to avoid using reference corpora, which are theoretically problematic (See Section 3.2.7).

Wordlists

`corpkit` is designed to be used in tandem with a series of included wordlists. These wordlists were adapted from numerous resources, including the *Process Type Database* (Neale, 2002) and `pattern.en` (De Smedt & Daelemans, 2012). This makes it possible to match or remove words of certain types from further analysis. Wordlists include:

1. Closed class words:

- Pronouns
 - Prepositions
 - Articles
 - Determiners
 - Connectors, conjunctions
 - Modals
2. Systemic-functional Process Types:²⁶
- Mental
 - Verbal
 - Relational
 - Material
3. Conversion/normalisation
- UK/US spelling conversion
 - Manual lemmatisation (to augment/correct automatic lemmatisation errors)
4. Grammar conversion (where possible): from Universal Dependency labels (see Nivre, 2015) systemic-functional labels

Each wordlist is a `Wordlist` object, which has methods for generating verb inflections (via `WordNet`) or generating regular expressions to match list items:

```

1  from corpkit.dictionaries.process_types import processes
2  print(processes.verbal.lemmata[:5])
3  # ['accede', 'add', 'address', 'admit', 'advise']
4  print(processes.verbal.words[:5])
5  # ['accede', 'acceded', 'accedes', 'acceding', 'add']
6  print(processes.verbal.lemmata[:5].as_regex(boundaries='line'))
7  # '(?i)^accede|add|address|admit|advise)$'

```

Every list can be used as criteria for interrogating corpora or editing results. In the graphical and interpreter interfaces, new wordlists can be interactively created, inflected and saved for later use.

Additional features

`corpkit`'s core classes and methods are complemented by a number of other functions for saving and loading data, and building regular expressions from wordlists. These are documented online, and in Appendix B.

The tool also has broader uses beyond those showcased in this thesis. Language models can be generated from corpora, allowing classification of arbitrary texts by their similarity to texts found in a subcorpus. Such methods make it possible to automatically categorise and interrogate new data. The methods could therefore be applied in near real-time to popular online communities, rather than to a community that has completed its lifecycle.

Module dependencies

`corpkit` relies heavily on a number of other modules, the most important of which are outlined in Table 4.11. Notably, `Stanford CoreNLP` is used to parse texts, and either `Tregex` or `nltk_tgrep` can be used to query parse trees.²⁷ `pandas` is another key dependency, being used to store both parser output and search results.

Task	Tool
Tokenisation	<code>CoreNLP</code> , <code>NLTK</code>
Lemmatisation (dependencies)	<code>Stanford CoreNLP</code> , <code>WordNet</code>
XML manipulation	<code>CoreNLP_XML</code>
Parse tree traversal	<code>Tregex</code> , <code>nltk_tgrep</code>
Synonyms, hypernyms, hyponyms	<code>NLTK</code> , <code>WordNet</code>
Verb inflections	<code>pattern.en</code>
Linear regression	<code>scipy</code>
Visualisation	<code>matplotlib</code> , <code>pandas</code> , <code>mpld3</code>
Result manipulation	<code>pandas</code>
Multiprocessing	<code>jobjlib</code>

Table 4.11: Tasks performed in `corpkit`

4.4.4. Alternative interfaces

Though the API was used for the analysis of the Bipolar Forum Corpus, two other interfaces were also created, in order to maximise the utility of the tool for researchers without expertise in computer programming, or, more specifically, in Python. The first is a graphical application. The second is a natural language interpreter, which parses commands entered by the user and invokes the API. These are described in the following two sections.

Graphical interface

The simplest of the three developed interfaces is the graphical interface, built using **Tkinter**. It is reminiscent of other standalone graphical tools such as **AntConc** and **UAM Corpus Tool**, providing interfaces for file viewing, corpus searching and concordancing. It extends upon **AntConc** by adding the ability to parse and search parsed texts, and by allowing the user to work with structured or metadata-rich collections of text. It extends on **UAM Corpus Tool** by providing a greater variety of query languages, as well as result editing and visualisation, and more comprehensive project management, with previous processes stored in memory, so that they may be viewed, saved or discarded. Figure 4.10 shows the main tabs. Not shown are the *Build* tab (which also visualises parsed sentences), and numerous pop-up windows for wordlist creation, thematic category building, and project management.

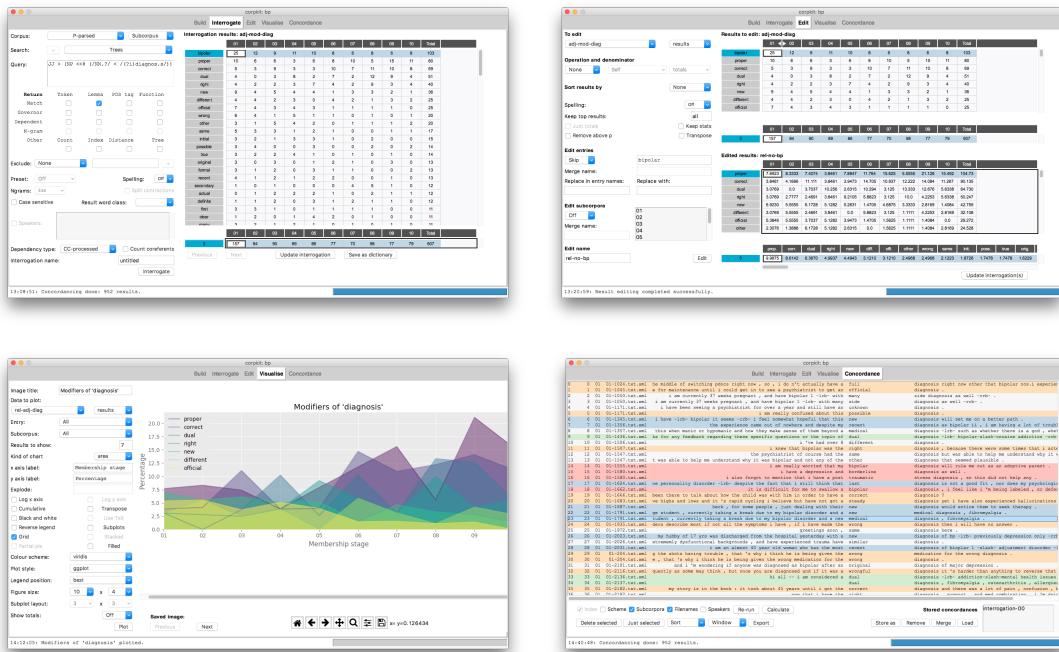


Figure 4.10: Screenshots from corptkit's graphical interface

Interpreter

The other alternative interface is a natural language interpreter, which allows the user to type in imperative commands for `corplib` to perform. It is reminiscent of the CWB. Like the CWB, it allows the user to search corpora and annotations in complex ways. The user can create macros and variables, export output, and write scripts that call the tool's functions and methods. It extends upon the functionality of the CWB by allowing the user to search fully parsed, rather than POS/lemmatised text.

```
1 > set bipolar as corpus
2 > search corpus for governor-lemma matching processes:verbal \
3 ... excluding lemma matching "[write, email, pen]" \
4 ... showing lemma and pos
```

The interpreter occupies a middle ground in complexity between the graphical interface and the Python API. This makes it particularly suitable as an introduction for corpus linguists to the command line, and programmatic approaches to corpus analysis. Further examples of its syntax can be found throughout Chapter 9.

4.4.5. Open source development

A key limitation in previously available CL tools addressed by `corplib` is in the manner in which the tool is provided to the community. In terms of existing CL tool, it is notable that most do not provide the end-user with access to the source code. At the same time, many are maintained by a single developer. This creates a number of potential problems. First, the software can potentially become a *black box*, where researchers feed in data and interpret output, without understanding what has been done to their data, or how the output was extracted from the input. This can be more or less benign, as when a developer has simply not documented the workings of a feature. At worst, however, the tool may mask its own shortcomings, or remove results that may be important, but apparently uninteresting, or contrary to intuition. At the same time, in closed-source software deployment, software users can report, but not fix bugs themselves. This slows down the process of improving a tool, especially in cases where the software is single-authored. The second issue that arises in single-author, closed-source software development is what has been referred to

as the *bus factor*, which measures *how many people would need to be hit by a bus in order to make a project unable to proceed?* (Cosentino, Izquierdo, & Cabot, 2015). For many users, switching tools takes significant effort, and involves a learning curve. Making this investment for a tool with an uncertain future may be undesirable. `corplib` addresses these problems by being completely open-source. Users can not only report bugs, but also fix them. Users may also extend the tool to meet their needs, and request that a developed extension be merged into the master version of the tool. The tool can therefore respond to community needs in a more organic way.

4.4.6. Contributions of the toolkit

The toolkit is a contribution to the CL and SFL communities, as well as the digital humanities more broadly. It expands on the functionality of many currently available tools, allowing more sophisticated kinds of querying and manipulation than other tools to date. As an example, by allowing the user to forgo reference corpora and stopword lists, its approach to keywording resolves tensions between theoretical orientation of critical/functional linguistics (which tend to problematise the notion of balanced/representative texts) and keywording (which typically relies on frequency data drawn from reference corpora). `corplib` also brings tasks that have typically fallen under the domain of computational linguistics (e.g. lemmatisation, normalisation, parsing, linear regression modelling) to researchers interested in discourse. Despite the notable opposition to lemmatisation and parsing by e.g., Sinclair (2004), such perspectives are now uncommon, with increasing recognition that computational methods can enhance the nuance with which the lexicogrammar of texts can be probed, and thus the certainty with which claims about discourse-semantics of texts in corpora can be made. Looking at the tool from a computational linguistic perspective, it serves to bring state-of-the-art developments to a new user base, including corpus linguists and discourse analysts. This addresses a problem articulated by De Marneffe and Manning:

A major problem for the natural language processing (NLP) community is how to make the very impressive and practical technology which has been developed over the last two decades approachable to and usable by everyone who has text understanding need. . . . That is, usable not only by computational linguists,

but also by the computer science community more generally and by all sorts of information professionals including biologists, medical researchers, political scientists, law firms, business and market analysts, etc. . . . [T]he availability of high quality, easy-to-use (and preferably free) tools is essential for driving broader use of NLP tools (2008, p. 1).

Finally, `corppkit` is aligned with the emerging areas of digital and programmatic humanities research. It increases transparency and reproducibility of methods and findings: **Jupyter Notebooks** can be kept under version control and easily shared with others; the graphical interface stores interrogations and edited results in memory during each session, and allows for saving and loading of generated data between sessions. In being free, open-source and publicly available, it can be adapted by other researchers, simplifying the process of extending previous research and generating new theory.

4.5. Approach to data analysis

The process of data analysis follows on from the affordances of the developed tools, and from the output of the parsing pipeline: each subcorpus is interrogated via both constituency and dependency parses; results are edited, sorted or merged where necessary to translate findings into meaningful systemic-functional terms (insofar as is possible). These findings are then visualised or presented as tables; key points of interest then become the focus of further interrogations. Concordancing and plain-text examples is used where necessary to better understand a given lexicogrammatical phenomenon in co-text. As the key area of interest is longitudinal change, linear regression is performed on many interrogation results to calculate trends. This is followed by sorting of the entries by slope, in order to unearth parts of the lexicogrammar that become more or less frequent over time, are turbulent, or remain static.

The approach to analysis can be characterised as *systematic, exploratory and recursive*:

Systematic progression through data

Interrogation progresses along the cline of instantiation, from broad/shallow features to more specific lexicogrammatical realisations. Where possible, investigation of MOOD and TRANSITIVITY features are also interrogated separately. This ‘from-above’ approach is unusual in automatic analysis of text (Matthiessen et al., 2010); it is made possible by the combination of full parsing, which gives access to broad features, and the use of theory, which allows targeting expected linguistic sites of change.

Exploratory components

Given the fact that `corpkit` allows rapid interrogation of the dataset, testing hypotheses is often trivial. As such, features of the lexicogrammar that are not expected to be particularly salient (change in determiner use over time, for example) can be tabulated nonetheless. Much of the work of the analysis consists of iteratively exploring the corpus, guided by the findings of previous investigations: If past tense appears to be a key feature undergoing longitudinal change, a follow-up investigation may count Predicators in past tensed clauses, or look for keywords in these clauses’ Adjuncts.

Recursivity

Programmatic methods simplify and expedite the process of querying large datasets. As such, a great deal of recursivity is possible. For example, code can be written that locates key heads of Participants in the corpus, searches for processes in which these words are Participants, and so forth. This kind of interrogation was occasionally performed, and is for the most part documented in the accompanying Notebook, rather than in the findings section, as description of methodological steps is unwieldy, compared to simply reviewing/rerunning the code.

4.5.1. IPython & the Jupyter Notebook

IPython is an extension of the Python programming language, designed to allow quick access to Shell commands, to store code output to file and to easily access the output of previous commands (Pérez & Granger, 2007). **Jupyter Notebooks** are web browser based displays of **IPython** code, as well as code output, headings, text, and multimedia. **Jupyter Notebooks** are increasingly popular within scientific communities, as they allow code to be contextualised and explained multimodally, and can be easily run by other researchers, ensuring reproducibility of results. Much of the corpus investigation took place via **IPython** and **Jupyter Notebooks**. Many key findings from the investigation in this thesis are therefore available in Notebooks available at: <https://github.com/interrogator/thesis>. This Notebook presents methodological processes and code in more depth, and can be used to manipulate the corpus, and edit findings and results.

4.5.2. Limitations

The methodology has some limitations. First, it is bound by the speed and accuracy of its dependencies, with shortcomings of other tools inherited into every investigation. Second, it is strongly oriented toward quantitative insights into corpus texts, at times homogenising data prior to analysis. This means that bottom-up, grounded-theory-based interpretation of the dataset is not always possible. As such, the emergence of theory from data is perhaps less likely. The methodology also relies on ad-hoc translation between three grammars of English—constituency, dependency and the SFG. This means that linguistic notions may be simplified, and categories conflated when distinctions are not shared by all grammars. A further consequence is that it is not possible to engage deeply with concepts in SFL/SFG that make it a particularly attractive framework for locating discourse-semantic phenomena at risk in a structured corpus.

Second, the methodology is also based on relatively simplistic statistical methods. Though the output of interrogations is amenable to complex statistical analysis, here, most results are generated via simple relative frequency or keyness calculation, ordered using least-squares regression. This precludes, for instance, predictive ap-

plications of the methods, which could reveal which members are likely to progress to veteran stages, or to drop out.

Finally, the methodology is limited to querying the annotated data, and manipulating results. Though strategies for querying the data are more complex than is typical of CL/CADS research, the methodology does not harness a number of emerging methods in computational linguistics for categorising and understanding the content and context of digitised texts. Machine learning approaches to text analysis, for example, allow unsupervised classification of documents based on automatically determined features that theoretically motivated human researchers may not consider. The need for later research to apply state-of-the-art computational techniques is described in Chapter 9.

4.6. Situating the thesis methodology

The methods outlined here incorporate components from a number of convergent fields of study, such as text mining, corpus linguistics, computational linguistics, and the digital humanities. There is no clear line delineating where one of these fields ends and another begins. There is, however, a tendency for digital humanities and CL to focus on interpreting specific datasets, while text mining and computational linguistics are primarily interested in the development of tools. A mixed emphasis on methodological innovation and data analysis means that the case study straddles a broad interdisciplinary space. For discourse analysts and corpus linguists, the methods present new ways to engage with well-established computational linguistic practices and tools, allowing automation of work that has previously been possible only through manual analysis. At the same time, text mining and computational linguistics are often concerned with extracting entities and entity relationships, with little consideration of grammatical metaphor, or theories of language and grammar more generally. The developed method thus aims to combine the speed and automation of text mining approaches with the grammatical sensitivity seen in qualitative research. This opens up future research agendas that integrate discourse analysis and state-of-the-art computational linguistic methods.

4.7. Chapter summary

In this chapter, I have outlined and justified the elements of the data collection, approach to analysis and tool development. The following chapters operationalise the methodology in order to investigate language features in the Bipolar Forum. The analyses is followed by a discussion of findings in Chapter 8.

5. An introduction to the data: generic features of the Forum

In the previous chapters, I described the context, theoretical background and potential approaches of a case study designed to understand longitudinal change in language use in *Bipolar Forum*, a bipolar disorder OSG. In this and the next two chapters, I present the finding of the case study. The Bipolar Forum Corpus, and its Membership Stage Structure representing ten stages of membership, is the main dataset used. This chapter provides an introduction to the corpus from both qualitative and quantitative perspectives. First, I present a genre analysis of a small selection of posts, which highlights text structure beyond the level of the clause, and familiarises the reader with the kinds of language in the corpus. This is followed by a presentation of shallow quantitative features in the corpus' ten subcorpora. Chapter 6 contains findings from the MOOD and MODALITY analysis. Chapter 7 contains findings from the TRANSITIVITY analysis.

The case study was data-driven and exploratory: results from one interrogation could largely determine what next needed to be explored. At other times, it was driven by the results of prior studies of language use in OSGs, in communities generally, and in healthcare institutional settings. It was also theoretically informed, based on relationships between lexicogrammar and meaning provided by SFL. In this and the following two chapters, therefore, findings are generally accompanied by a brief discussion of their discourse-semantic significance. There are two reasons for this structural choice. First, because the analysis was exploratory, with more delicate querying following broader querying, discourse-semantics needed to be presented *in situ* in order to justify movement from one component of the lexicogrammar

to the next. Analysis of discourse-semantics was an integral part of the research process as it unfolded, rather than a task to be undertaken after the investigation had concluded. Second, the co-presentation of wordings and their meanings more closely models the way humans use and understand language: it makes little sense to keep discussion of meaning separate from discussion of wording, as readers naturally interpret the meaning behind wordings as they encounter them.

Accompanying these chapters is a series of `Jupyter Notebooks`, containing the figures and concordance lines presented here, as well as the code used to generate them. They also document the functionality of `corpkit`, the toolkit used for analysis, in more detail, provide further examples from the auxiliary corpora, and document raw findings (i.e. frequency tables, etc.) that cannot be included in this chapter for reasons of space. The Notebooks are designed to facilitate reproducibility and transparency, and allow easy future applications of methods developed here to new datasets. They can be accessed via GitHub at <https://github.com/interrogator/thesis>.

5.1. Genre and first contributions

Though the case study analysis is for the most part a quantitative, computational one, the analysis begins with a qualitative, genre-oriented analysis of some individual contributions to the Forum. I unpack generic features of a small selection of first posts and replies they receive, using the approach to genre analysis outlined by Eggins and Slade (2004) and reviewed in Section 3.3. Foregrounding the qualitative analysis serves two main purposes. First, it provides a context for the abstracted lexicogrammatical findings that follow: before analysing a corpus, it makes sense for the researcher to have familiarised him/herself with some of its texts. Second, it foregrounds limitations inherent to the automated, computational searching of the corpus. Contributions to the Forum follow generic constraints—these, however, are not considered during automated searching. Later chapters discuss this problem, and potential solutions, in greater detail.

One particular thread, entitled ‘**am i bipolar and what should i do?**’ is analysed in more detail than others. It is comprised of a new user’s first post, and replies from

existing members. It was chosen because it contained posts from users spread across the ten stages of membership, and because it contained discussion of key social actors in the Field of discourse, such as friends and family and health professionals. Based on simple reading of Forum threads, it also appeared to be relatively typical in terms of length, content and participating users.²⁸

5.1.1. A user's first contribution to the Forum

On May 30, 2011, jessff1989787, a 20 year old female from England, posted a new thread, '**am i bipolar and what should i do?**', to the board:

1 hi im jess new to this site, umm... well im currently 20 years
2 old and have been diagnosed with depression from a young age
3 but in the last 5 years or so i have been feeling very odd having
4 some extreme highs which include loss of appetite concentration
5 using drugs and alcohol spending sprees and also sex, i can not
6 control this when i get impulses like the above it is impossible
7 though i have tried. I also suffer with depression which is quite
8 severe most of the time i find it hard to get out of bed or to
9 even be able to connect with anyone including my partner who
10 lives with me, i am hurting him so much but i dont feel like i
11 can do anytjing about it
12 i have asked my doctor to test me to see if i am bipolar as my
13 antidepressants do not work even though they have been changed a
14 million times!! he said no that he wont test me and i also asked
15 for counselling and he also declined that, at the moment i feel
16 that im loosing control of everything and its getting worse, i
17 want to change my doctor and have been telling my partner i would
18 but im scared of finding out that i am bipolar, but i really feel
19 like either an extreme high or an extreme low is on the way and
20 im quite scared i dont know what to do
21 please help

The post is reminiscent in form and function to posts analysed in other work on OSG: Jess indicates a need for general social support (*im quite scared*, line 20) and requests advice (*what should i do?*, title; *i dont know what to do*, line 21). As noted by Smithson et al. (2012) and Varga and Paulus (2014), in order to legitimate herself as a potential member, and add perlocutionary force to her request, Jess also offers a narrative designed to demonstrate that she likely suffers from bipolar disorder, and thus fulfills the membership criteria of the board. Also notable is the sense of urgency, which has been conveyed through an apparent lack of planning within the post, spelling errors (*anytjing*, line 11; *loosing control*, line 17), and the pervasive joining of clauses through conjunction. As noted by Horne and Wiggins (2009), this may be a strategy for increasing the likelihood of response. In the context of bipolar

disorder, it could also connote the onset of a manic episode, which would serve to bolster the claim to membership.

Generic stages in Jess first post

The post appears to borrow from both SELF-INTRODUCTION and NARRATIVE genres identified by Labov and Waletzky (1967). Indeed, in many ways, the post conforms to the NARRATIVE genre, which has the structure:

(Abstract) ^ Orientation ^ Complication ^ Evaluation ^ Resolution ^ (Coda)

where

- ABSTRACT is an encapsulation of the point of the story
- ORIENTATION orients the listener to the circumstances of the story
- COMPLICATION is temporally sequenced event which culminates in a problem
- EVALUATION is the attitude of the speaker toward the COMPLICATION
- RESOLUTION is how the story's protagonist resolved the COMPLICATION
- CODA makes a point about the text and may reorient the listener to the present (Labov & Waletzky, 1967, p. 32).

As shown in Table 5.1, the ABSTRACT stage is realised by the post's title; ORIENTATION, COMPLICATION and EVALUATION do follow, though they appear to be recursive. Also different is the focus of the EVALUATION: here, it refers to an EVALUATION of the previous complication, rather than an evaluation of the narrative itself. The most significant deviation is that rather than a RESOLUTION (and optional CODA), there is a REQUEST to other members for 'help', presumably with the questions posted in the title: whether or not she is bipolar, and how she should respond to her presented self-evaluation.

Within a possible 'first-post genre', explicit REQUESTS are potentially optional (Vayreda & Antaki, 2009), though other members may interpret descriptions of problems and the act of posting themselves as warranting the provision of advice (Goldsmit, 2000). Some researchers (e.g. Herring, 1996b; Weber, 2011) have noted that users may not begin introductions to online communities with an explicit SALUTATION: as all messages are attributed to the writer's username multimodally, and even new users are likely familiar with the way in which the site presents posts, SALUTATION, if not rendered explicitly in prose, is in some sense embedded within

Genre stage	Sentences in Jess's first post
ABSTRACT	am i bipolar and what should i do?
SALUTATION	hi im jess
ORIENTATION	(I am) new to this site, umm... well im currently 20 years old and (I) have been diagnosed with depression from a young age but
COMPLICATION (1)	in the last 5 years or so i have been feeling very odd, (I have been) having some extreme highs which include loss of appetite concentration using drugs and alcohol spending sprees and also sex,
EVALUATION (1)	i can not control this when i get impulses like the above it is impossible though i have tried
COMPLICATION (2)	I also suffer with depression which is quite severe most of the time i find it hard to get out of bed or to even be able to connect with anyone including my partner who lives with me,
EVALUATION (2)	i am hurting him so much but i dont feel like i can do anything about it
COMPLICATION (3)	i have asked my doctor to test me to see if i am bipolar as my antidepressants do not work even though they have been changed a million times!! he said no that he wont test me and i also asked for counselling and he also declined that,
EVALUATION (3)	at the moment i feel that im loosing control of everything and its getting worse,
COMPLICATION (4)	i want to change my doctor and (I) have been telling my partner i would but im scared of finding out that i am bipolar, but
EVALUATION (4)	i really feel like either an extreme high or an extreme low is on the way and im quite scared i dont know what to do
REQUEST	please help

Table 5.1: Genre stages in Jess's post

the architecture of the forum mode itself. Given that all posts must be titled, and that titles almost always summarise the content of the post, ABSTRACT is an obligatory stage. The generic structure provided by Labov and Waletzky (1967) could thus be adapted for this case as:

Abstract ^ Salutation ^ Orientation ^ [Complication ^ Evaluation]^n ^ (Request)

To characterise the extent to which Jess's post was representative of the genre, the two first posts with the smallest wordcount were located and a basic genre stage analysis was performed (Table 5.2). It was assumed that small first posts would contain only obligatory genre features. These short posts indicate that EVALUATION is an optional stage, and confirm that recursion of COMPLICATION and EVALUATION is optional (though perhaps a key feature in developing a sense of urgency). *Short Post B* shows that the optional CODA noted by Labov and Waletzky (1967) appears to

be possible within the first-post genre. This leaves a finalised generic structure for thread-initial first posts within the Bipolar Forum:

Abstract^Salutation^Orientation^ [Complication^ (Evaluation)]^n^ (Request)^ (Coda)

Genre stage	Sentences in Short Post A	Sentences in Short Post B
ABSTRACT	<i>Lamictal</i>	<i>Depakote and hair thinning - any suggestions?</i>
SALUTATION	Hi-	Hi
ORIENTATION	I'm on my 3rd day of Lamictal and	My daughter has been on Depakote for 4 months and
COMPLICATION	it's giving me weird almost headache like pains in my head	(she) has experienced some (h)air thinning
EVALUATION		
REQUEST	Has anyone else had this happen while taking lamictal and when does it go away?	Has anyone out there found an effective way to counteract this problem?
CODA		She is on 500mg per day. Any suggestions really appreciated.

Table 5.2: Genre stages in short first contributions

It is not enough, however, to simply look at the features of the post: as Eggins and Slade (2004) explain, perhaps the best indicator of genre conformity is the way in which other members respond to the presented text. Thus, at this point, the investigation turns to consider replies to Jess's first post.

5.1.2. Replies to a first post

Jess's first contribution received six replies over the course of seven hours, from members with post counts ranging from five to over 5,000. Due to limitations of scope, only two of these replies have been selected for further analysis. They were selected based on the post count of the writer at the time of posting: one is a newer user (Subcorpus 03) and the other is a very senior member (Subcorpus 10). They also bookend the replies: the newer user is the first to respond to Jess, and the veteran member the last. The qualitative analysis, therefore, provides a description of language use at multiple stages of the membership course. The complete thread, including the other four replies, is available in Appendix A.

The first reply was a response from *Luvssoccer*, who had a total of five posts to the board:

1 It sounds like you might have bipolar to me. You need to change
2 Drs. One with more knowledge apparently. The reason the
3 antidepessants are not Working is because If you are bi polar
4 and they put U on an antidepressant alone it can make things
5 worse... I know from experience. Don't be scared it's treatable.
6 Find you a dr that can make a correct Diagnosis and go from there.
7 Good luck!

Following *Luvssoccer* were four replies from other members at differing membership stages, unanalysed here. The final post in the thread was by *Emz45*, who at the time of data collection had authored 5071 posts:

1 Hi, welcome to the boards, hopefully we can help you out and be
2 a support system for you. First off you're not A bipolar, *l*
3 we're not things, it's a condition. From what you say, if
4 sounds very likely that you might have Bipolar disorder. I would
5 go to a psychiatrist for testing and find out. You don't have to
6 have your docs permission to do this. If you're already seeing a
7 psychiatrist and that's who's doing all the denying, then find a
8 different one, because he's not doing his job, nor is he
9 considering your best interest. That's all that you can really
10 do in the beginning, find out what's what. We aren't docs here
11 and can't diagnose you. But I think it would definitely be smart
12 to go and get a diagnoses.
13
14 Take care, and please keep in touch, let us know how you're
15 doing, okay?
16
17 Emz

The original poster did not contribute to the thread again, but posted to the board on three other occasions in the next week, interacting again with *Emz45*.

Analysis of replies

A first notable feature of the replies is their functional similarity. Both *Luvssoccer* and *Emz45* offer social support (*Don't be scared*, line 5; *take care*, line 14), and encourage Jess to find a new doctor. Both also hint at a likely diagnosis based on the symptoms she has presented. In both cases, key ideological tenets of the Forum are represented delicately. Both disparage the work of Jess's current psychiatrist (*One with more knowledge*, line 2; *he's not doing his job*, line 8). Both navigate the Forum's conflict between an inability to provide official diagnoses and an orientation toward supporting those who appear to have bipolar disorder. In fact, both provide near-identical wordings:

1. *It sounds like you might have bipolar to me*
2. *From what you say, it sounds very likely that you might have Bipolar disorder*

Both statements essentially amount to a lay diagnoses, with modulation, modalisation, embedding and dummy Subjects used to reduce certainty and to avoid attribution of the diagnosis to the speaker herself. Both further hedge the lay diagnosis by stressing the need for a professional diagnosis, and, in doing so, deny that what they offered was any kind of diagnosis at all.

The relationship between first-post and reply differ across the hierarchy of stratification. At the stratum of genre, unlike first posts, the two replies do not conform rigourously to an identifiable sequence of clause functions. At the same time, the fact that the two responses overlap in content may be a useful indication that first posts do constitute a genre that is recognised by other members of the community. In terms of register, while the dimension of Field remains largely consistent (aside from Emz's avoidance of the topic of specific medications), the most dramatic differences are within Tenor: Jess is a prototypical newcomer, describing a medical history and current problem, positioning herself as vulnerable, lacking agency, and unable to participate effectively in her own care. In contrast, to Jess's descriptions of the past and present, Luvsoccer and Emz offer explanations and potential actions for the future.

Within the lexicogrammar, we can see differences in how the first and non-first-posts construe reality via the system of TRANSITIVITY. All participants position themselves as *Sensers*, but the kinds of mental processes vary in their representation of subjectivity, reasoning and control. Jess characterised herself through subjective mental processes over which she has no agency (*i have been feeling very odd, i dont feel like i can do anytjing about it, i feel that im loosing control of everything, i really feel like either an extreme high or an extreme low is on the way*); Luvsoccer, in contrast, *know[s] from experience*, while Emz *thinks* that diagnosis is the most important next step. Jess construes the health professional as an Agent, and most often, a Sayer (*He said no; He also declined that*), whose actions negatively affect her ability to obtain needed treatment. This contrasts with the other members, who reformulate health professionals as Goals (*You need to change Drs; Find you a dr; I would go to a psychiatrist*

(*for testing, etc.*) Jess positions her partner, on the other hand, only as a Goal or Target, and never a Participant that puts processes in motion (*i find it hard to [...] to connect with anyone including my partner; i am hurting him so much; I ... have been telling my partner*). For this reason, the partner is not construed at all in either response.

There are also differences in the way first and non-first posts instantiate the system of MOOD. All of Jess's major Mood choices are declarative and congruently work to provide information, until a final imperative, *please help*, which shifts the function of the post from the recursive narrative stages of ORIENTATION, COMPLICATION and EVALUATION, making a modulated demand on readers to address her declared fear and need for information (*im quite scared i dont know what to do*). The Subject of most of Jess's clauses is *I*, emphasising the self as the one invested with modal responsibility, and thus, the one who will honour any advice that others choose to provide. Modalisation of the self as Subject is used to stress an inability to modify behaviour (*I can not control this, I don't feel like I can do anything about it*). This is in contrast to the responses, which use interrogative and imperative Moods to request further information and to command the new user to respond. You is by far the the most common Subject in the two replies, as the responses keep modal responsibility on Jess herself. Modalised declarative choices, meanwhile, do not always provide information, but may also issue directives in the form of advice: Emz does this twice, casting herself as the hypothetical actor in the first (*I would go to a psychiatrist for testing and find out*) and using rank-shifted and non-rank-shifted modulation strategies to hedge the second (***I think*** it would ***definitely*** be smart to go and get a diagnoses).

Also observable are subtle differences between the two replies, caused primarily by membership stage. As mentioned earlier, Luvsoccer's post is a part of Subcorpus 03; Emz45's is in Subcorpus 10. First, in terms of choices of Field, Luvsoccer attempts to explain the reason that antidepressants are not helping, while Emz avoids the topic completely. This is related to the gradual rise in general offerings of social support and the decrease in references to specific medications (Wang et al., 2012): veteran users may refrain from offering specific advice on medication choice and dosage, as within a normative biomedical ideology, these are domains restricted to

the health professional (Vayreda & Antaki, 2009). Another key difference between the two replies is in the source of knowledge underlying the advice. Luvsoccer foregrounds personal experience by construing herself as the Senser (*It sounds to me; I know from experience*); Emz, on the other hand, only represents herself as a member of the board (*hopefully we can help you, we aren't docs and [we] can't diagnose you*), as the Subject within hypothetical advice (*I would go to a psychiatrist*), and within rank-shifted modulation of another advice instance (*I think it would definitely be smart to go and get a diagnoses*).

A third notable difference is in the ways the two replies are concluded. After offering advice, Luvsoccer simply writes *Good luck!*, effectively signalling the end of her part in the interaction. Emz, on the other hand, attempts to maintain dialogue (*please keep in touch, let us know how you're doing, okay?*) by commanding the user to report back, hedging through the use of a tag question seeking agreement/permission. Emz's conclusion opens up space for further exchange within a community where all interpersonal exchange is understood to contribute to wellbeing (c.f. Althoff et al., 2016).²⁹

Emz45's reply also shows us that veterans' preference for jargonisation (see Chapter 7) is not absolute. Here, presumably because she is interacting with a new user, she opts for lay terms (*bipolar disorder, psychiatrist, diagnoses*). This contrasts with Luvsoccer, who uses the developing shorthand forms for health professionals (*dr, drs*), but not the jargonised variants that distinguish between particular kinds of health professional.

A final important difference between the two replies is the way in which they advocate changing doctors/psychiatrists. Compare:

LUVSOCER

You need to change Drs. One with more knowledge apparently.

EMZ45

If you're already seeing a psychiatrist and that's who's doing all the denying, then find a different one, because he's not doing his job, nor is he considering your best interest

Though Emz is by far the more senior member, her advice is conditional, and sensitive to an ambiguity concerning the type of health professional Jess is currently seeing. Emz is also the only one who provides an explicit reason for the need to change. Luvsoccer's reasoning must be inferred from the statement that the doctor lacks knowledge, and has prescribed what she believes to be inappropriate medication.

The similarities and differences between the two replies highlight the way in which expertise and social status develop over the course of membership. It takes only a handful of prior contributions to the board to adopt the role of the expert when interacting with a newcomer. Imperatives and jargonised lexis are used to reinforce this role in earlier stages of membership. What may take time to develop are the subtle routines for providing information and support while maintaining an inclusive, non-hierarchical space. The main point of contrast between the two replies is that Luvsoccer's style is terse, lacking almost entirely in elaborations and pleasantries, while Emz disperses similar health information and advice within a text that is rich in incongruence and modulation, with the ultimate aim of establishing rapport. This phenomenon will be shown quantitatively in the following two chapters.

What remains, at this point, is to demonstrate the extent to which the thread analysed here is prototypical of the Forum's contents as a whole. More specifically, the remainder of this chapter, as well as the next two chapters, chart language choices quantitatively over ten stages of membership. It is useful to remember that Jess's first post is a part of the first subcorpus, Luvsoccer's reply is in the third, and Emz's in the 10th and final subcorpus, for all 560th posts or above.

5.2. Shallow lexicogrammatical features

The first quantitative part of the case study involves a short analysis of shallow features of the corpora, derived from subcorpus features such as word count, clause count, and the distribution of word classes/POS tags. Features analysed here are only very general approximates of register, crossing metafunctional lines and ignoring the entire dimension of lexis.

A secondary purpose for general frequency counting is for use in relative frequency calculation in the following two chapters. Rather than discovering the frequency of a given noun compared to all words in the dataset, it is more instructive to compare the noun to all nouns, or all nouns of a given class. In this way, analysis remains sensitive to broader grammatical changes that occur throughout membership: if texts become more highly nominalised, when comparing frequencies with all words in the dataset, a particular noun may seem to increase in frequency, when in reality, it is less and less often chosen ahead of other related nouns.

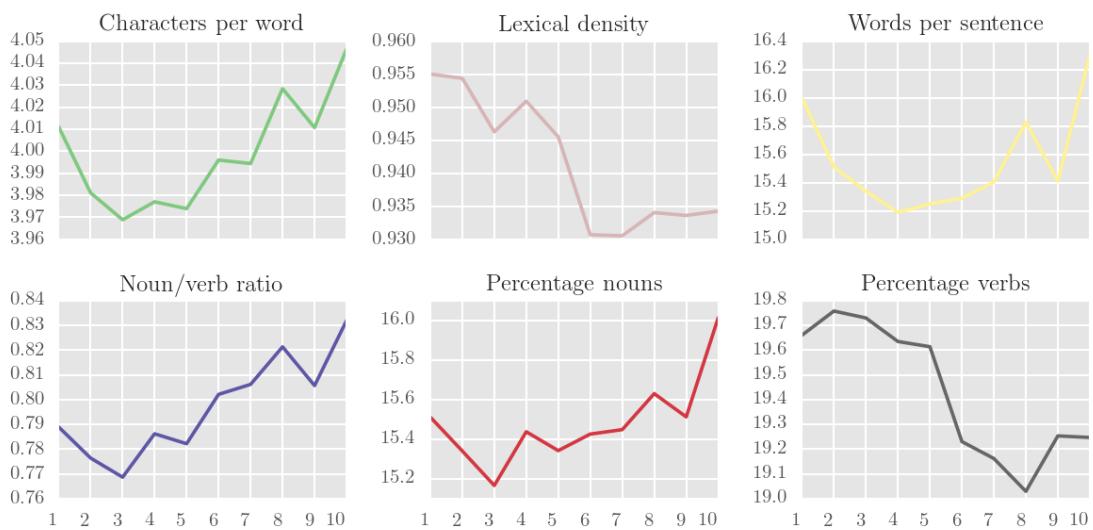


Figure 5.1: Derived shallow features for the Membership Stage Structure

A number of features exhibit relatively consistent longitudinal change (see Figure 5.1). Many of these features point toward an increasingly ‘scientised’ register (Harvey, 2012), such as nominalisation and word length, which are related, as nominalisation typically involves the addition of derivational morphemes to non-nominals (Simon-Vandenbergen, Ravelli, & Taverniers, 2003). In the same vein, verbs become less frequent over time. That said, *noun* and *verb* are formal categories that can only approximate semantic notions of *Things* and *Events*, or, at a higher rank, *Participants* and *Processes*. More attention will be paid to these in the chapters that follow.

Another trend visible in Figure 5.1 is that the first subcorpus—that is, first posts to the Forum—has features that often differ from the overall upward or downward trend. Lexical density (defined here as the ratio of lexical words to clauses) is one

example: peaking in first posts, it remains relatively stable thereafter. First posts also deviate from general trends in terms of average character per word, words per sentence, noun/verb ratio, and the relative frequency of nouns in general. These features indicate that first posts are considerably more packed with content than later contributions. Higher lexical density in first posts is to be expected: users are under no time constraints, as they are not responding to an interlocutor, and because other members are not aware that a message is being prepared. At the same time, users' initial contributions are more formal because it is in these contributions that new members first make interpersonal demands (solicitation of responses) on other members, who are by definition more senior figures in the community. Many users may also have longer-term considerations in mind: their first posts constitute an initial bid for acceptance by the community, and thus are tailored to conform to others' expectations of politeness. After the user is welcomed into the community, lexical density quickly stabilises, and many other features begin a stable trajectory. These are early quantitative indications of a difference between the first and non-first posts at the strata of *genre* and *register*. This discussion will be picked up in Chapter 8, following the presentation of findings from more delicate querying of the lexicogrammar in Chapters 6 and 7.

5.3. Controlling for self-selection bias

The Membership Stage Structure has the potential for self-selection bias: users who become veterans may use language differently even in their first contributions. To control for this, the Future Veteran Structure (which contains no posts from early dropouts) and the Comparative Structure (which has subcorpora for Dropouts and Future Veterans) can be used. In the section below, I briefly demonstrate that the language use of those users who progress to veteran stages is not quantitatively different from the language of users who drop out during early contributions. The implication of this finding is that future veteran members' language choices do in fact change in predictable ways over time.

The Future Veteran Structure contains no posts from members who dropped out before contributing 30 times. It is structured identically to the Membership Stage Structure. When using this corpus, the first four stages of membership are collapsed into a single stage, in order to make sure that each subcorpus contains a quantifiably reliable amount of text.

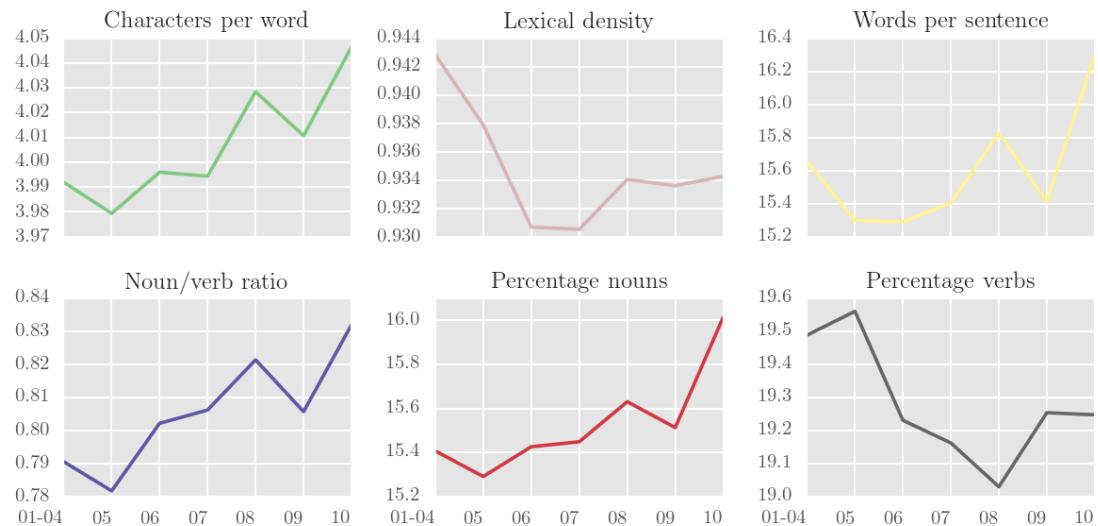


Figure 5.2: Derived shallow features for the Future Veteran Structure

There is generally little difference from the Membership Stage Structure (compare with Figure 5.1), suggesting that future veteran users use a similar register to non-future members in early posts.

The Comparative Structure has two subcorpora: the first contains posts from any member who posted fewer than 30 times; the second contains the posts of members who posted 30 or more times.

Figure 5.3 shows a great deal of similarity between the posts of Dropouts and the early posts of future Veterans. This means that broad register features change over the course of membership.

5.4. Mapping membership stages to Forum history

The Longitudinal Structure takes threads as the unit of analysis, rather than posts. Each thread is grouped by the year in which it was created, so that linguistic change

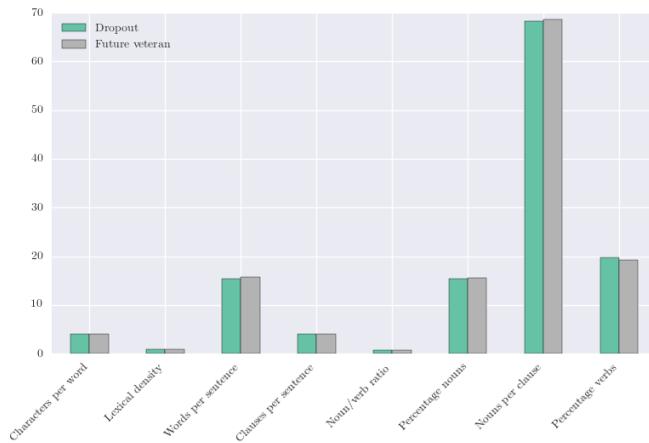


Figure 5.3: Derived shallow features for the Comparative Structure

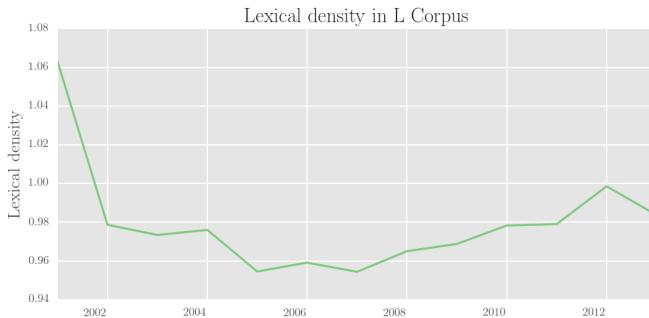


Figure 5.4: Lexical density in the Longitudinal Structure

across the Forum's history can be examined. One key feature exhibiting consistent change is lexical density (Figure 5.4), which steadily rises throughout much of the Forum's history—fluctuations observed in the first and last subcorpora could well be the result of chance (due to their smaller sample size—see Section 4.3.2). That said, it is not impossible that the first ever contributions to a forum adopt a slightly more formal tone, as users air on the side of (polite) caution until the normative linguistic practices of the community take shape. The finding also hints at phylogenetic change, where linguistic practices of one generation of users within the Forum can have a lasting effect on the practices of the next (Danescu-Niculescu-Mizil et al., 2013).

Though limitations of scope prevent a larger analysis of the alternative corpus structures, each helps to build a richer understanding of language use in the Forum. Lexical density in the Longitudinal Structure, for instance, provides preliminary

support for the notion that veteran users may have lasting change on the discursive orientation of an online community. Methodologically, the advantages brought by such alternative corpus structures is clear: the same data, differently organised, can be analysed with identical methods in order to answer a different set of research questions, or to shunt between analysis of different kinds of linguistic change.

5.5. Chapter summary

In this chapter, I have provided preliminary analysis of the Bipolar Forum Corpus from two complementary perspectives. First, I have approached three posts within a single thread from below, looking at realised words and wordings in sequence. This analysis highlighted differences in how Forum members use language over the course of membership, echoing in many respects existing findings in discourse-oriented OSG literature: newcomers' first posts may conform to a first-post genre in which medical histories and current problems are outlined, with the dual aim of both legitimating the self as a potential member and eliciting responses from readers. Second, I have looked at the corpus from above, looking at broad changes in shallow grammatical features, without consideration of the role of lexis, or of the distinction between metafunctions. Findings here showed that some features vary in more or less consistent ways over the membership course, and may be related to overall shifts in the register of users as they progress toward veteran status within the community. What remains, however, is to fill in the middle ground between realised samples of text and broad grammatical features. This analysis, segmented by metafunction into MOOD and TRANSITIVITY parts, is performed over the next two chapters.

6. MOOD and MODALITY choices in the Forum

With an overview of shallow linguistic features in each corpus, as well as a basic qualitative understanding of posts and their generic properties, the investigation can begin to shift toward individual grammatical systems and the lexical choices at the most delicate end of each system’s instantiative cline.

In this chapter, I present findings from an analysis of MOOD and MODALITY choices, which are responsible for making interpersonal meanings—that is, they are used to negotiate the roles and responsibilities of interactants. Frequencies for choices of Mood and Indicative Type are calculated first. This is followed by an account of modality, Mood Elements, and brief descriptions of TENSE and POLARITY systems.³⁰

6.1. Mood and Indicative Type

The broadest features of the MOOD system that is reliably annotated by constituency parsing are the distinctions between Indicative and Mood Type (see Section 3.3). Because dedicated tags for Mood Types are not available in either the constituency or dependency grammars, however, accurately locating them by automated searching alone is an unintuitive task. For this case study, *Tregex* expressions were developed, in order to discern major Mood and Indicative features from constituency parse trees (see Figure 6.1). These queries must not just be designed for ideal, well-formed cases, but must also handle false positives, such as sentence initial vocatives and salutations, which are often parsed as subject NPs. For interrogative matching, after many attempts to exploit Wh-pronouns, Subject–Finite order, it was determined

```

1 # salutations and exclamations that cause parser errors
2 badwords = ['-l.b-', 'hi', 'hey', 'hello', 'oh', 'wow',
3             'thankyou', 'thanks', 'welcome', 'thank']
4
5 # mood types as dict object
6 m = {'Mod. declarative':
7       r'ROOT < (S < (/NP|SBAR|VP)/ $+ (VP < MD))',
8       'Unmod. declarative':
9       r'ROOT < (S < (/NP|SBAR|VP)/ $+ (VP !< MD))',
10      'Interrogative':
11      r'ROOT << (/?/ !< __)',
12      'Imperative':
13      r'ROOT < (/S|SBAR)/ < (VP !< VBD !< VBG !$ NP !$ SBAR < NP
14      !$-- S !$-- VP !$ VP) !<< (/?/ !< __) !<<- /-R.B-/ !<<, /%s/' %
15      as_regex(badwords)}

```

Figure 6.1: Tregex queries for Major clause Mood Types

	Mod. declarative	Unmod. declarative	Interrogative	Imperative	Coverage
01	7.012	68.350	9.521	1.168	86.051
02	7.849	66.143	8.618	1.366	83.976
03	8.643	66.324	8.527	1.466	84.959
04	9.132	65.576	8.967	1.579	85.254
05	9.023	65.115	8.777	1.577	84.492
06	9.880	63.367	9.183	1.732	84.163
07	9.515	63.234	8.780	1.615	83.144
08	9.792	62.474	8.640	1.710	82.616
09	10.904	61.347	8.853	1.899	83.003
10	12.645	60.592	9.283	1.704	84.224

Table 6.1: Relative frequency of Mood types

that counting sentence final tokens containing at least one question mark was the most accurate method. When developing the constituency queries, accuracy was preferred over coverage. For this reason, only main clauses were analysed. To test the coverage of the queries, the total number of Mood Types found was compared to the total number of sentences (i.e. parse trees), minus any tree that did not contain a VP headed by a verb found in VerbNet. This removed a number of common, non-major clausal parse trees, such as those provided for a minor clause like *Hello!*.

The frequency of imperatives rises steadily through membership length (see Table 6.1 and Figure 6.2), as veteran members may explicitly direct new members to do certain things: *take care*, *find the right meds*, or *watch alcohol consumption*, for example, are common commands issued in veterans' posts. Though advice given

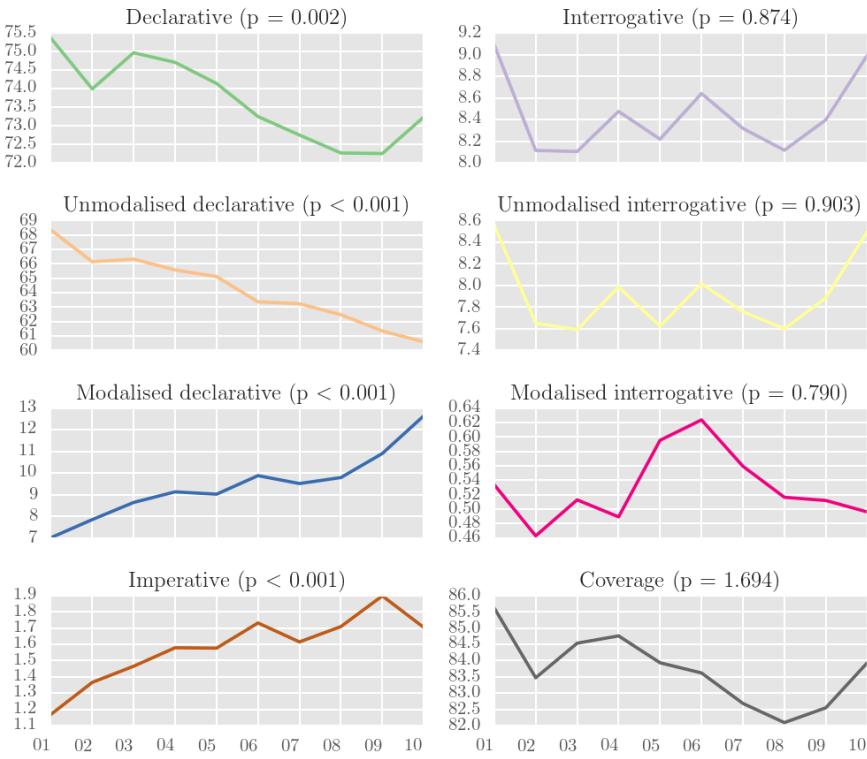


Figure 6.2: Mood features in the Membership Stage Structure

through imperatives provides suggested behaviour (as in the cases above), heavy grammatical constraints on the imperative Mood (lack of Subject, tense, etc.) make elaboration within this kind of advice uncommon. Modality is also difficult to encode, as modal Finites cannot grammatically modify imperative Predicators (**Could go to the doctor!*). Imperatives thus rarely carry information regarding the addressee's level of obligation or the speaker's level of certainty or source of knowledge, and accordingly, can be face-threatening for newcomers (Goldsmith, 2004; Hudson, 1990). For this reason, in veterans' posts, unmodulated imperatives often function as general markers of social support (*take care of yourself*; *keep up the good work!*) rather than as suggested behaviour coupled with health information or marking of the source of the knowledge. Thus, there is no one-to-one relationship between the extent to which Forum users make interactive demands on others and choices of Mood Type.

As shown in Table 6.1, declaratives shift over time, from unmodalised toward modalised type. This indicates a correlation between membership stage and the

extent to which users insert judgements of likelihood of propositions. This pattern can be interpreted as evidence of increasing social status over time. At the same time, however, the MODALITY system can be used as a resource for signalling incongruence between Speech Function and Mood Type (see the following section). Interrogatives, on the other hand, undergo no significant trajectory shift, as demonstrated in the significance calculations shown in Figure 6.2. Questions therefore comprise the same amount of talk at all stages of membership. As will be discussed below, while the overall frequency of the interrogative Mood remains stable, the kinds of propositions being negotiated in these questions undergo observable longitudinal change.

The calculation of *Coverage* (Table 6.1/Figure 6.2) shows that the proportion of sentences for which Mood Type information could be obtained fluctuated, but in no clear direction. This increases confidence that observed changes in Mood Type proportions are not the result of consistent language changes that cause an increase or decrease in parser accuracy.

6.2. Modalisation

Modalisation construes uncertainty within the clause as exchange. It is a more delicate feature of the MOOD system than Mood Type, because its meaning changes depending on Mood selection: in declaratives, it congruently involves a speaker judgement; in interrogatives, it demands judgement from the addressee (Halliday & Matthiessen, 2004). Different modal lemmata are responsible for communicating obligation, certainty, probability and usuality; at the same time, choices in modal lexis distinguish between low, median and high uncertainty, with negative and positive polarity at either pole (see Section 3.3.5: MODALITY).

Over the course of membership, in general, Modalisation as a feature of Major clauses increases consistently (from 6.67% to 8.83% of all clauses). The major Mood Type accommodating this increase is the modalised declarative (Figure 6.2). A primary cause of this change is a strategy for advice provision that develops over the membership course. Rather than simply issuing commands with imperatives,

- | | |
|---|---|
| <ol style="list-style-type: none"> 1. Let us know how your appointments go!! 2. Keep us posted with how things go. 3. Keep on posting and let us know how you are doing. 4. Try to pretend that you have a suit of armor on and try to allow as much of this to bounce off of you. 5. Get a good night's sleep, eat right eliminating caffeine & sugar from your diet, take your meds as prescribed, avoid stress as much as possible, get plenty of exercise and hold onto your sense of humor. | <ol style="list-style-type: none"> 1. I would certainly make a point to follow up 2. I would definitely have your daughter pay her 3. I would highly recommend it 4. I would DEFINITELY recommend seeing a psychologist 5. I would definitely make mention of this 6. I would strongly suggest that you discuss 7. I would be very careful with just the Zoloft 8. I would highly recommend that you take your meds on a daily basis. |
|---|---|

Figure 6.3: Advice provision via imperatives and modalised declaratives

much advice dispensed by veterans is realised by declaratives, generally featuring modalisation (see Figure 6.3 for examples).

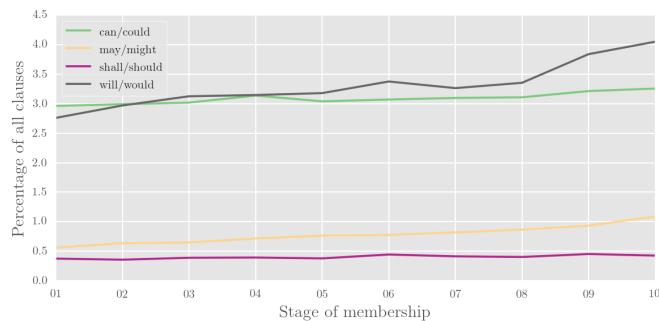


Figure 6.4: Relative frequencies of modal lemmata

Figure 6.4 shows that over the length of membership, there is a particularly marked increase in the frequency of *would/will* modals, in comparison to other modal lemmata. Concordancing of declaratives modalised by *would* was performed in order to understand their typical contexts of use. This revealed three things. First, veterans commonly dispense advice through hypothetical *I would* statements (as in the examples above, and as seen in the qualitative analysis of Emz' post in Section 5.1.2). Second, the modalised declarative is also often modulated by an Adjunct, in order to reconstitute perlocutionary force that was diminished through the incongruent selection of Mood Type (*I would seriously/really/certainly consider quitting*). Third, the same grammatical construction also exists in new member talk, but very rarely as a means of giving advice. Rather, *I would* (+ adjunct) in first posts often introduces

a construal of past behaviour (*I would get seriously drunk every night*—see Figure 6.6 for contextualised examples).

As a result of this ambiguity, all *I would + adjunct* declaratives in new and veteran members' posts (261/143 total matches) were coded according to five inductively developed functional categories (outlined in Table 6.2). Three false positives (where 'd as a contraction of *had* had been incorrectly parsed as a contracted *would* modal) were excluded from analysis. The categories are not codified in the SFG; rather, they are based on the apparent meaning potential of the register of Forum talk. Categories are also not intended to be discrete: since a core purpose of modalisation is to express inclination, most instances of *I would + adjunct* are on some level statements of inclination. Not all statements of inclination are descriptions of past behaviour, or provisions of advice, however.

Category	Description	Example
Past Behaviour	Habitual (generally negative) actions in the past	<i>And was around the time I would occasionally go to sleep for as much as 24 hours</i>
Advice	Suggesting what another should do	<i>I'd really encourage you to just call your psychologist back</i>
Request	Requesting actions (typically information/support provision) from others	<i>I'd really appreciate any feedback from anyone on here</i>
Inclination and hypothetical	Preferences and inclinations, often in irrealis scenarios (within if-clauses)	<i>I would never want to be without him; I wish I were sicker so that I wouldn't even worry about it</i>
Hedged salutation	Self introduction, explicit salutation	<i>I'm new here so I thought I'd just start out with a big ol' hi</i>
Past sense of will	Talk of the future within narratives about the past—usually in Verbiage	<i>My father always told me that I would never mount to anything</i>

Table 6.2: Thematic categories of *I would + adjunct*

Figure 6.5 shows that new and veteran members employ the *I would + adjunct* construction for different purposes. In newcomer talk, it is commonly used to make incongruently realised requests (*I would like to know ...*). This is a politeness strategy—veteran members seldom hedge their requests in this way. Also more common in newcomer talk are the uses of *I would + adjunct* to index *Past behaviour*, and as the *Past sense of will*, an example of which can be found in Jess's first post (*i want to change my doctor and have been telling my partner i would*—see Section 5.1.1). From the perspective of genre, as highlighted in the previous Chapter, initial contributions conforming to generic structures include a substantial medical history, told in the

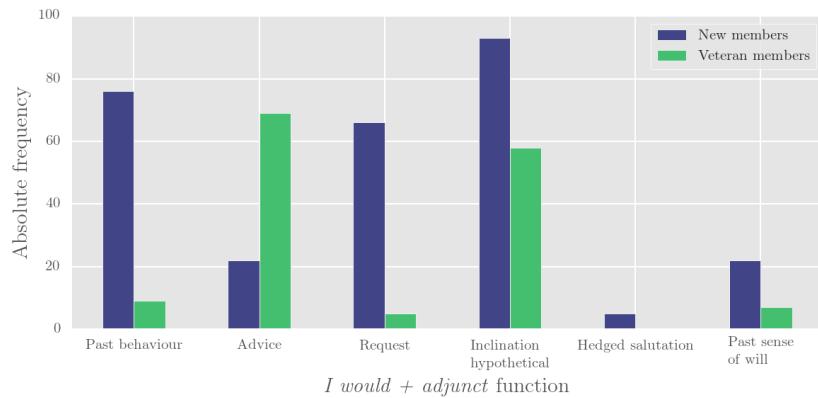


Figure 6.5: Functions of '*I would*' in new and veteran posts

past tense. For veterans, on the other hand, *I would + adjunct* is predominantly used to signal advice. Aside from the general category of *Inclination and hypothetical*, other functions are very uncommon in posts made during the final stage of membership.

Advice realised by an *I would* reveals much about the ways in which role relationships within the community are discursively constructed. In contrast to health professionals' advice, *I would* puts the veteran hypothetically in the position of the newcomer, highlighting their shared role as people living with bipolar disorder. At the same time, however, it stakes claim to higher social status: the construction implies not just a burden on the addressee to follow the advice, but also some more subtle normative social/ideological values—that veterans know more, and that the course of action they would personally take in such a situation is the right one.

An experiential analysis of the *I would* style of advice provision complements the interpersonal perspective here. The advice giver is construing him/herself experientially as the main participant in the figure, substituting the newcomer out of the construed reality entirely. This main participant is almost always an Agentive one: the most common processes include *suggesting*, *talking*, and *going*. To emphasise agency further, veteran members may embed advice within a material process of *going* (see Table 6.3), especially when the addressee is to be the Medium in the presented figure (*I would go and get evaluated ...*). In turn, this emphasis on agency highlights for the addressee the need to take action in order to gain control over his/her health problems.

i wouldn't just	go	and quit ... perhaps it is time to look for a less stressful job
i'd	go	right away and ask for major help from your pdoc
i would most definitely	go	and be evaluated by a psychiatrist, that is absolutely the only way
n your shoes i would	go	and get a second opinion from a doctor you've not been to before.
i would	go	and get evaluated for bipolar if you think that is what you have
and then i would	go	talk with the pdoc and ask him why these meds .
but i would definitely	go	for that eval.

Table 6.3: Emphasising action in veterans' advice to newcomers

This foregrounding of subjective knowledge is very much at odds with Harvey's 2012 finding, where adolescents used a medicalised register to legitimate their claims. In the Bipolar Forum, legitimacy is claimed explicitly through lay experience. While newcomers typically enter the community with specific requests for information (about a side-effect of a medication, or a problem with insurance, for example), veterans demonstrate an understanding of the illness course as a whole, with advice often taking into account a 'bigger picture' that the newcomer is assumed to be not yet privy to.

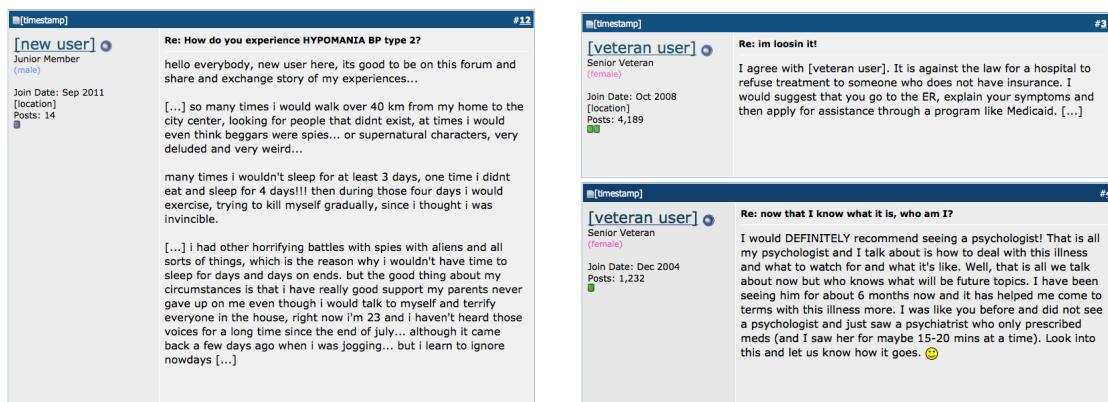


Figure 6.6: Examples of typical new and veteran members' 'I would' constructions

6.3. Mood elements

In SFL, the Finite grounds a proposition in time and space, making it *arguable*. Arguability can be centred on primary tense, where propositions are debated with respect to when they occurred, in relation to the time of speaking. Alternatively, Modality can be argued. The Subject, meanwhile, is the thing charged with ensuring

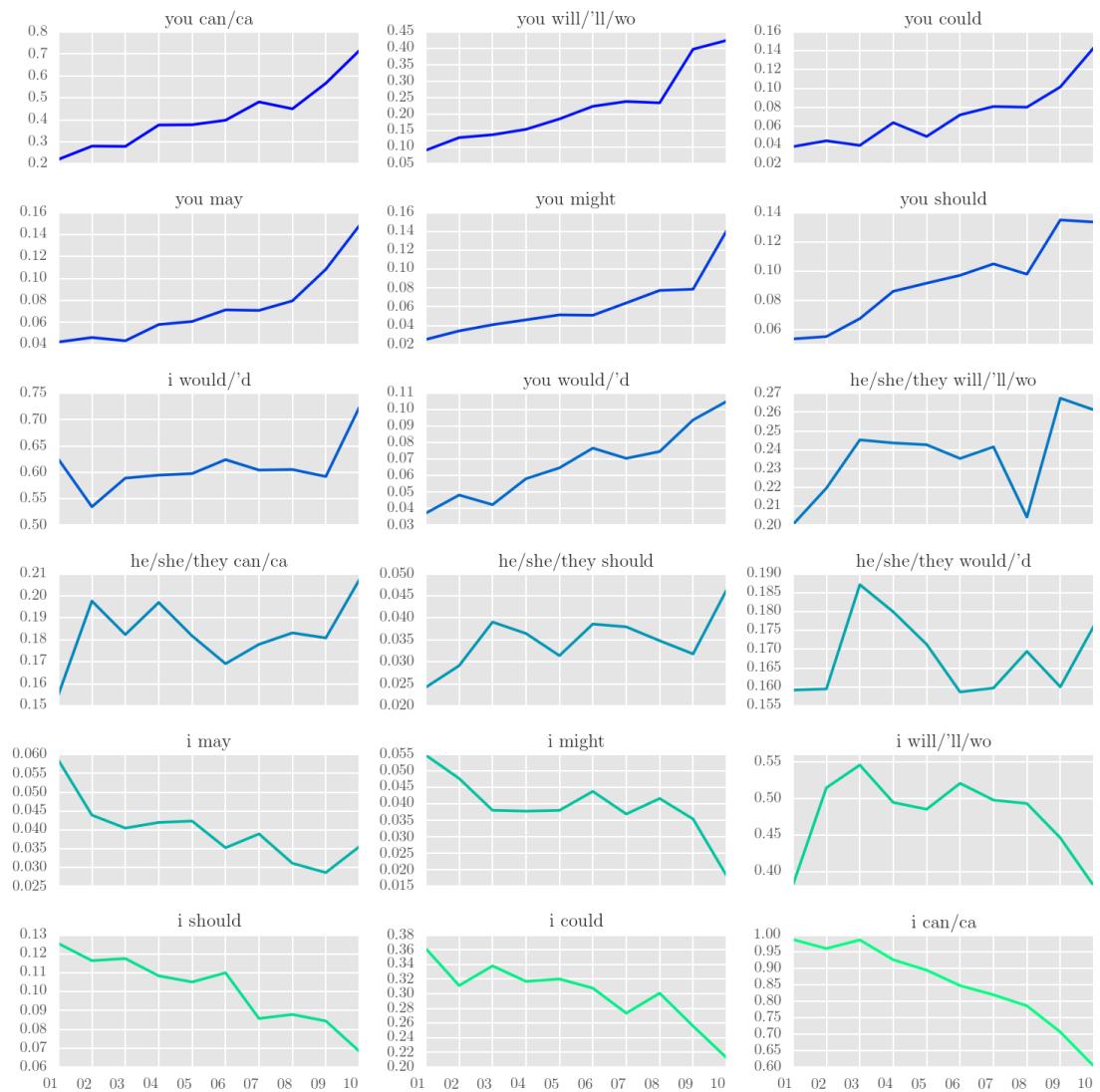


Figure 6.7: Pronoun-Subject + Modal-Finite blocks as a percentage of all clauses

the carrying out of the proposition. By interrogating the corpus for combinations of Subjects and Finites (see Eggins, 2004), it is possible to chart change in who is made responsible, in the clause-as-exchange.

The frequencies and trajectories of configurations of Pronominal subjects and modal Finites can be used to determine the kinds of proposals and propositions being debated. Figure 6.7 shows that generally, *I + modal* is displaced by *you + modal* over the course of membership. In fact, *I would* stands out as the only *first person + modal* block on an increasing trajectory. This is due to its frequent use as a means of giving

advice, as discussed above. Though modals negotiate a diverse range of concepts, such as certainty, probability and obligation, the *Subject + modal* configuration still places an interpersonal burden on the Subject as the one charged with bringing the Predicator about. Accordingly, newcomers self-assign the burden, while veteran members turn to direct interpersonal demands on their interlocutors. This is not to say, of course, that the veteran/newcomer relationship is one that mistreats the newcomer. The dynamic is a symbiotic one. Veteran membership comes with a kind of expertise, which is being exchanged with other community members through advice. Veterans answer others by explaining what is possible, feasible, likely or uncertain.

Mood Blocks can also be viewed through the lens of community membership conditions. In early posts, users narrate their medical history, setting out credentials for acceptance by the user base. Veteran members act as interviewers, requesting further information, or grafting the presented narratives to the community's normative biomedical ideology. Newcomers are under observation, and are expected to assist in veterans' assessments by providing a narrative account of their illness course. This is particularly important in first posts, but extends generally to cover early contributions, in which the new user is primarily focussed on the self.

Figure 6.7 also shows that the trajectory of *third person + modal* constructions is generally inconsistent or stable. This is because role-relationship negotiation is carried out by making or fulfilling the demands of interlocutors, rather than things and people being spoken about. There is therefore little difference over the course of membership in the way non-present participants are modalised. Rather, as we will see later, non-present participants undergo a number of changes within the system of TRANSITIVITY.

Of all constellations, *I + can* (and its derivatives and interrogative inversion) undergoes the most radical change. In a question, this kind of Mood Block expressly asks others for information or permission. As can be seen in Table 6.4, in first posts, in declarative form, *I can* is very often negated, with contributors narrating difficulties carrying out healthy mental processes (*cope, focus, make up mind, handle/stand it, stand it, remember own name, have no emotion, feel normal*), controlling harmful behavioural

processes (*can't stop crying, can barely stay awake, can't breathe*), and performing vague material processes (*get things done, nothing I can do about it, get out of everyday life*). Over the membership course, the construction falls into disuse: unlike *I would*, it is very rarely employed as a proposal (e.g. *I can PM you this if you like*).

0	antly stressed , i had migranes and	i could n't sleep , i could n't go out for to long be
1	ad migranes and i could n't sleep ,	i could n't go out for to long because i just always
2	s just started to get abit better ,	i could go out again
3	nd i did want to end it all because	i could n't take the fear anymore
4	i 'v also noticed that	i can get really angry , really easily these days too
5	- if anyone has any tips about how	i can lose weight and not gain anymore whilst on depa
6	st but i 've coped with it the best	i can
7	i 'm not working right now , but	i ca n't seem to get anything done around my house
8	t have tutors come to my house when	i could handle them
9	f out of bed in the morning and all	i can think about all day
10		i can not stop crying .
11		i can barely stay awake .
12	i feel so fat and disgusting , but	i ca n't stop eating
13	i just wish	i could make them feel how i 'm feeling , just so the
14	al and i feel like there 's nothing	i can do about it
15	erally just keep thinking about how	i can go back to the psych ward
16	go back to the psych ward , just so	i can get out of everyday life for a while
17		i can not get along with any of my coworkers , as wel
18		i ca n't process any of my rapid thoughts and have ju
19		i ca n't think of any reason why this occurs and do n

Table 6.4: Concordancing *I + can* Mood Blocks in first posts

This pattern can also be approached in terms of membership entry conditions. Newcomers often stress the seriousness or urgency of their case, based on the existence of some ongoing mental or physical harm.

6.3.1. Tense

If not by MODALITY, arguability happens along the lines of TENSE—the relationship between a proposition and the time of speaking. Figure 6.9 shows different configurations of pronominal Subject and Tense. More specifically, it shows that tense of modals is not a feature that exhibits consistent change over time: the trajectory shifts are caused for the most part by changing frequencies of pronominal Subject choice.

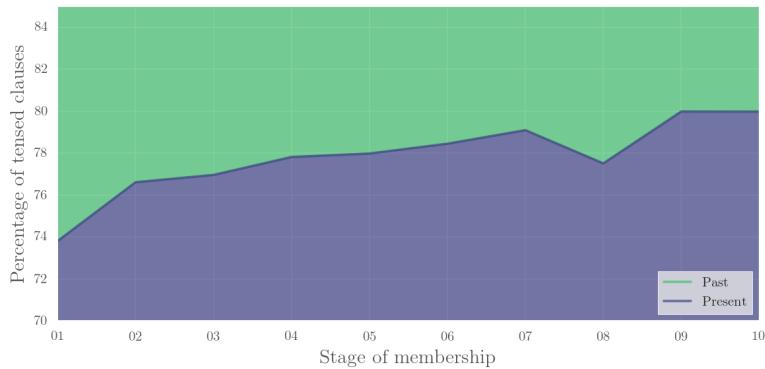


Figure 6.8: Tense of tensed clauses over the course of membership

The system of TENSE can be isolated from its co-text, however. Figure 6.8 plots primary tenses of non-modalised clauses over the membership course. The steady transition from past toward present (from 74 to 80 per cent of all tensed clauses) demonstrates a shifting focus from setting out medical history narratives as membership credentials to helping others with their current problems. At an ideological level, the shift also represents a gradual refocussing on present and future actions, both of the self and of addressees.

6.3.2. Polarity

The final component of the interpersonal analysis is the system of POLARITY.³¹ Clauses may have positive or negative polarity. The default polarity is positive; negative polarity is realised through a small group of lexical items (*n't/not, no, never*) that appear in close proximity, or fused with, the Finite. Therefore, POLARITY can be interrogated using Tregex queries:

```
Negative = 'VP !> VP <+(VP) (VP !< VP <<# (/VB.?/ < /$VERBLIST/)) \
           ' <<# /(VB.|MD)/ < (RB < /(?i)^n.t|never|no$/)''
Positive = 'VP !> VP <+(VP) (VP !< VP <<# (/VB.?/ < /$VERBLIST/)) \
           ' <<# /(VB.|MD)/ !< (RB < /(?i)^n.t|never|no$/)''
```

POLARITY figures into other places within interactions, such as as minimal responses to polar interrogatives. These kinds of minor clauses are not considered here; only clauses containing a verb-form in VerbNet are considered. Figure 6.10

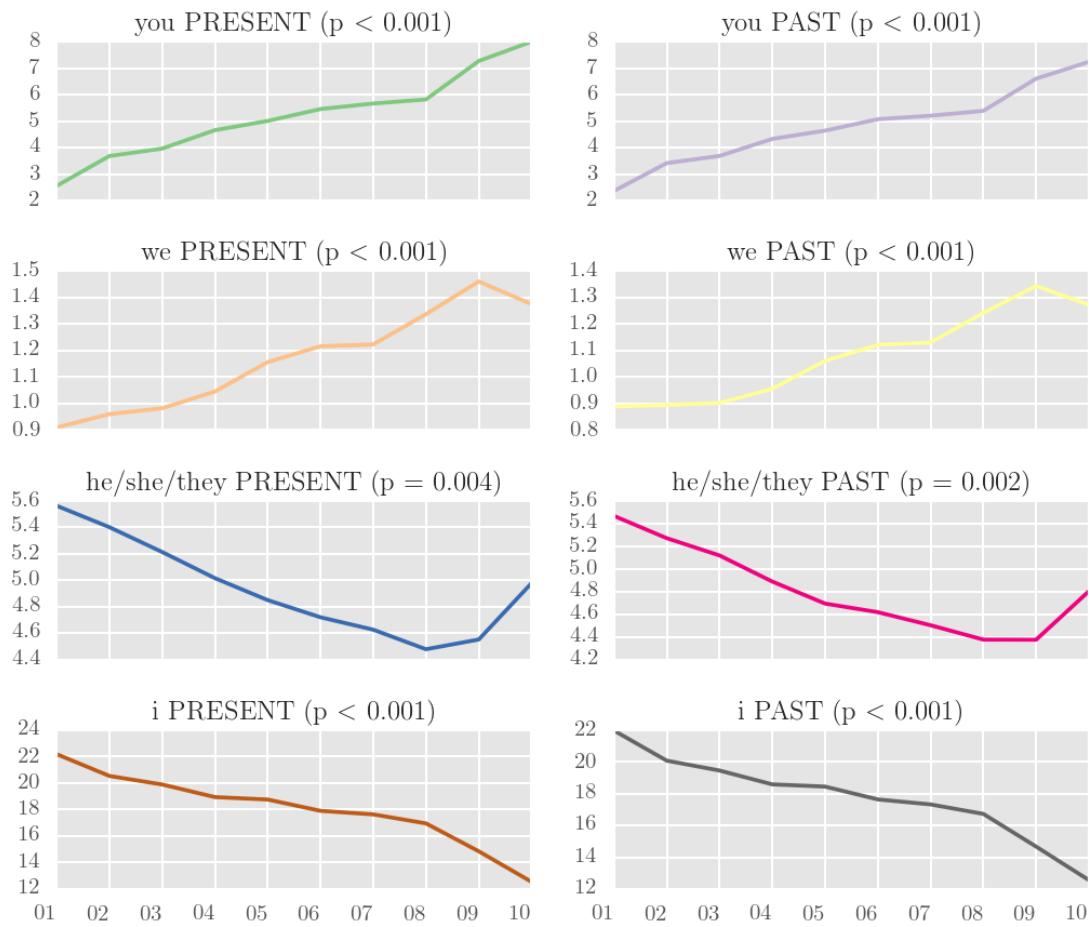


Figure 6.9: Pronoun–Subject + Past/Present tense blocks as a percentage of all Major clauses for the Membership Stage Structure

shows that clause-level polarity choices have an unstable trajectory over the membership course, decreasing until the 8–15th post, and then increasing. The fact that this result is difficult to interpret is unsurprising: the meaning of polarity is to some extent dependent on Mood Type, and on the previous move(s) in the exchange. Moreover, unlike a linguistic feature like passivisation or grammatical metaphor, the ratio of positive/negative polarity clauses does not have a direct or clear relationship with discourse–semantics within a clause, except perhaps in a few rare cases (i.e. a list of rules). As noted in Section 3.3.5, POLARITY is more likely to provide insights into discourse when examining entire interactions, rather than individual posts. Even so, one point of note is that negative polarity is approximately

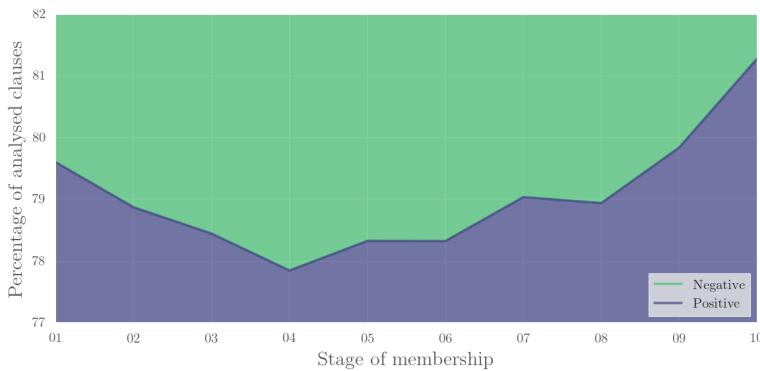


Figure 6.10: Clause polarity over the course of membership

twice as frequent as the ratio suggested by ten per cent of all clauses—see, Halliday and Matthiessen (2004).

It is also possible to search for polarity in combination with other Mood features, in order to determine whether combinations of Modality + Polarity, or Pronominal Subject + Polarity, are undergoing change. In Figure 6.11, modals are grouped into four types. Distinctions are also made between first and second person Subjects, and between positive and negative polarity. Again, this kind of analysis shows, when compared to choices of Subject and modal, polarity choices play only a minor role in the longitudinal register shift of members of the Bipolar Forum.

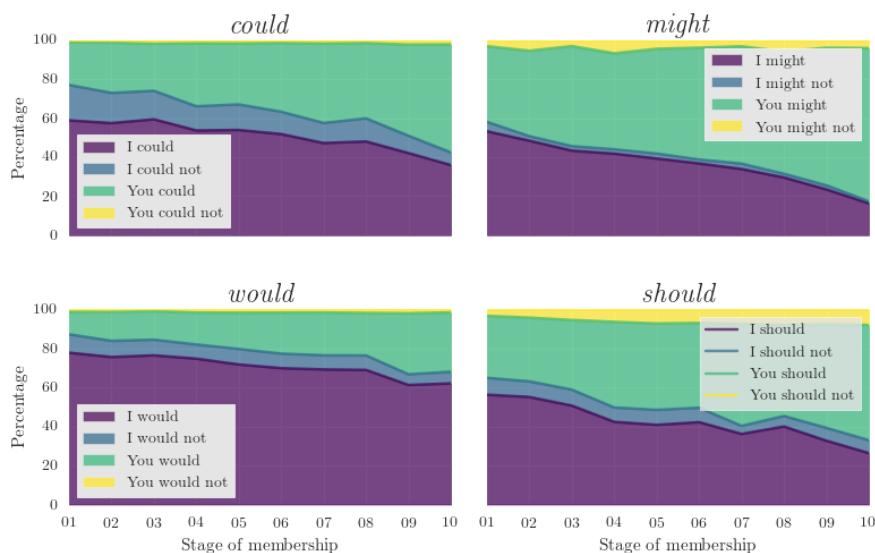


Figure 6.11: Constellations of Subject, Modal and Polarity

6.4. Summary

Mood and Indicative Types shift across the course of membership: imperatives and modalised declaratives become more frequent as users gain the social capital necessary to dispense advice or make (semiotic) demands on newcomers. Interrogatives do not exhibit a clear trajectory, because the community roles of both veterans and newcomers involve asking questions. For newcomers, questions are a means of obtaining information from the longer-term users; for veterans, questions are used to elicit information that may aid in advice provision. Questions can be used to demonstrate interest and investment in the narratives of others, and to hold others accountable for perceived absences from the group (*Where did ya go?*); alternatively, they can be used to invite newcomers to return to the community and/or to contribute again (Paulus & Varga, 2015).

MODALITY is used throughout membership, with increases in later stages. Exploration of *I + would* clauses reveals shifts in the socio-semantic activities in which Forum members engage, from narration of past circumstances (*sharing*) toward hypotheticals and suggestions for further behaviour (*advising*—see Matthiessen, 2013, 2015a). This, however, is a registerial insight that blends components of Field and Tenor (Matthiessen, 2015b). Looking at Mood Blocks more generally, a clear transition takes place over the course of membership: new and veteran members alike co-operate to hold the less senior members modally responsible in discourse. The system of TENSE exhibits change when modal tense is excluded, showing a longitudinal orientation toward the present and future, away from the past. POLARITY is shown to vary inconsistently when analysing posts alone, and to generally play a subordinate role to other components in the MOOD system.

In the next chapter, I use dependency parses to investigate longitudinal changes in experiential meanings made in the Forum, via choices of TRANSITIVITY.

7. TRANSITIVITY choices in the Forum

In Chapter 6, I analysed MOOD and MODALITY features of the corpus, discussing findings with respect to interpersonal meaning-making in the community. This chapter uses similar methods to focus on experiential meanings, and their realisation within TRANSITIVITY choices. This allows observation of how inner (mental) and outer (physical) states are represented by Forum members. TRANSITIVITY analysis of a clause involves distinguishing between Process Types or ergativity (represented in the main verbal group), as well as participants (the arguments of the verbal group) and circumstances (which modify the meaning made by the process). In this chapter, the analysis of participants is performed first, and processes second. That said, more delicate queries often involve consideration of both, as well as their attendant modifiers and circumstances. Rather than using constituency grammar, this part of the investigation relies on dependency parses, which more closely model the TRANSITIVITY system (Costetchi, 2013). Collapsed, conjunction-processed dependencies are used in almost all cases (see Manning et al., 2014).

For both participants and processes, the workflow is similar. Heads of groups are extracted from the corpus and transformed into relative frequencies and keyword tables. Linear-regression based sorting is used to determine which results are on increasing, decreasing, static and turbulent trajectories, in terms of both relative frequency and keyness. A subset of results undergoing obvious trajectory shifts are then analysed in greater detail, by probing their lexicogrammatical behaviour, and by concordancing. Profiles for these results, and for concepts indexed by sets of related lexis, are constructed in order to highlight differences in the way they are deployed in the subcorpora of each corpus under investigation.

Because of the strong methodological focus of the thesis, results are typically presented without manual removal of parsing errors, or of phenomena that are vague or difficult to interpret. Where necessary, the cause of errors is discussed before moving forward to analysis of the generated result. The argument being advanced, both here and in later chapters, is that doing discourse analysis or CL/CADS can be automated to a greater extent than it thus far has been: targeted searches of the lexicogrammar, exclusion of instances based on non-arbitrary information, and trajectory-based sorting can leave the analyst with summarised lexicogrammatical information with relatively obvious links to discourse-semantics. Unedited result lists are therefore presented in order to demonstrate both the promise of partially automated CL methods and the severity of limitations in the method that are discussed in Chapters 8 and 9.

7.1. Participants

To locate participants, a corpus can be searched for the lemma forms of dependency nodes of particular types, roughly corresponding to distinctions made within the SFG. For participant, the corresponding labels are *acomp*, *agent*, *appos*, *csubj*, *csubjpass*, *dobj*, *iobj*, *nsubj*, *nsubjpass* and *xsubj*. There are some inconsistencies between the grammars that are difficult to resolve, however: as one example, the Universal Dependency grammar used by Stanford CoreNLP annotates the Range in a process-range configuration (*I took a shower*) with the same label as it does a Goal (*I built a shower*). Even so, there is a great deal of overlap between the grammars, especially in the case of the general distinction between process and participant.

Because closed-class words have not at this point been excluded, the most common participant heads are generally pronominal (Table 7.1). Even so, the shift from first toward second person participants, and the relative stability in third person pronoun usage, are both observable, echoing findings from the investigation of MOOD choices, where first and second person pronouns feature prominently as Subjects, and where the latter come to displace the former in veteran talk. Within the domain of interpersonal meaning, the increasing second person as Subject indicated a shift in

Subcorpus	i	it	you	he	she	that	they	we	what	this
01	32.22	6.31	3.48	4.44	2.87	1.88	1.84	1.29	1.61	1.32
02	30.13	7.10	5.09	4.12	2.87	2.17	2.16	1.39	1.63	1.14
03	29.49	7.40	5.46	4.36	2.47	2.25	2.18	1.47	1.73	1.12
04	28.08	7.49	6.44	4.11	2.22	2.29	2.25	1.59	1.68	1.12
05	27.67	7.62	6.91	3.80	2.18	2.33	2.35	1.77	1.77	1.15
06	26.51	7.54	7.52	3.41	2.36	2.47	2.29	1.92	1.71	1.10
07	26.02	7.69	7.82	3.59	2.07	2.44	2.38	1.96	1.66	1.08
08	25.18	7.76	8.05	3.34	2.06	2.54	2.32	2.11	1.70	1.05
09	21.91	7.79	10.06	3.26	2.39	2.78	2.24	2.37	1.73	1.05
10	18.65	7.50	11.14	2.96	3.32	2.88	2.44	2.30	1.66	1.19

Table 7.1: Relative frequencies of common participant heads at each stage of membership

which interactant is charged with interactive demands. Within the system of TRANSITIVITY, however, the shift in meaning is an experiential one: *who is being spoken about* changes over the course of membership, from the self toward the addressee (Table 7.1).

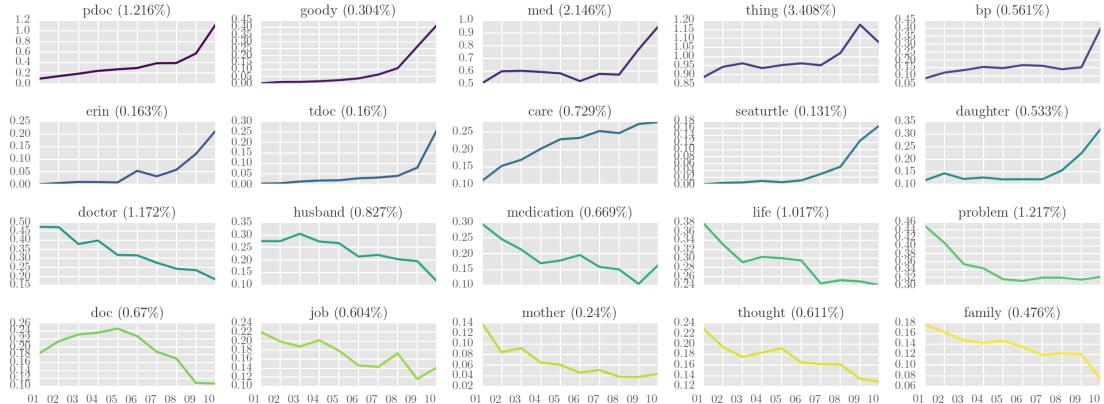


Figure 7.1: Trajectory of common participants undergoing change

Figure 7.1 shows the longitudinal trajectories of common open-class participant heads, as a percentage of all participants. From the 1000 most common heads, the ten on the most sharply increasing trajectory (top half), and the ten on the most decreasing (bottom half) are shown. Proper nouns denoting veteran members and their family/friends (*seaturtle*, *goody*, *erin*, *daughter*) dominate the ‘increasing’ results because veteran members refer to one another and commonly discuss those

close to them by name. Similarly, the declining frequency of *husband* and *mother* may have more to do with the particular circumstances of the small group of veteran members than it does to do with a normative value governing the ways in which the world ought to be construed by veteran members of the group. It is important to keep in mind, therefore, that the small sample of veteran users can cause an over/under-representation of particular participants. Fortunately, removing proper nouns is a trivial task, as they are annotated by Stanford CoreNLP with distinct POS tags (**NNP**/**NNPS**).

Table 7.2 shows the top participants, excluding pronominal and proper-nominal words. It uses log-likelihood keyness, rather than simple relative frequency, and collapses the ten subcorpora into three (*Early stages*, *Mid stages* and *Late stages*). Discursive insights become clearer here—negative emotional lexis appears as key in early contributions, while more positive terms and jargon appear as key in later posts. Before interpreting these, however, it is important to make note of the issue of parser accuracy, which has an impact on what is classified as a participant. *Welcome* and *thanks*, for example, are in some cases misannotated as participants. This is caused by short, minor clause sentences or sentence fragments, with which the parser model used is unfamiliar. Another problem is that some proper nouns have not been correctly removed: *whiskey* and *kait* are shorthand forms of two veterans' usernames, often misannotated as common nouns or adjectives. The misannotation is caused by the fact that the names may be written without capitalisation in the Forum—uncapitalised proper nouns, like minor clause sentences, are essentially absent in the parser training data.

A final view of overall participant frequencies is provided in Figure 7.2. This figure only charts trajectory changes among the 200 most frequent participants. This has the useful effect of automatically excluding many erroneous results. Exploratory concordancing of various lexical items in Table 7.2 and Figure 7.2, revealed four key themes: *jargonisation*, *metadiscourse*, *vague language* and *the construal of instability*. These themes are explored in the sections below.

Early stages		Mid stages		Late stages	
bipolar	709.26	person	114.96	pdoc	2002.63
new	692.95	doc	97.41	tdoc	735.35
doctor	396.43	hard	87.91	med	339.99
medication	254.38	luck	73.50	glad	316.05
mother	229.52	be	66.70	hug	315.12
psychiatrist	178.40	good	61.09	stability	287.02
dr	161.91	lol	60.07	daughter	278.64
swing	159.18	illness	59.41	able	229.33
medicine	139.93	whiskey	55.67	son	208.47
problem	136.31	kindness	54.89	thing	193.58
angry	135.12	happiness	47.38	important	188.02
mg	125.84	bp	45.07	okay	176.41
scared	125.47	dh	39.75	cycling	170.74
episode	122.30	positive	39.68	stabilizer	162.11
old	121.93	sorry	35.97	sorry	128.38
life	120.06	lot	34.37	welcome	120.52
crazy	115.24	peace	33.37	care	120.31
husband	110.02	spouse	32.32	support	112.70
kill	109.40	thanks	30.39	news	111.06
depression	107.70	ok	29.78	better	101.82
take	104.63	right	29.18	group	98.96
normal	100.55	care	28.60	hear	93.55
high	99.87	wendy	28.49	imbalance	93.49
worse	96.02	time	27.11	treat	91.19
attack	93.26	stuff	26.96	adjustment	90.46
adjustment	-31.53	sticky	-25.49	becuase	-35.54
easier	-31.64	kait	-25.49	normal	-37.82
step	-33.05	assessment	-25.49	cos	-38.51
bper	-34.94	earn	-26.43	nt	-39.22
able	-36.00	incorrect	-26.43	swing	-39.91
bf	-36.68	prn	-26.43	scare	-40.78
right	-37.27	marijuana	-27.38	moodly	-41.95
stable	-46.18	hypersexuality	-28.32	god	-45.27
okay	-48.81	mil	-28.32	dad	-46.83
post	-55.83	disorder	-28.80	hyper	-46.92
bp	-62.22	cycling	-28.88	horrible	-48.12
thread	-65.28	ed	-29.27	bipolar	-48.65
hear	-65.44	son	-29.57	medication	-62.73
tee	-67.39	flashback	-30.21	dr	-65.39
important	-78.59	psychiatrist	-30.60	crazy	-66.38
news	-80.07	typeing	-32.10	thanks	-68.69
stabilizer	-84.75	titrate	-32.10	mother	-74.03
welcome	-90.33	impulsivity	-32.10	scared	-78.99
care	-92.95	daughter	-34.83	mad	-79.00
stability	-103.53	pdoc	-34.99	husband	-79.01
sorry	-131.56	stability	-35.74	doc	-84.92
hug	-133.61	hi	-44.81	mg	-110.55
glad	-210.66	bipolar	-86.69	medicine	-123.90
tdoc	-309.27	new	-90.07	doctor	-196.93
pdoc	-1126.49	tdoc	-93.25	new	-241.95

Table 7.2: Key and unkey participants in three stages of membership

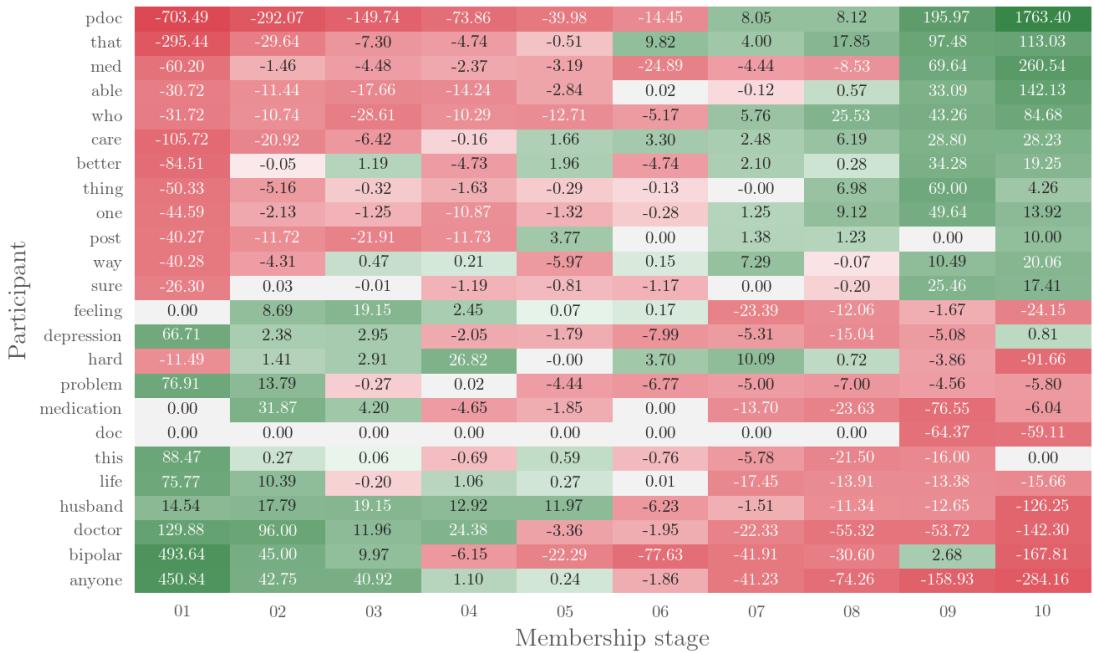


Figure 7.2: Keywords on increasing and decreasing trajectories (symmetric logarithmic colour scale)

7.1.1. Jargonisation

The first theme of interest based on participant frequencies is jargonisation. In systemic-functional terms, jargon serves important roles within both interpersonal and experiential metafunctions. Interpersonally, jargon demonstrates familiarity with community norms and expectation (Martin & White, 2005), and can demonstrate a (lay) expertise that enhances legitimacy. Experientially, jargon can also achieve more delicate distinctions between important participants in a given Field of discourse: the shift from construal of *mood swings* to states of *mania* and *hypomania* highlight the potential for jargon to facilitate more advanced taxonomisation of bipolar disorder and its symptoms.

Over the course of membership, jargon becomes significantly more common: *meds*, *pdoc*, *tdoc*, *bp* and *mania* are all Forum jargon, each of which ranks among the most key participants in late stages of membership (Table 7.2) and the top ten participants in veteran member talk in terms of relative frequency. Jargon terms also steadily displace non-jargon variants: *medication* becomes *med(s)*, and *bipolar* becomes *bp*. Figure 7.3 shows this pattern in four subcomponents of the Field of

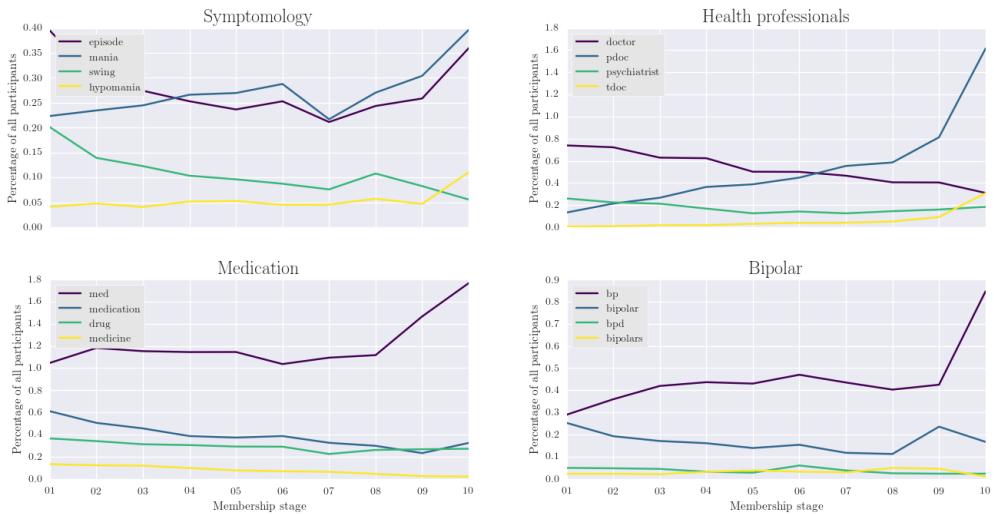


Figure 7.3: Jargon term use by postcount

discourse overall. Within symptomatology, *mood swings* decrease, as discussion of *mania* and *hypomania* emerge. Terms for health professionals are shortened: psychiatrist and doctor become *pdoc* at later stages. For medication, *med(s)* rises in frequency while *medication* and *drug* decline. For bipolar disorder itself, veteran members show a dramatic shift in preference toward *bp*. Interestingly, some changes are multi-stage: Figure 7.1 shows how *doctor* becomes *doc* in middle stages, before finally splitting into *pdoc* and *tdoc* in late stages.

A final point to note here is that the relationship between participant and lexical item is not perfectly one-to-one, as many agnate terms (often in the form of jargon) exist at different stages of membership to denote the same Thing. Therefore, to more accurately gauge shifting Fields of discourse, it is necessary to attempt at least a basic collapsing of participant taxonomies, and of jargonised and non-jargonised terms. This is performed later in the chapter.

7.1.2. Metadiscourse

As can be seen in Figure 7.4, *board* and *group* are the main way that members refer to the community itself (0.044 and 0.065 per cent of all participants, compared to *forum* at 0.009 per cent). *Board* and *group* become more frequent over the membership course, as do other terms that denote features of the OSG such as *post* and *thread*.

While new members typically commend the board and explicitly mark the fact that they have recently found it, veteran members speak on its behalf and outline its goals and orientation (see Table 7.5 for instances of *board*, not limited to participant roles). Over the course of membership, users thus increasingly construe the community as a Thing that can bring about Events and Goals. Most importantly, the Forum is capable of acting *for* users in the service of providing information and support. Construal of the Forum as a static entity acted upon by its members, or as a non-essential circumstance of location, decreases over time.

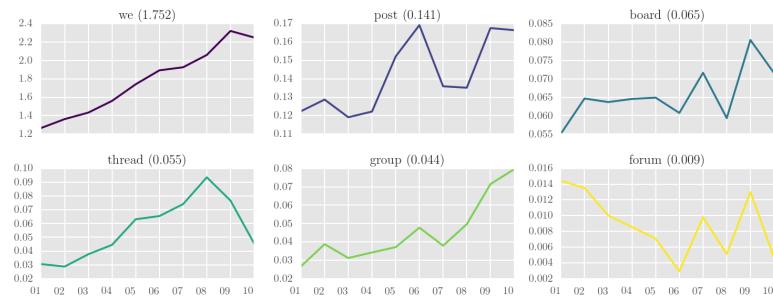


Figure 7.4: Key metadiscourse words

New users

1. *I'm so happy to have found this **board**.*
2. *I'm not new to the [website], but new to this **board***
3. *Thanks for this **board**, everyone's posts have been really helpful to read.*
4. *I am hopeful regarding being on this message **board** and sharing with all of you.*
5. *i just joined this message **board**.*

Veteran users

1. *I'm glad you told him about the **board** and about NAMI.*
2. *Part of the purpose of the **board** is for venting.*
3. *Hello, Welcome to the **board**.*
4. *on this **board** at least , pdoc is psychiatrist and tdoc is shorthand for therapist...*
5. *Your sage comments are missed by all on the **board**.*

Figure 7.5: References to *board* in new and veteran talk

Though experiential meanings, strictly speaking, construe information about events in the world, rather than role-relationships between participants in interactions (Eggins, 2004), salient role-relationship negotiation is performed through the differing discursive function of metadiscourse according to membership length. New members' discussion typically highlights the division between the community and the self (*I'm so happy to have found this board*), whereas veteran members construct

a shared identity with the board (*Hello, welcome to the board*). Further, only veteran talk about the board is elaboratory or explanatory (*Part of the purpose of this board is for venting*).

Also shown in Figure 7.4 is that the use of *we* increases over the course of membership. Concordancing of Subcorpus 10 shows that *we*, like *board* and *group*, commonly occurs during explanation of the function(s) of the community to newer members. As can be seen in Figure 7.6, another common use of *we* in veteran talk is in generalisations about the lived experience of all people living with bipolar disorder—a discursive strategy that simultaneously stakes a claim to knowledge about bipolar and highlights the role of lay-experience as the source of this knowledge.

- | | |
|---|--|
| <ol style="list-style-type: none"> 1. <i>Hang in there and know that we are all rooting for you.</i> 2. <i>Keep on venting and taking one thing at a time and know that we are here.</i> 3. <i>So what happened, [username] ... we were sooo worried about you the other night.</i> 4. <i>None of us can diagnose you since we are n't medical professionals, but it very well could be bipolar.</i> 5. <i>Also remember that we are all here for you anytime you need us.</i> | <ol style="list-style-type: none"> 1. <i>For whatever reason, we want to think that the brain we were born with is just fine thank you very much.</i> 2. <i>With bipolar disorder the rock does edge off of us - it just seems to take forever and we are so powerless beyond our meds to make it go away.</i> 3. <i>I think sometimes we are just too overwhelmed to act.</i> 4. <i>I think part of the reason we crash is that we are simply exhausted.</i> 5. <i>For after all, we are all the sum total of our experiences.</i> |
|---|--|

Figure 7.6: Two functions of *we* in veteran users' language

7.1.3. Vague language

The appearance of *thing* as an increasingly prominent participant in veteran talk warranted further analysis. Initially, lemmatisation had conflated *thing* and *things*, with the latter being the site of the most change. Concordancing of *things* in new member posts showed that *things* often described past circumstances (*things were getting really bad*). In contrast, veteran members used *things* in expressions of general support (*things will get better!*):

1. *I hope things improve for you soon*
2. *I am glad that things are looking up for you too*
3. *call your pdoc if things do not improve very soon!*

4. decided to hold off a semester until you get **things** more under control
5. as soon as you get back on your meds **things** will be better, just wait and see!!!

To measure this difference, the tenses of clauses in which *things* takes the position of experiential subject (generally Actor, but, more rarely, Sayer, Senser, or Token/Possessor) in new and veteran members' posts were tallied. Figure 7.7 shows that *things* is more commonly used by new members to describe action in the past. In veteran talk, *things* tends to be vague, and is more commonly focussed on the present and future. This change echoes the more general shift toward present tense, as outlined in the previous chapter. Though more in-depth analysis is perhaps warranted, this feature in veteran talk could potentially be a strategy for providing advice that may be of benefit to the entire community, rather than simply the thread initiator. An alternative analysis is that *things* may allow veterans to provide social support without the need to ask follow-up questions. Given the high dropout rate of new users, with most new users posting fewer than three times, requests for clarification of specific circumstances may often go unanswered. In order to complete more exchanges, advice is accordingly dispensed as soon as possible, with little clarification sought. Another manifestation of this strategy can be seen in Emz's reply to a new member in Chapter 5: rather than attempting to clarify unexplained parts of the newcomers' narrative, Emz instead simply highlights that her beliefs and suggestions are based on incomplete information (*From what you say, if sounds very likely that you might have Bipolar disorder; If you're already seeing a psychiatrist [...], then find a different one*).

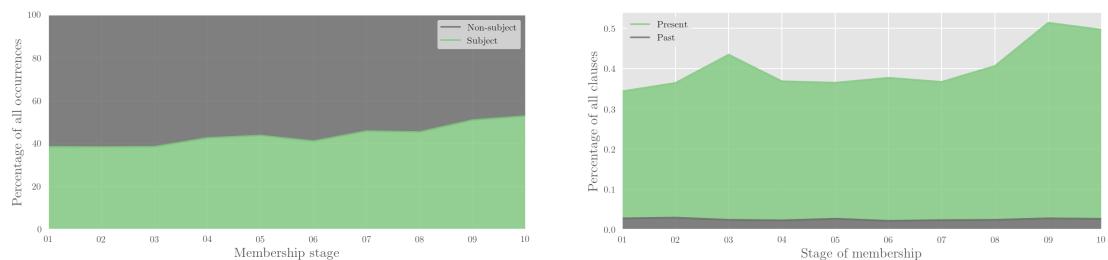


Figure 7.7: *Things* as an increasingly common left participant, and as an increasingly common left participant in present tense clauses

7.1.4. Construing (in)stability

Other results highlight changes in experiential semantics that can be mapped to the biomedical orientation of the community. One striking pattern evident in Table 7.2 is that newer members commonly construe negative mental states and desires that index instability and volatility (*swing, angry, scared, crazy, kill*, etc.). For veterans, such terms are unkey, being reconfigured under an attitudinally neutral or potentially medicalised nominalisation, *imbalance*. As seen in the qualitative analysis and noted in related literature (e.g. Horne & Wiggins, 2009), new members produce medical narratives connoting urgency in order to legitimate their membership bid and elicit responses. The framing of bipolar symptoms in emotive language, however, is dispreferred, with veteran members avoiding attitudinal lexis in favour of a nominalisation, *stability*, which is represented as the possible result of proper care.

terms of making sure that she finds	stability	.
t your finding and maintaining your	stability	as well .
are important as far as maintaining	stability	.
to try different avenues to achieve	stability	.
that way you will reach	stability	so much quicker .
those of us who have finally found	stability	for ourselves or our loved ones it
If so that you will be able to find	stability	and identify the stressors and deve
he has given me more	stability	over 2 months than i 've had in 1.5
our " seasoned " bper who has had	stability	for over 35 years .
between of something resembling "	stability	" or " normalcy . "

Table 7.3: *Stability* in veteran posts

7.1.5. Construing human agency

Because jargonisation disperses the potential lexical realisations of participants, wordlists must be created that can collapse the distinction between participants with multiple realisations in lexis during the corpus interrogation process. The main human participants (or social actors—see Van Leeuwen, 1996) in the Field of discourse can be divided into four : *The Self, Other Members, Friends/Family and Health Professionals*, with wordlists developed to match jargonised and non-jargonised variants (Table

7.4). For the remainder of the thesis, these capitalised terms denote the results of the four search queries, while lowercase variants refer to the social actor in general. Of course, such an approach means that many instances will go uncaptured, due to grammatical metaphor, misannotation, pronominalisation and so on. Even so, there is sufficient data for exploring how these four participant types behave over the membership course. Figure 7.8 shows the relative frequency of four kinds of human participants over time. Clearly, Other Members come to steadily displace The Self as the ideational social actor being represented. Friends and Family, meanwhile, are on an uneven trajectory, occupying a larger part of the semantic space in early and late stages of membership than in the middle. Construal of Health Professionals rises dramatically in the late stages of membership.

Participant type	Realisations
The Self	<i>i, myself, me</i>
Other Members	<i>you, yourself, kat, seaturtle</i>
Friends/Family	<i>friend, mother, husband, daughter, son, wife, father, dad, mom, mum, erin</i>
Health Professional	<i>doc, docs, doctor, doctors, dr, dr., drs, g.p., gp, gps, nurse, nurses, pdoc, pdocs, psych, psychs, shrink, shrinks, tdoc, tdocs</i>

Table 7.4: Human participants and lexical realisations

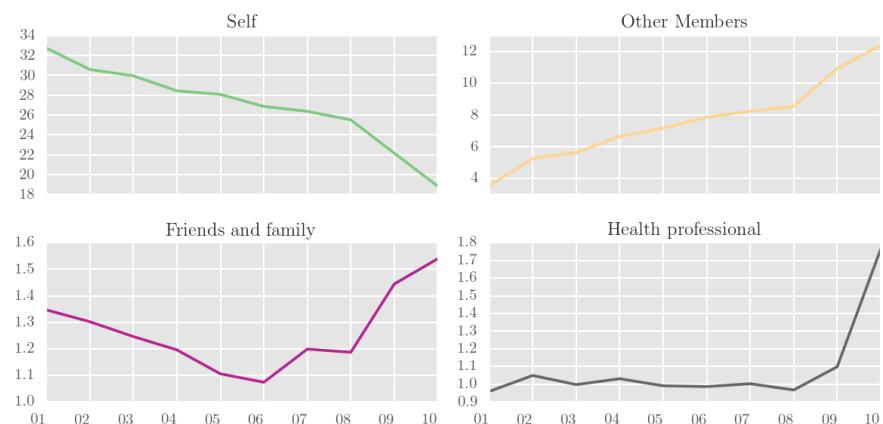


Figure 7.8: Relative frequency of four participant types over the course of membership

Next, the ergative model of TRANSITIVITY can be used to determine the extent to which these human participant types are construed as Agents, and how this shifts

with membership length. To measure this, each occurrence of participant type occurring in an Agentive dependency role can be divided by occurrences of the same participant type filling any participant role (Figure 7.9). This controls for differences in the overall frequency with which certain participant types are mentioned.

Most notable in Figure 7.9 is the wide gap between health professionals and non-health professionals in early contributions: when professionals are mentioned as participants in first posts, they are in positions of experiential agency over 75 per cent of the time. Non-health professionals, on the other hand, are more commonly positioned as Media—that is, as the entity through which a process comes into being. This is a representation of a world where health professionals are granted more control over medical processes than healthcare consumers, and where change in the world is enacted through the consumer, who may be willing or unwilling: patients are treated by doctors; doctors tell patients how to manage their condition.

The clearest individual trend is the increasing agency granted to Forum members generally. Veteran users construct a discourse of consumer-centredness by positioning their interlocutors as able to actively make changes in their inner and outer states. Representation of The Self changes in a similar, but not identical way. Though The Self is less often charged with ensuring the completion of the clause as an intersubjective proposition (a responsibility increasingly delegated to the addressee), The Self is also increasingly positioned as an entity that can bring about change in the world. Finally, there is a modest decrease in the proportion of *Health-professional-as-Agent* over membership length, closing an ideological divide present in early contributions, where health professionals carry out processes through the Medium of the sufferer.

7.2. Key processes

The next focus of the analysis of TRANSITIVITY choices is on processes, as typically realised by verbal groups with at least one nominal group argument. Of primary interest is the head of this Process, which corresponds to the Event in the SFG. In a constituency grammar, this is generally the rightmost verb in a VP; in the Universal

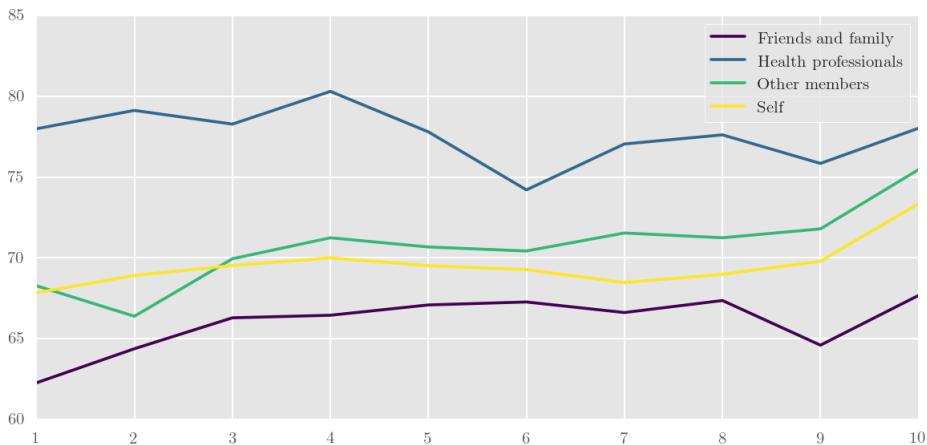


Figure 7.9: Proportion of each participant type in Agent role

Dependency grammar, Events are annotated with *ccomp*, *cop*, *advcl* and *root* labels. Therefore, an analysis of processes can begin in the same way that the analysis of participants was carried out, but with process-like labels substituted for participant-like ones.

Figure 7.10 shows which Events are key in each subcorpus according to a log-likelihood comparison of the Forum contents as a whole. By far the most key process in any subcorpus is *diagnose* in Subcorpus 01, as diagnosis is both the catalyst for visits to the site, and the main entry condition for participatory information and support exchange. Similarly, *thank* is key in second posts because users thank others for replies to their first. Other key Events in first posts reveal different kinds of motivations for joining the community, including being *told* that they may have bipolar disorder, *dates* with bipolar people, and *beginning/start*ing new medication regimens or manic/depressive cycles.

As with the analysis of participants, where veteran members orient toward more positive lexis, positive processes related to social support also become more common in veteran posts (*hug*, *thank*, *love*, *welcome*). Another focal point is processes urging others to carry on (*continue*, *remain*, *improve*), which also have positive connotations. This contrasts with the negative sentiments inherent in *suffer* (as in, *to suffer from bipolar*), which is key in first posts (see below for a more thorough treatment of the ways in which bipolar is ascribed to the self and others across membership stages). Finally, it is notable that *thank* and *appreciate* are unkey in the final stage

	01	02	03	04	05	06	07	08	09	10
Event	Subcorpus									
diagnose	906.92	66.42	0.67	-5.92	-23.07	-73.27	-75.73	-97.46	-67.38	-34.53
hug	-152.27	-71.36	-42.90	-21.31	-10.91	0.65	-9.87	-0.89	523.98	60.84
start	181.20	6.15	5.62	2.16	-0.84	-4.03	-7.21	-6.37	-25.08	-35.78
thank	-60.59	355.11	79.20	17.76	0.41	-0.25	-3.68	-12.89	-45.84	-237.36
please	117.15	-0.73	-0.30	-1.39	-0.10	-0.09	-13.68	-4.10	-1.67	-4.32
suffer	173.54	16.60	2.77	-2.32	-1.86	-1.12	-7.16	-11.54	-33.34	-56.88
continue	-17.54	-8.66	-2.41	-5.21	-2.59	-2.27	-1.69	14.37	10.59	84.21
love	-4.75	-14.38	-0.62	-1.25	-4.60	-0.01	-0.18	13.29	73.25	2.46
welcome	-127.23	-103.89	-65.66	-41.55	-2.16	-3.76	11.60	16.10	237.81	141.11
eye	-30.31	-19.95	-16.77	-12.39	-5.88	3.81	57.53	90.67	-9.03	-8.46
allow	-15.64	-3.43	-1.31	-10.12	-0.19	-0.23	-0.67	1.05	18.06	53.60
want	42.53	5.40	12.42	1.78	-1.31	-1.43	-0.47	-2.42	-1.44	-14.01
put	62.22	5.58	7.16	-0.17	2.96	-9.48	-2.96	-7.04	-9.25	-10.14
need	-45.55	-17.82	-6.51	-0.02	-0.03	0.34	4.64	13.30	38.53	46.85
call	-6.68	-4.30	-0.02	0.93	-1.91	-0.00	0.00	0.20	1.55	43.26
suggest	-1.51	0.96	-3.04	-1.76	-0.43	0.06	-0.00	-0.14	-0.38	38.84
begin	46.25	2.09	-0.91	0.02	0.01	-6.27	-6.01	-0.52	-2.76	-0.17
plan	-3.66	0.50	-0.35	-5.46	-1.62	-0.31	1.38	0.29	-4.60	45.39
date	48.88	1.19	0.05	-4.37	-3.90	-2.55	-0.60	-1.05	-2.64	-3.62
ask	-10.87	-1.03	-0.89	0.36	-1.75	4.89	-0.60	2.16	-0.00	38.91
concern	-1.31	-0.00	-4.93	-0.13	-2.48	-2.82	-0.13	-0.46	5.87	37.35
work	-12.16	0.01	4.25	0.29	-0.07	-2.18	1.94	-0.10	-0.00	36.90
take	-0.59	6.40	2.30	0.02	0.00	0.00	0.08	-6.21	0.31	26.54
feel	132.97	20.21	16.66	20.40	6.77	0.13	-16.62	-15.12	-32.05	-108.40
dream	-9.97	-3.19	-0.82	-0.38	-18.73	4.62	38.47	-0.69	-2.10	17.50
use	-8.57	-0.49	-0.23	1.43	-0.04	0.17	-0.45	13.66	-4.40	23.46
hear	-4.08	-3.05	0.08	-4.66	-0.07	0.38	-0.17	-0.02	20.09	16.06
tell	15.70	16.37	16.56	0.00	1.21	-2.21	-2.23	-3.47	-19.22	-0.19
remain	-2.58	-1.17	-0.12	-1.32	-2.42	0.36	-0.00	-2.02	2.86	28.84
find	-5.11	0.03	-0.01	-1.74	-0.02	-0.79	9.59	-0.04	6.94	12.88
give	-4.55	-0.00	-0.83	0.01	-0.45	-1.27	5.66	0.62	0.65	21.77
mother	33.54	-0.99	-0.25	-1.42	2.66	-0.23	-1.84	-1.20	-0.38	-8.69
mention	-4.51	-0.23	-0.01	-4.38	-3.70	-0.02	-1.95	16.99	2.52	16.41
recommend	-0.67	-0.56	1.30	-1.95	-0.72	-2.27	-0.17	-0.63	1.83	24.78
lie	28.67	-0.23	-0.50	-2.24	-2.79	0.00	-0.33	-2.40	0.18	0.05
go	5.88	5.36	-0.01	1.17	5.22	-3.07	0.04	-3.75	-0.24	9.65
imagine	-9.83	-2.82	0.06	-0.37	-0.95	-0.14	1.13	-1.15	5.98	28.06
consider	-2.87	-1.61	-0.43	-4.25	-0.07	2.45	-0.91	16.85	-1.00	11.51
appreciate	99.83	56.31	21.17	-1.56	-2.55	-1.68	-22.46	-3.69	-40.59	-85.36
improve	-0.64	-1.11	-1.72	0.11	-3.17	-0.06	-3.79	-0.12	5.84	23.85

Figure 7.10: Key and unkey Events in each subcorpus

of membership; the increased social status in the community means that platitudes, for veterans, are no longer obligatory to the same extent.

7.2.1. Construing diagnosis

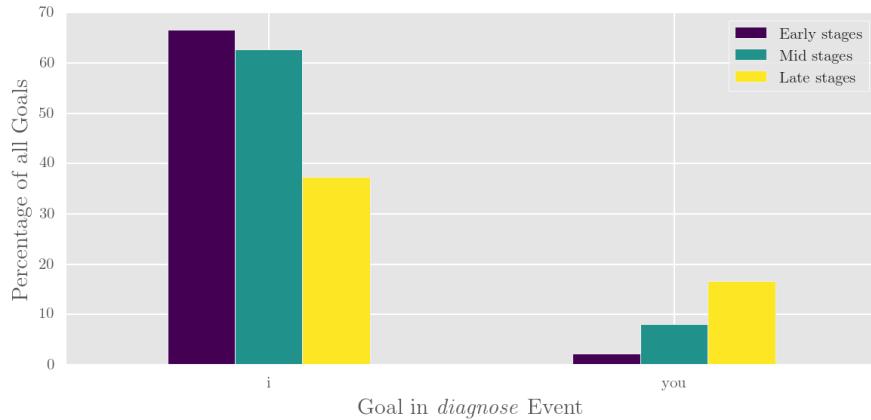


Figure 7.11: Goal of *diagnose* processes in three stages of membership

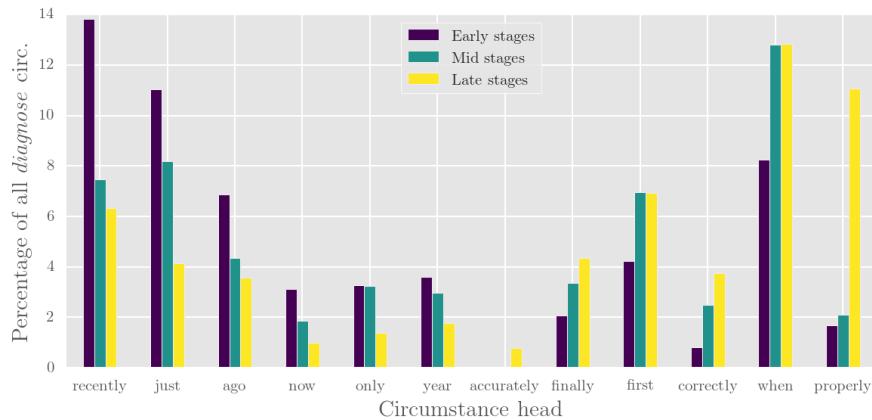


Figure 7.12: Circumstances in *diagnose* processes in three stages of membership

Diagnose is a very prominent process in the Forum, and is by far the most key of key processes expressed in first posts. By contrasting early, middle and late stages of membership, clear differences emerge in how the process of diagnosis is configured over time (Figure 7.11). Veteran members, for example, are more likely to represent the health professional as the Actor in the *diagnose* process. In terms of circumstances (Figure 7.12), there is a shift in focus away from temporal mean-

ings, with veteran members instead framing the *diagnose* process with regard to its correctness, accuracy and legitimacy. New members often enter the community because of a recent or possible future diagnosis. Veterans, on the other hand, seek to ensure that the diagnosis is reliable. This has the dual purpose of ensuring the legitimacy of the new member's 'entry ticket', and ensuring conformance with the biomedical model, where successful treatment is predicated on accurate diagnosis. Indeed, a number of users enter the community not because of a diagnosis from a health professional, but because they are attempting to ascertain whether their self-diagnosis, or their informal diagnosis of a friend or loved one, can stand up to the scrutiny of lay-experts. Veteran members, however, are reluctant to legitimise these kinds of strategies, due to their deviation from a normative conceptualisation of the 'correct' consumer journey.

Notably, veteran members may deliberately undermine the biomedical model of diagnosis. While expressing conviction that diagnosis must be performed by a qualified health professional, they may simultaneously hint that the newcomer is *probably* bipolar, or that non-bipolar diagnoses are in error. As was shown in the qualitative analysis in Chapter 5, two separate replies to an undiagnosed newcomer relied upon the same lexicogrammatical means of hinting (*It sounds like you might have bipolar to me; if sounds very likely that you might have Bipolar disorder*), while nonetheless insisting on seeking diagnosis through mainstream channels (*Find you a dr that can make a correct Diagnosis; We aren't docs here and can't diagnose you*). As shown in the following section, in veteran–veteran interactions, health professionals' ability to correctly identify and treat people living with bipolar is occasionally into question. This keeps the membership entry point open for those who have initially failed to meet the core criterion of a legitimate diagnosis, while pushing undiagnosed, suspected bipolar users toward actions that are in line with mainstream medical norms.

Diagnosis and grammatical metaphor

Grammatical metaphor entails the use of one grammatical component to do the work that is congruently performed by another. Nominalisation is one of the most common examples (Simon–Vandenbergen et al., 2003). What is congruently an action,

process or Event (e.g. to *applaud*) may be reconstrued as a participant (*applause*), allowing denser packaging of information. Turning the process into a participant also facilitates taxonomisation and classification: the lexical component of a nominal group may include Classifier, Epithet and Numerative, in addition to the Thing; in the verbal group, the only lexical component is the Event (Halliday & Matthiessen, 2004). This kind of grammatical metaphor also opens up the reconstrued process to deixis. For these reasons, it is a key characteristic of scientific English (Halliday & Matthiessen, 1999).

Over the course of membership, the process of diagnosis undergoes a steady shift toward metaphorical realisation as a participant. In formal terms, it is more often nominalised. Figure 7.13 shows the strong relationship, but inexact, relationship between nominalisation and grammatical metaphor in the case of diagnosis: nominal and participant realisations become more frequent, while verbal and process realisations decrease. Charting experiential roles shows us that *diagnose* is not limited to participant and process roles: commonly, it is a part of a circumstance (*we received more help in terms of diagnosis and treatment*) or modifies a Thing (*it is extremely common for those with undiagnosed bipolar disorder to self medicate*). The increasing extents to which *diagnose* is nominalised, and to which *diagnose* is represented as something other than a process, demonstrate an important discourse-semantic shift. By moving away from *diagnose-as-process*, it becomes possible to represent diagnosis as a possession (*it's great you have a final diagnosis and have started medication*), and therefore as something that can be acted upon or thought of in a particular way (*i finally feel like i've accepted my diagnosis*). Another possibility opened up by nominalisation is the potential for modification through Epithets (*this is a frightening diagnosis, particularly if you don't know anyone who has it*). Classification of *diagnosis* through adjectival modification is also made possible (*accurate/correct diagnosis*), but, of course, as shown in Figure 7.12, agnate meanings can be made through circumstantial modification of *diagnose* as a process (*accurately/correctly diagnosed*). A final grammatical affordance is that diagnosis as a participant in a relational process can highlight its potential to be incorrect (*my current diagnosis is schizo-affective disorder; bp is becoming a catch-all diagnosis, frequently made by a well-*

meaning family doctor). In this way, grammatical metaphor not only contributes to an increasingly scientised register, but also expands veteran members' ability to explain medical processes and their relationship to other members (Heyvaert, 2003).

Calculating the relative frequencies of common modifiers of diagnose as verb (*diagnose, diagnosed, diagnosing*) and diagnose as noun (i.e. *diagnosis/diagnoses*) clearly shows a preference for temporal modification of verbs and veracious modification of the nominal form (Table 7.5).

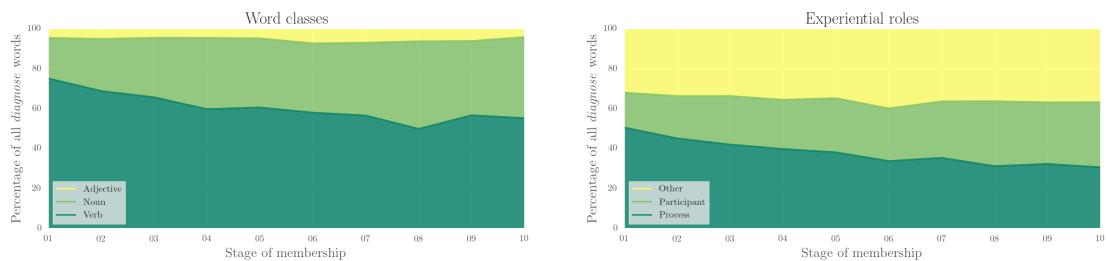


Figure 7.13: Grammatical metaphor in *diagnosis*, via word classes and experiential roles

<i>Diagnose</i>	Rel freq.	<i>Diagnosis</i>	Rel. freq.
not	11.80	bipolar	11.13
ago	11.10	proper	8.82
just	7.60	correct	7.49
recently	6.68	dual	5.07
when	4.16	right	4.41
now	3.54	new	3.96
only	3.11	official	2.75
properly	2.75	different	2.64
also	2.27	wrong	2.20
finally	2.12	other	2.20
so	2.09	same	1.87
never	2.09	initial	1.65
then	1.93	possible	1.54
correctly	1.50	true	1.54
well	1.47	recent	1.43

Table 7.5: Most common modifiers of *diagnose* and *diagnosis*

7.2.2. Construing the relationship between people and bipolar

As noted by Harvey (2012), people attribute depression to themselves in a number of different ways. The same set of grammatical constructions are available to those afflicted by many health issues, including bipolar disorder. In the Bipolar Forum, people may *have*, *be*, *feel* or *suffer from* bipolar. For reasons of scope, however, only the two most common clause-level ascriptions—*having* and *being*—are examined in detail here, and group/phrase level ascriptions (i.e. *a bipolar friend*) are not considered. To look for longitudinal changes in being/having constructions, the corpus was interrogated for relational processes with *bipolar* (or jargon variants *bp*, *bpi*, *bpii*, *bp1*, *bp2*, etc.) as a first object argument, and with human pronouns as the leftmost nominal group, returning lemma forms of the located relational Events.³² Generated concordance lines were used to remove false positives caused by parser errors. Results were then recalculated from the concordance lines using `corpuskit's` `Concordance.calculate()` method.

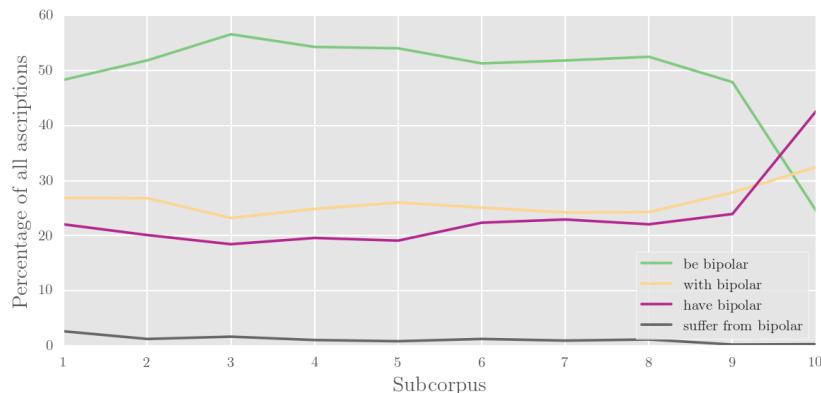


Figure 7.14: Processes with bipolar as Medium, combined

Figures 7.14 and 7.15 show changes in the way Forum users construe the relationship between bipolar, themselves and others. Most strikingly, *having* forms overtake *being* forms as ways of ascribing bipolar disorder to the self and others. This change occurs very late in the membership course. Similar is the growth in attribution via circumstances, as in *a person with bipolar*. To *suffer from bipolar* decreases steadily over the course of membership.

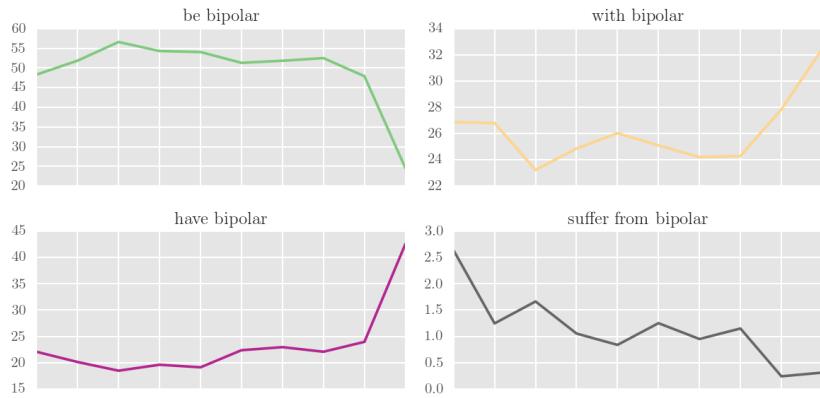


Figure 7.15: Processes with bipolar as Medium, separated

Understanding these changes requires analysis of the grammatical properties of each kind of ascription. First, the shift away from *suffering from* and *feeling* is away from mental and toward relational processes. This shift includes the increasing frequency of attribution via *with*, as grammatical circumstances are in fact partially articulated relational processes (Matthiessen, 1995). Both *being* and *having* constructions are attributive relational processes, meaning that in both cases an attribute is ascribed on some level to the experiential subject. The difference is in the nature of the attribution. *I have bipolar* is a **possessive attributive**, in which the experiential subject functions as a possessor of the attribute, which is inherently possessed. Possession itself is foregrounded in the process of having; ownership of bipolar is ascribed to the subject. Inversion of the clause highlights the dominance of the possessor over the possessed: *bipolar has me* conveys an (undesirable) lack of control over the condition. *I am bipolar*, on the other hand, is an **intensive attributive** construction, where bipolar forms a class in which the subject is a member. Halliday and Matthiessen's (2004) use of the term *Carrier* for the subjects in these kinds of processes highlights the ambiguity concerning the Carrier's willingness to be ascribed the Attribute. This construct is ultimately dispreferred by veteran users, as it creates a subclass of person characterised chiefly by the ascribed bipolarity, rather than foregrounding the process of ownership, and, by extension, some control over their possessed condition.

The salience of the two constructions is evident in the fact that veteran users of the *Bipolar Forum* may explicitly draw attention to the *being/having* distinction. Recalling the qualitative analysis of a new user's interaction with two more senior members (see Section 5.1.1 for the complete reproduction), we can see that each interactant construes Jess as having bipolar disorder using different relational process lexis:

Jess: [...] i have asked my doctor to test me to see if i am bipolar as my antidepressants do not work even though they have been changed a million times!! [...] i want to change my doctor and have been telling my partner i would but im scared of finding out that i am bipolar [...]

Luvsoccer: It sounds like you might have bipolar to me. You need to change Drs. One with more knowledge apparently. The reason the antidepressants are not Working is because If you are bi polar and they put U on an antidepressant alone it can make things worse... [...]

Emz: Hi, welcome to the boards, hopefully we can help you out and be a support system for you. First off you're not A bipolar, *!* we're not things, it's a condition. From what you say, if sounds very likely that you might have Bipolar disorder. [...]

Jess exclusively uses *being* forms, while *Luvsoccer* switches between *having* and *being*. Emz, however, uses only *having* forms, except when explicitly drawing attention to the fact that copula ascription is dispreferred. To make this point, she highlights a potential reading of the *being* construction as an *intensive identifying* process, whereby the experiential subject has *bipolar* assigned to it for the purposes of identification, rather than quality attribution. This is undesirable among the community: as explained by the veteran user (Figure 7.16), it is preferable to be a member of a class of *people* possessing a particular quality of *bipolar*, rather than a relationship of equivalence between the Identifier and Identified.

Figure 7.16: A veteran user discussing the *being/having* distinction

Some of the rare instances of *being* constructions in Subcorpus 10 are in fact the result of veteran users recasting new users' claims of *being bipolar* as *having*

bipolar, as can be seen in Emz' response to Jess. Note the veteran's insertion (and capitalisation) of the indefinite article during her recast: *First off, you're not A bipolar, we're not things.* Here, the veteran member is highlighting a second potentially undesirable reading of the *being* construction, where rather than functioning as an Epithet, bipolar disorder becomes a Thing. This transforms the attribute into an entity, rather than Quality, and renders the class of bipolar (as the veteran member points out) a Thing to which a person can be equated, rather than a trait that may (currently) characterise someone.

Aside from recasts, in veteran posts, *being* constructions are exceptionally rare. In the highest postcount group, only 11 matches were found, with eight cases involving ascription to others, rather than the writer him/herself. One post, in which the self was identified as *being* bipolar three times was selected for further analysis (Figure 7.17).

The screenshot shows a forum post by a user named 'Veteran user'. The user is identified as a 'Senior Veteran (female)' who joined in Dec 2004 and has 1,232 posts. The post is titled 'Re: Movies in head?' and contains the following text:

The pdocs at the first hospital said I was bp1. Then at the state hospital they said I was bp2, then they said I was schizoaffective bipolar type.

Current Meds:

- Abilify 30mg
- Lamictal 200mg
- Wellbutrin xl 450mg
- Lexapro 10mg
- Ativan 0.5mg, then 1mg at bedtime

Figure 7.17: A veteran user employs *being* forms

Here, pivotally, it is pdocs, rather than the speaker himself, who (twice) invoke the *being* construction, through an embedded clause in the Verbiage. Given that elsewhere in the Forum the same member had employed *having* constructions, analysis of entire threads was performed, in order to better understand veteran's motivations for using *being* forms. In multiple cases, veterans appeared to use *being* constructions to help establish a negative characterisation of the health professionals in the process of diagnosis. In an earlier post, the user had told the story of his misdiagnosis:

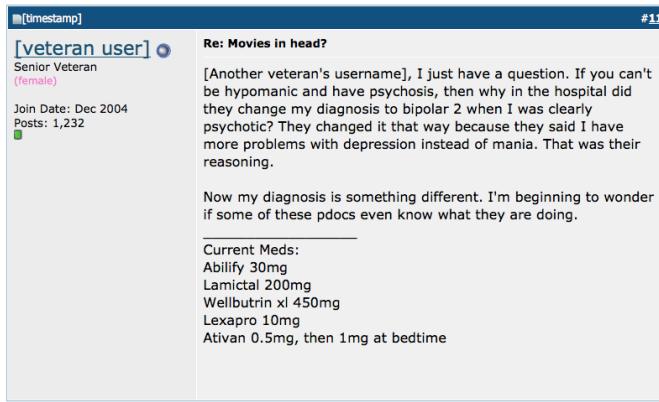


Figure 7.18: A veteran user's characterisation of pdocs as incompetent

The writer's uses of the dispreferred *being* constructions add weight to his positioning health professionals as incompetent: in the narrative, health professionals can neither offer correct diagnoses for their patients nor adequately conceptualise the relationship of ownership and possession between person and condition that functions as an important normative value in the community.

7.2.3. Process-participant type configurations

Earlier, I analysed the frequency of four kinds of participants in the Forum's Field of discourse (*The Self, Other Members, Health Professionals and Friends/Family*, and the proportion of occurrences of each participant within the role of Agent (within an ergative interpretation of the system of TRANSITIVITY). The transitive interpretation of the system can be used to investigate the processes each Agent is involved in. Figure 7.20 visualises the keyness of lexical heads of processes.

Finally, the individual processes can be collapsed by Process Type (Figure 7.19). A four-type system, as described in the Cardiff Grammar, is used here, due simply to the availability of the Process Type Database (Neale, 2002). Processes are defined broadly. *Feel* and *smell*, for example, are counted within both *mental* and *relational* types (though they could be differentiated by counting the number of participants). As the Process Type Database did not contain some common processes found in the corpus (e.g. *diagnose*), the top 200 processes in the corpus were added to lists

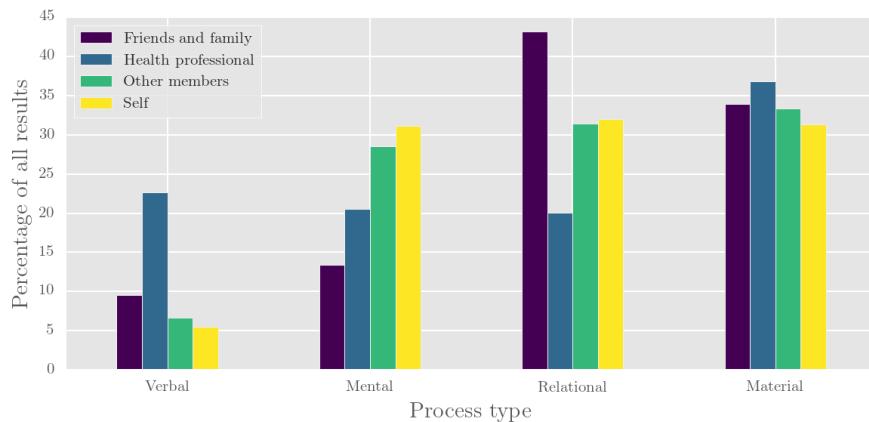


Figure 7.19: Four participant and Process Types

manually if absent based on intuition and grammatical tests. Material processes are simply those that do not occur in the other three lists.

It can also be observed that different participant types are typically construed as participating in different kinds of activities. Health professionals are most commonly Sayers, as language is the means through which mental health treatment is predominantly provided: diagnosis, elicitation of symptoms, prescription, advice, referral, and most kinds of therapy are accomplished through talk. The Self is positioned as a Sayer far less often: speakers do not construe their own speaking within the Field of discourse, because speakers' talk is primarily an interpersonal, rather than an experiential phenomenon; its function is to enact and negotiate with the addressee.

Mental processes of thinking and feeling are most commonly performed by the Self. This is perhaps expected, because users have unmediated access to their own thoughts; others' mental processes are often accessed through other Process Types. Other members are commonly construed as Sensors in irrealis scenarios, and/or during the provision of advice or health information (see Table 7.6). Formulaic or idiomatic expressions such as *if you know what I mean* and *you know* also play a role. Rarely, however, do veteran members construe newcomers' manic/depressive episodes.

Friends and family, meanwhile are most commonly represented within relational processes. Two reasons for this are that many users enter the community seeking

Process	Friends and family	Health professionals	Other members	The self
	Participant			
hope	-310.15	-322.15	-1516.58	2458.99
think	-451.25	-120.10	-692.05	1507.95
know	-511.84	-275.42	-179.97	899.85
say	104.04	1248.77	94.73	-825.15
guess	-158.76	-152.69	-924.20	1405.31
feel	-469.99	-666.58	163.16	52.34
prescribe	-9.08	1156.98	-41.99	-374.39
see	-15.53	1391.61	4.18	-308.38
call	59.47	1070.94	-17.06	-283.50
die	583.52	-13.68	-55.17	-90.75
talk	27.02	818.31	6.20	-329.95
wonder	-93.68	-71.12	-382.35	634.20
tell	58.40	905.14	-33.34	-112.09
thank	-13.37	-45.71	1152.63	-841.59
be	721.43	-309.21	-285.85	285.77
sound	64.18	0.12	412.23	-485.94
need	-49.14	-102.52	1973.44	-1134.59
hate	-11.83	-38.84	-386.09	481.41
try	-54.42	-8.73	-170.96	312.56
ask	3.49	580.98	-1.67	-113.88
diagnose	0.05	441.18	0.14	-155.65
mention	-3.95	123.71	221.41	-309.08
read	-59.44	-109.26	-41.90	182.23
believe	-43.78	-15.03	-158.22	267.68
sleep	-30.05	-100.77	-63.17	182.16
remember	-41.40	-41.59	-85.42	197.12
describe	-0.00	-0.20	397.98	-339.90
put	-7.05	509.09	-19.01	-37.70
appreciate	-23.56	-25.60	-112.57	192.16
start	-2.32	-10.69	-173.56	235.58

Figure 7.20: Keyness of processes involving four participants

1660 you do n't want him taking something that effects the chemstry of his brain
1370 do you know what has helped me with self-esteem ?
4064 i know how you feel about your son being on such strong medications .
4827 you have to be responsible in letting your pdoc know that your meds are n't working
2125 you just have to find what helps you the best .
4327 i hope that you will find the program very beneficial .
632 you have to train your mind to allow alot of this external stimuli to bounce off
4401 all will be okay , and it will be over before you know it .
1111 you have come to the right place to learn about your disorder , to find support
3018 can you help me understand why you can only contact your pdoc 1 time per month ?

Table 7.6: Other Members as Senser in veteran posts

a classification of a friend/family member's symptoms or behaviour as being or not being consistent with bipolar disorder. Relational processes are the congruent grammatical means of performing such classificatory work. Another reason is that friends and family are construed through sequences of relational processes that function as a description of medical history and demographics. In the example below, a new user describes her son and his actions via a sequence of relational processes, with the son in the position of Token. The text concludes by relating both her son and Self to Bipolar disorder using an Identifying relational.

I'm the mother of 3 sons. My middle son **is** our problem. Ugh, problem isn't even close to the appropriate word. But, I guess I'll go with that. Where to start.....as of now he's a chronic runaway, self medicates, he's **been** in trouble with the law, He spent 72 hours in the psych ward as a 5150 for threatening to kill himself, his emotions are up and down, he's nasty, hateful, on the other hand loving and warm. Towards me in particular, he blames me for everything and says if he's bipolar so am I.

This brief account of participant and Process Type constellations is a useful place to conclude the case study analysis. On one hand, the findings challenge Widdowson's (2000) argument that CL reveals findings contrary to expectation: we would expect that the Self is more likely to be construed as engaging in mental processes, simply because humans have constant access to their own thoughts. Likewise, we would also expect friends and relatives to be construed relationally (so that addressees understand the role of third parties in the discourse), and pdcs to provide diagnosis and treatment through talk. The expectedness of results is, foremost, an indication that the methods presented are capable of automatically developing an account of

discourse that is not at odds with intuition. At the same time, however, the role of the discourse analyst has not been erased: the subtle distinctions made within and across individual posts often elude search queries; the automated mapping of what is found to what can be found through close reading remains, for now, well beyond the power of computational tools and methods.

7.3. Chapter summary

In this chapter, I have explored TRANSITIVITY choices in the Forum, and how they change longitudinally. By progressing from frequency and keyness calculations of participants and processes toward analysis of salient items and their syntagmatic behaviour, I showed how Forum members construe the world differently at different stages of membership. In the next chapter, I discuss the findings presented in the previous three chapters, mapping lexicogrammatical changes to discourse-semantics, and relating both to findings of the body of literature reviewed in Chapter 2.

8. Discussion: meaning-making in the Bipolar Forum

In the previous three chapters, the findings of the case study were presented. Following a brief qualitative analysis in Chapter 5, findings were organised according to lexicogrammatical system. MOOD and MODALITY were investigated in Chapter 6, showing how Forum members use language as a resource for role-relationship formation and maintenance, and for the giving and receiving of information and support. This was followed by an analysis of TRANSITIVITY choices, which highlighted longitudinal change in how Forum users construe the healthcare journey, both in terms of its processes (*diagnosing, being/having bipolar*) and participants (including Health Professionals, Friends/Family, Other Members and the Self).

In this chapter, I discuss key findings with reference to online health discourse literature reviewed in Chapter 2, and with respect to key concepts in SFL, as outlined in Section 3.3. This chapter therefore focusses on directly addressing Research Questions 1 and 2, which concern lexicogrammatical, discourse-semantic and registerial change over the course of membership in the Forum. Because discussion was presented alongside findings throughout the previous three chapters, the discussion takes the form of semantically organised summaries. Following this discussion, I provide a critical reflection on the findings of the case study, the theories that inform them, and the tools and methods that generated them.

8.1. Addressing Question 1: Lexicogrammatical features at risk

Answering the first question involves simply summarising what was presented in the previous three chapters. New information is not being presented here. The summary is necessary, however, because the second research question is predicated on the first: lexicogrammatical change must be identified in order to discuss discourse-semantics and register in the Forum.

8.1.1. MOOD and MODALITY

Chapter 6 detailed MOOD and MODALITY choices over the course of membership in the Forum. Imperatives and modalised declaratives rise in relative frequency, while unmodalised declaratives decline. Interrogatives, both modalised and unmodalised, undergo no clear trajectory shift.

The investigation of pronominal Subjects showed a clear shift from first toward second person. Modality was also shown to figure prominently in the Forum. Absent consideration of other MOOD features, the proportion of clauses modalised by any of four types (*can/could, will/would, shall/should* and *may/might*) increases. In terms of combinations of pronominal Subject and modal Finite, clauses with *you* as Subject, combined with any modal, rise steadily. The major exception is *I would*, which is more common in later stages of membership. The TENSE system undergoes less shift, but there is nonetheless an observable trend from a focus on the past toward a focus on the present. POLARITY does not show readily interpretable trends in either direction.

8.1.2. TRANSITIVITY

Chapter 7 detailed shifting patterns within TRANSITIVITY choices. Key participants in early contributions are often emotionally charged (*kill, crazy, scared*). *Anyone* figures prominently, as a general addressee. In later contributions, jargon terms for health professionals and for medications become very key (*pdoc, meds, tdoc*). References to

the community itself also rise in relative frequency over the course of membership. Shifts in the frequency of lexis denoting The Self, Other Members, Health Professionals and Friends/Family also take place, with each of these participant types being represented as Agents to differing extents longitudinally.

In terms of key processes, *diagnosing* and *thanking* are important within early contributions, while *welcoming*, *hugging* and *loving*, among many others, become more common. The *diagnose* process was investigated in greater detail, and changes in its typical attendant participants and processes uncovered: the healthcare professional figures more prominently in the configuration in later contributions; temporal circumstances become displaced by circumstances of veracity (*recently diagnosed* → *correctly diagnosed*). Looking at the configuration of *person + relational process + bipolar*, a shift from *being* to *having* was uncovered.

8.2. Addressing Question 2: Discourse-semantics and register

With an understanding of which words and wordings undergo change over the course of membership, it becomes possible to discuss what these words and wordings *mean* and *do*—that is, to summarise the semantic and discursive content being realised through the lexicogrammar.

Most existing linguistic accounts of OSGs have centred on analysis of the discourse-semantic stratum of language and above. For this reason, it is possible to connect many findings from the case study to key claims made in OSG literature. As noted in Chapter 2, however, little OSG research has drawn upon the systemic-functional framework, which differentiates between discourse-semantics as the most abstract part of the content plane, and register as a component of context—that is, the situation type in which the text is produced (Halliday & Hasan, 1989). Accordingly, it is necessary at this point to determine the most suitable heading under which to treat key themes uncovered during the analysis. For the purposes of this discussion, therefore, phenomena such as *representation of social actors*, *advice*, *references to*

(*in*)stability or the *construal of diagnosis* are treated as primarily discourse–semantic; member roles and identities more generally are registerial.

8.2.1. Discourse-semantics in the Forum

During the previous three chapters, lexicogrammatical results were generally presented alongside a brief explanation of their discourse–semantic significance, with findings organised by lexicogrammatical feature, and separated along metafunction lines. The main limitation of this structure is that the discussion of meanings being made by words and wordings becomes disjointed: meanings are not compartmentalised within different grammatical systems and subsystems; rather, meanings are made through the deployment of elements of the grammar together in text as it unfolds. In fact, language is structured around meaning, and simply realised through words and wordings (Halliday, 1978). The progression in each of the previous two chapters through increasingly delicate components of grammatical systems therefore does not allow sufficient space for a discussion of the meanings made by hierarchical or agnate linguistic structures. At the same time, there are a number of moments where meaning-making transcends the systemic division of metafunctions. The division of the analysis into genre, MOOD and TRANSITIVITY chapters makes this kind of multifunctionality difficult to present. To coherently discuss meaning-making in the Forum, therefore, a discussion organised semantically is now required.

Social actors and interactants in the healthcare journey

A useful starting point in a summary of discourse–semantics in the Forum is to broadly characterise the kinds of social actors and interactants that occur, and the kinds of processes they are construed within, without respect to longitudinal change (Van Leeuwen, 1996). By collapsing participant heads into four groups, the TRANSITIVITY analysis differentiated between four main categories of social actor: *the Self*, *Other Members*, *Health Professionals* and *Friends/Family*. These participant types are construed within different processes: the *Self* is very commonly a *Senser* (engaging in *appreciating*, *thinking*, *guessing*, *hoping*, *knowing*), but is rarely construed as the

Sayer. Health professionals, on the other hand, do a great deal of saying, telling, and asking—in fact, they are involved in verbal processes more than twice as often as any other participant type. Of course, this distribution of Process Type over participant type aligns with any sensible expectation of how consumer health discourse serves to represent the stakeholders in healthcare journeys. Because language mediates most critical events in clinicians' treatment of mental health issues (diagnosis, prescription, therapy and directives, to name just a few examples), health professionals are often construed as Sayers. Friends and relatives are unsurprisingly construed relationally, as their relevance within Forum users' medical narratives is largely determined by their relationship to the speaker. The Self is construed more often than other participant types as thinking because people only have direct access to their own cognitive processes; others' thoughts must be inferred from their verbal, behavioural and material actions.

The way these four types of Things are construed undergoes change over the membership course. First, in terms of their overall prominence as experiential participants and interpersonal Subjects, there is a steady decline in the relative frequency of first person, and a steady increase in the relative frequency of second person/references to Other Members. Interpersonally, this can be analysed in two ways. First, it can be seen as a shift from users charging themselves with modal responsibility to charging their addressees, over whom they have a kind of authority conferred by their membership stage. In this analysis, users make cases for their own membership within communities (Stommel & Meijman, 2011; Vayreda & Antaki, 2009), and, upon the ratification of their bid, shift toward examining the claims made by later newcomers (Paulus & Varga, 2015). Such a shift would be more dramatic still if the rise in imperative clauses were taken into account: in imperatives, it is the second person, despite being absent in the lexicogrammar, that is held responsible for meeting the speaker's demand. The second, complementary analysis is that the interpersonal burden is always on newcomers—as a member exits newcomer membership stages, he/she begins to charge newer members with the interactive demands that he/she has recently shed. The newcomers, in occupying the Subject role, are always the dominant 'resting point of the argument' (Halliday

& Matthiessen, 2004, p. 118) within Forum exchanges. Ideationally, the world being represented through language is, in the same way, the world of the newcomer: as members progress toward veteran roles, their representations of their own health journeys become less frequent, and shift in purpose, serving more often as exempla from which newcomers can learn about common mistakes or useful strategies in the management of bipolar disorder (Pfeil et al., 2011). To borrow terminology from Bauman and Briggs (1990), users shift longitudinally from entextualisation of the Self toward entextualisation of the Other.

Though the Self and Other Members are on divergent trajectories in terms of relative frequency as participants, both are positioned as Agents to very similar extents (69.31% and 71.19%, compared with Health Professionals and Friends/Family, at 77.68% and 65.78% respectively). This shows us that while propositional validity rests predominantly on the newcomer, and while the newcomer's journey is typically the thing being construed by newcomer and veteran alike, both newcomer and veteran, as social actors in healthcare journeys, are represented as causing change to approximately the same degree. The Self and Other Members occupy distinct roles within the interactions of the Forum, but more or less the same position within the world of healthcare being represented by Forum members. The Self and Other members also undergo very similar increases in the extent to which they are positioned as Agents: in newcomers' talk, both are positioned as media through which processes happen; in veteran talk, increasingly often, The Self and Other members are construed as Things that say, think, act or do. This represents a shifting understanding of the role of the consumer within the healthcare journey, from passive experiencer of healthcare systems to navigator of the journey's trajectory. Simply put, the longer a member contributes to the Forum, the more likely he/she is to construe both him/herself and others as 'active patients'.

Meanwhile, Health Professionals undergo very different kinds of change. In terms of overall frequency, they rise rapidly in the final two stages of membership, indicating an increasing emphasis on formal healthcare institutions and their role in bipolar management in veteran talk. At the same time, however, Health Professionals are less and less commonly positioned as Agents. Importantly, the disparity

in agency between Forum members (i.e. The Self + Other Members) and Health Professionals lessens over time. Such a shift indicates that veterans construe a less hierarchical professional-consumer dynamic than newcomers, and an increasing level of consumer empowerment generally. Rather than being Mediums through which processes operate (*My doctor diagnosed me with bipolar*), veterans are instead more likely to cast the consumer as interpersonally responsible for making the proposition valid, and, simultaneously, as experientially responsible for progression through the healthcare journey (*You should go to your doc and get a diagnosis*). Veteran members therefore tend to advocate a number of current components of the dominant biomedical ideology, such as *the active consumer*, *shared decision making*, and *consumer-centredness*.

Discursive shifts

Other kinds of longitudinal changes in semantics can be identified through shifts in how particular grammatical constituents are typically configured within a text. One of the most striking of these is the way in which users construe the process of diagnosis—a more or less mandatory component of the bipolar disorder journey within the biomedical ideology of the Forum (see Section 7.2.1). *Diagnose* as an Event was found to be a very key process in first contributions, leading to more delicate analysis of its behaviour in the corpus. This analysis turned up an increasing rate of grammatical metaphor (in the form of nominalisation, as *diagnosis*), and shifts in attendant participants and circumstances. Each of these changes reflects a change in the discourse-semantics of the Event. First, the metaphorical realisation of the diagnosis Event as a participant opens it up to possession, deixis and more precise kinds of classification. At the beginning of membership, diagnosis is an Event that the speaker experiences, often without his/her volition, and often without an explicit Actor. The Event is situated in the past, represented as catalysing the new user's decision to sign up or contribute to the board. Diagnosis, as others have remarked, can function as a kind of entry ticket, or marker of legitimacy, for membership in mainstream OSGs (Stommel & Meijman, 2011). Within the overarching ideology of the Forum, treatment (including the talk therapy provided by Forum interaction

itself—see Kaufman & Whitehead, 2016) is reserved for those who have an official mandate that warrants it (Vayreda & Antaki, 2009). As such, it is no surprise that the *diagnose* process is increasingly construed as a Thing that can be given, possessed and inspected. In the community, bipolar disorder is understood to be treatable, but not curable—the amount of time between the diagnosis and the present is therefore more or less immaterial when inspecting the validity of the entry condition. Instead, what becomes important is the veracity of the diagnosis: *suspected* and *unofficial* diagnoses must become *proper* and *official* diagnoses, because it is correct diagnosis that steers the newcomer on the optimal path within the bipolar disorder journey.

The distinction between *being* and *having* bipolar disorder (recall Section 7.2.2, and the shift toward the latter over the membership course, functions differently to the change in the configuration of diagnosis. Employing the *being* form does not put the validity of a membership claim at risk; instead, new members' transgressions of the *having bipolar* norm provide an opportunity for the ideological values of the community to be made explicit (Weber, 2011). More specifically, veteran members can use the moment to semiotically reconfigure *bipolar* into something that can be controlled, and thus, can be managed, be it by medication, visits to a health professional, or the talk therapy that constitutes Forum interaction itself.

Another area of change is the increasing frequency of advice. Advice is best understood as a discourse-semantic phenomenon: it is congruently delivered within individual clauses (or potentially clause complexes), rather than within individual constituents, or across a text more generally. Its deployment may involve explicit lexicogrammatical features (*Let me give you some advice ...*), but is far more commonly signalled through combinations of MOOD and MODALITY choices, particular Process Types, and so on. It is most easily recognised by its semantics, where the addressee is being told that a particular course of action is desirable in the eyes of the speaker. That said, it does have common lexicogrammatical realisations, in imperatives, modalised declaratives, and, more rarely, interrogatives (Locher, 2006). One point of discussion is whether the membership stage of the veteran member may determine the congruence of chosen realisations of advice (DeCapua & Huber, 1995). On one hand, there is indeed a steady growth in the number of imperatives over the

membership course. On the other, hypothetical statements, in the form of modalised declaratives, also become much more common over time. The reason for this is that what develops over the course of membership is not simply an authority over newer members that manifests in more direct realisations of commands—rather, it is a nuanced grammatical repertoire designed to best motivate the newcomer to adopt community values that are intended to facilitate better health outcomes. Advice is a coupling of experiential and interpersonal content: suggested action, but also a negotiation of the power dynamic between speaker and addressee (and the wider audience of potential posters and lurkers) (Harrison & Barlow, 2009, 1). Given that a clause must have interpersonal and experiential value, this is an expectation within any systemic-functional interpretation of text.

A final kind of discourse-semantic change is in general preference for construals of stability over instability, and of positive over negative attitudes. These tendencies are the most diffuse of identified semantic change, and fall far closer to the pole of lexis (or instance), than to grammar (or system). This is to be expected: as Martin and White (2005) explain, the systems of APPRAISAL are not lexicogrammatical, but semantic, in the first place. For this reason, they did not receive sustained treatment from a lexicogrammatical perspective. Even so, the shifts are striking, occurring as very key participants and processes in the first and last stage of membership. The shift toward construals of *stability* can best be understood with reference to the phenomenology of bipolar disorder itself, and with an understanding of the Forum's therapeutic orientation: because bipolar disorder involves oscillation between periods of mania and periods of depression, the achievement of stability (of emotions, behaviours, relationships, etc.) is, in a sense, the point at which bipolar disorder has been successfully managed. As shown in the previous chapter (Table 7.3), veteran members commonly construe stability as the Goal of a material process, with Forum members functioning as the Actor who will eventually *find*, *Maintain*, *reach* or *achieve* it. Meanwhile, the gradual elimination of attitudinally negative lexis reflects an increasing tendency to construe the ideal goals of the healthcare journey, rather than previous or potential future problems. This reinforces to newcomers a conceptualisation of bipolar disorder as manageable through diligent self-regulat-

ory behaviour. The increasing construal of positive action within the community (*hugging, welcoming*) builds a representation of the Forum as a supportive space in which the journey toward stability can be carried out.

8.2.2. Register

In the sections above, I have summarised some key parts of a normative semantics that is drawn upon more and more as users progress toward veteran membership. The task that remains is to connect the identified changes in the content plane of interactions within the Forum to the contextual variables of Field and Tenor.

In SFL, *register* is the stratum above, or *realised by*, discourse-semantics. It is, in effect, a type of situation, where the patterns and probabilities of instantiation of particular linguistic features are set (Halliday & Matthiessen, 2004; Lukin et al., 2011). Following Matthiessen (2013), any given healthcare interaction can be understood as existing at the most delicate end of the cline of instantiation—as an instance, or realisation, of the meaning potential of the linguistic system, related to the journey through a common ideational focus on a health-related issue, and a common interpersonal interactant, the self. By investigating patterns in a corpus, we progress from the (delicate) realisation end to the (broad) system end, locating possibilities and probabilities in the grammar, and relating these to meaning-making across the interpersonal, experiential and textual metafunctions. From the sketch of relevant choices along the cline of instantiation, we can then produce a more abstract register description, covering who is talking, what language is being used to construe, and what language itself is doing within the situation. The case study has already presented a quantitative overview of register, in the form of not only shallow features, but in the more delicate syntagmatic behaviour of some of the lexicogrammatical features that differentiate the Forum register from others. In the sections below, I add a qualitative description of the Field and Tenor of the Forum, with some cautious speculation regarding the unexamined dimension of Mode.

Tenor

In terms of the Tenor of discourse, the Forum is unlike most other possible elements of the consumer journey: unlike formal healthcare settings, where the consumer typically enters into a series of provider-consumer interactions (Slade et al., 2008), the community presents a space for intra-consumer interaction. In comparison to hospitals, members of the Forum are in a relatively equal role-relationship, sharing both an identity (as people living with bipolar disorder) and a goal (effective management of symptoms). Equality is in some respects facilitated by the architecture of the website, and text-based CMC in general, which may obscure a number of demographic distinctions that lead to unequal treatment (Walther, 1996). That said, in line with the work of Herring (2001), the gender, age, and nationality of many members are often easy to disambiguate: many users make their gender visible as a metadata feature that appears to the side of their posts; usernames and linguistic choices can also often communicate gender. Moreover, even though role-relationships are equal compared to typical provider-consumer interactions, Forum members still take on distinct, hierarchical positions within the community. The authority of the veteran over the newcomer, while not absolute, manifests in the lexicogrammatical strategies selected by veterans for advice provision: the modalised declarative *I would* construction is more common than the bare imperative command that may emerge when statuses are very unequal (DeCapua & Huber, 1995). At the same time, the *I would* construction flags the fact that the expertise of the veteran is of a lay, rather than professional type. As noted by Smithson et al. (2011a), the action content of veteran's advice is typically conservative. Most often, advice serves to reinforce mainstream medical expectations: where newcomers deviate from the biomedical norms, veterans offer advice that bolster the health professional's authority (*I would DEFINITELY recommend seeing a psychologist; I would highly recommend that you take your meds on a daily basis.*). Both newcomer and long-term members' interpersonal exchanges can be understood as attempting to negotiate a legitimate position within the social hierarchy of the community (Varga & Paulus, 2014; Koteyko & Hunt, 2015). For the newcomer, there is a legitimate need for

information and support; for the veteran, it is the existing status at the top of the Forum's social hierarchy that is preserved through interpersonal negotiation.

Field

Field is the variable with the most overlap between other components of a consumer's healthcare journey. In the Forum, like in a consultation with a doctor, the consumer is likely to talk about symptoms, diagnoses and treatments; related issues such as family, work and hobbies may also be construed. Even within this Field, however, for different situations, there will still be differences in the frequencies with which particular participants and processes appear, and the ways they are modified and arranged as configurations and activity sequences. In the Forum, for example, a great deal of talk centres on personal relationships and emotional states. This kind of talk is less likely in some healthcare encounters (e.g. a hospital stay), but more likely in others (e.g. within a psychiatrist's office).

The Field of discourse under discussion by Forum members also changes over the course of membership. Users engage in more metadiscourse, opt for jargon variants of common participants, and use vague language to provide advice that can be applied to not just the addressee, but other potential readers. As mentioned earlier, a number of lay and negative Events and Things are gradually displaced with more neutral, scientised language: talk about *mood swings* and *going crazy* declines, while references to *stability* and *balance* rise. Broadly, these changes reflect an increasingly faithful reproduction of the experiential semantics of a biomedical ideology, where the ideal healthcare journey involves a combination of engagement with formal healthcare institutions (for diagnosis, therapy, etc.) and self-monitoring for indicators of potentially oncoming manic/depressive cycles. In line with a biomedical model, the Forum is framed as a space for a kind of talk therapy that can foster stability by allowing people to vent, reflect, or engage in the exchange of information about bipolar disorder. It is an important space, but one that ultimately plays a secondary role to what can be accomplished through medication and psychotherapy: only doctors can diagnose or change medication regimens, but veteran members can encourage newcomers to speak with the doctor about their problems. Where

newcomers deviate from this normative ideology, as shown by Weber (2011), veteran members take the opportunity to make community norms explicit: (*part of the purpose of this space is for venting; we aren't docs and can't diagnose you; you're not *A* bipolar, we're not things, it's a condition*). The function here is to try to align the newcomer with the normative ideology, and the normative illness trajectory that exists within that ideology.

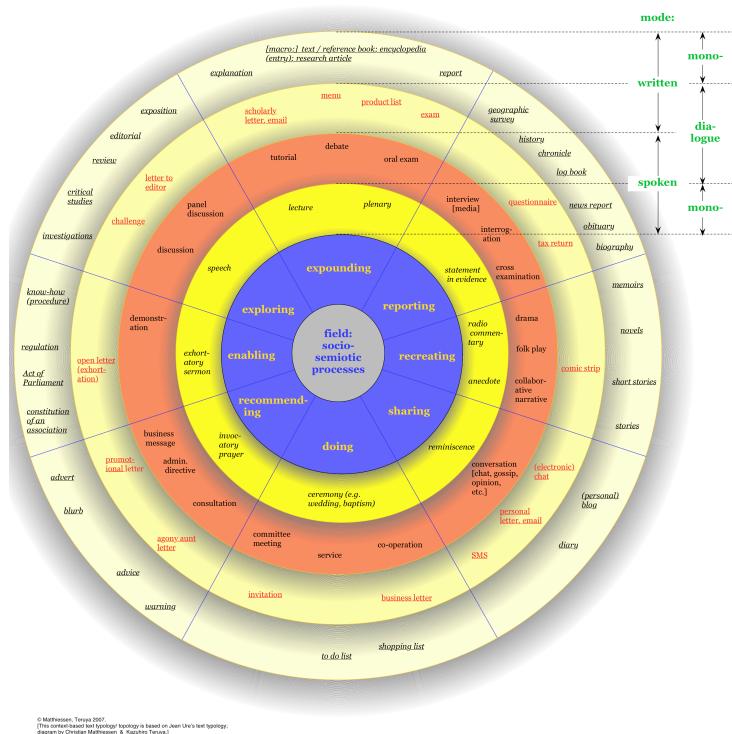


Figure 8.1: Socio-semiotic processes as topology (Matthiessen, 2015)

The overarching changes in Field can be represented as shifts in the kinds of socio- semiotic processes in which users typically engage (see Figure 8.1). New members engage most commonly in processes of *sharing* (their medical histories, current concerns and problems, etc.) and in information or support seeking. Though veterans, like newcomers, construe previous experiences from their healthcare journeys, more commonly, they are not simply sharing; at the same time, they are *recommending*—they provide, from their own lived experience, examples of previous problems and solutions (Kotevko & Hunt, 2015). *Recommending* can, of course, also be more

direct and explicit, with veteran members routinely issuing demands for material (*find a doctor*), verbal (*communicate with your doctor*) and mental (*consider it a learning experience*) action (Pfeil et al., 2011). It is important to remember, however, that a community of the size and scope of the Bipolar Forum is bound to become a space for a diverse range of socio-semiotic processes. Socio-semantic processes of all types were seen during qualitative exploration of the Forum: *enabling* can be seen in the sticky posts that list community guidelines, *exploring/expounding* are used to reflect on recent news articles and research about bipolar disorder; some users engage in *recreating* by writing poetry about their inner and outer healthcare journeys.

The healthcare journey

We also need to account registerially for the portion of the healthcare journey that takes place within the Forum itself. As shown throughout the case study analysis, over the course of membership, members' responsibilities come to complement those of the health professional, performing elements of the process of healthcare that cannot be carried out in clinical settings. Users encourage the undiagnosed to schedule consultations and change professionals if necessary to facilitate diagnosis and the beginning of treatment. At the same time, users gradually step back from encroachment on the role of the health professional, instead choosing to provide more general information about the effects of bipolar disorder, as well as its symptomatology and possible treatments, on everyday life. This decreases redundancy across the clinical and online treatment spaces, and limits the potential for conflicting advice. In first-post threads, veteran Forum users thus engage in a kind of *new-user-centred care*, where jargon terms are avoided or glossed, and where the new user's sense of inclusion and wellbeing, and the necessity of a continuing exchange, are the overarching concerns of the text.

8.3. Critical reflection on the investigation

With a completed discussion of the findings of the case study and their relationship to earlier literature, it becomes necessary to reflect critically on the investigation. In

the remainder of this chapter, I reflect on three kinds of shortcomings, each of which limits the explanatory power of the case study investigation. First are potential kinds of analysis that were not undertaken due to limitations of scope. Second are issues relating to choice of the Bipolar Forum as the dataset under investigation. Third are the affordances of the theories and methods reviewed in Chapters 2 and 3 as theoretical assumptions and applicable methods for the central aim of the case study, which was to investigate linguistic change over the course of membership in an OSG.

8.3.1. Unexplored linguistic phenomena

It is not possible, in a single thesis, to account for language across three metafunctions, and multiple ranks and strata, in a corpus of over eight million words, containing nearly 6,000 unique speakers and around 9,000 texts (i.e. threads). Though the approach involved systematic progression through metafunctions and the cline of instantiation, many lexicogrammatical features (and thus, the meanings they realise) went unexamined. In some cases this was simply due to constraints of time. In others, however, the main limitation is what is encoded, or searchable within, the automatic parses provided by **Stanford CoreNLP**. Below, I list a number of unexamined components of the SFG, each of which is responsible for realising important kinds of meanings.

Ellipsis

No attempt was made to identify or reintroduce elided components of clauses. Because elision is uncommon in formal texts, its existence in more casual interactions such as those within the Bipolar Forum presents a major obstacle for parsers, which are generally not trained to handle it.³³ Treatment of elision is necessary to analyse the meanings made in minor clauses especially (e.g. greetings, exclamations), which by definition are lacking full Mood elements, and which play important roles in the negotiation of role relationships. Extents and types of elision may also be useful statistics for approximating shifts in formality of language over the membership course, and could also help gauge parser accuracy.

Modulation

Modal Finites are not the only possible means of encoding modality: modal adjuncts, such as *certainly*, *possibly* or *of course* can make the same meanings, either independently or in combination with modals (Eggins, 2004). Unlike modals, however, there are hundreds of possible realisations of modal adjuncts. For this reason, modulation was not investigated to the fullest possible extent. In principle, however, nothing other than time constraints prevents the development of wordlists that group modal finites with modal adjuncts based on meaning, collapsing realisations of obligation, inclination, probability and usuality. This, alongside sensitivity to polarity and ellipsis, would lead to both a fuller understanding of the Bipolar Forum as a site of interpersonal exchange, and toward a computationally viable model of interpersonal semantics.

Group structure

Configurations of group heads was the main focus of analysis, both within the MOOD and TRANSITIVITY analyses. In part, this is due to the orientation of dependency grammar toward hierarchical lexical relationships instead of grammatical constituents. Group structures, however, may provide insight into how both Things and Events are discursively framed or appraised. It could have been useful, for instance, to more closely investigate the internal structure of both nominal and verbal groups. An analysis of Classifiers in nominal groups representing a health professional, for example, could lead to empirically informed taxonomies of possible participants. Epithets, on the other hand, could be used to uncover shifts in sentiment toward particular entities and happenings over the course of membership. Also within nominal groups is the system of DETERMINATION, which provides an alternative, unexplored means of relating illness to the self and to others via ascription or possession (*the bipolar* vs. *my bipolar*).

Verbal groups (generally realising processes), from a TRANSITIVITY perspective, are comprised of Finites, Auxiliaries and Events (Halliday & Matthiessen, 2004). Analysis of combinations of these components could lead to an understanding of how particular Events are construed in terms of their frequency or probability, or

in terms of their temporal relationship with the time of writing. Identification of phases of illness, as performed by MacLean et al. (2015), or the automatic generation of medical timelines for clinical use (Raghavan, 2014) could be aided by this kind of analysis, where processes of *being diagnosed*, *taking medication* and/or *seeing a doctor* could be situated on timelines based on finiteness, aspect and tense.

Mode

The case study did not involve an analysis of the register variable of Mode—that is, of how posts function as internally coherent messages, and as meaningful components in texts (i.e. threads). This decision was not a random one: Mode was expected to be the register dimension undergoing the least change, because the interface for writing messages did not change significantly over the period from which posts were collected, and because the popularisation of mobile access to the social web, which likely affects communicative practices in significant ways, began for the most part after the Forum’s peak in popularity. That said, some linguistic components of the dimension of Mode may be at risk over the course of membership. Semantically, because Theme choices mark a psychological focus in the discourse (Halliday & Matthiessen, 2004), an investigation of common Themes could do much to build an understanding of the structure of the text being produced. Based on the changes identified in Chapters 5–7, it could also be assumed that changes in the overall semantic complexity of messages may be realised by conjoined and embedded clauses. For this reason, Theme is a rich lexicogrammatical category, which can operate with independence from the grammatical structure of the proposition via thematic variation. In the case below, a user foregrounds his wife’s manic phases, the problems they cause, and the obviousness of their final outcomes, via marked Theme choices:

Part of my wife's problem is that her family completely enables her running away.

When pregnant she moved our whole 5000 sq ft house back to the old house that hadn't sold yet while I was at the office [...]. **Needless to say** I had to hire movers to move it all back 4 days later and her parents said nothing.

She always runs when manic [...]. **Of course** she then moves at the speed of light to end the relationship but **low and behold** crashes soon after saying she's sorry and wants to work it all out [...]. **When cycling back down** the relationship at least for my wife is a safe secure place for her and she wants to run back everytime.

I think it really comes down to dealing with 3 different people. The manic BP, the stable BP and the depressive BP. **When my wife's manic rocket boosters started to fire** it was amazing to see how she changed. **What mattered when she was stable** was all but out the window. [...] **The main problem was** I was the only one questioning her actions. I was of course the closest but no one around her would dare question her especially her family. **When manic** she never realized how much of a solid foundation I provided her. [...] **The problem is** I'm 1 person not 3 [...] .

It is therefore sensible to suggest that Theme choices could be exploitable for a number of analytic purposes, such as the location of adverse medical events and their causes (Chee, Berlin, & Schatz, 2011), or even delineate between subtopics within the board. Practically speaking, Theme is relatively easy to extract from a constituency parse, as its primary criterion in English is its location as the leftmost group within a clause. Already, much computationally driven research into mental health communities has used Mode phenomena to identify the current mental state of writers. Analysis of cohesion could play a key role in detecting shifts between illness states for individual members, as performed by MacLean et al. (2015). In fact, accurate semantic parsing, in needing to make connections between clauses in a clause-complex, is highly dependent on correct identification of given/new information within clauses.

8.3.2. Data source issues

Most CL research has focussed on general corpora, in order to make claims about a language generally. Also prevalent have been comparisons of different genres and registers within general corpora. This case study, however, presented a corpus that was largely homogeneous in terms of register. This has benefits and drawbacks. In terms of benefits, more homogeneous corpora make it easier to isolate an individual variable and determine the extent to which it is responsible for variation within the corpus as a whole. In this case study, lexicogrammatical and discourse-semantic changes can be easily linked to membership length, and compared to the evolution of the community itself. Another advantage of using a corpus from a single source is that it is possible to probe more delicate lexicogrammatical features that cannot be found in sufficient quantity in general corpora (Zinn & McDonald, 2015). At the same time, however, the homogeneity of the corpus can lead to potential issues in

terms of generalisability and representativeness. Though changes over membership and over time can easily be detected using the methods developed here, we can only hypothesise that such changes are common to other communities. In fact, given that other OSGs have explicitly different ideological orientations (pro-ana forums stand out as an example, as they run counter to mainstream medical opinion), we should expect differences (in the construal of health professionals and health institutions, for example). Accordingly, rather than concluding that certain kinds of language change are common over the course of membership in an OSG, we might instead conclude that changes appear to be normative.

When a single OSG is the sole dataset under investigation, it is very difficult to map findings to those generated in clinical settings MacLean et al. (2015). There are a number of reasons for this. First, unlike in face-to-face, clinical encounters, many demographic features cannot be collected and/or verified. Furthermore, we do not know where users go after ceasing activity within the Forum, or what caused their departure. Finally, it is difficult to use the dataset to draw conclusions about the general public: most members of the Forum have sought out a community to meet their needs, and as such, are likely unrepresentative of people living with bipolar disorder generally.

Small sample of users in veteran stages

Given that the final few stages of membership in the Membership Stage Structure represent the language use of a very small number of speakers (see Table 4.1), it is possible that what is being modelled as veteran user discourse could be better attributed to the roles and role-relationships developed by this particular set of users only. To account for this, brief investigations of the Comparative Structure and Future Veteran Structure were carried out. These showed that in most cases, there is little difference between Future Veteran members' early contributions and the early contributions of users who drop out early. That said, spatial constraints precluded a more exhaustive comparison of the corpus structures.

The fact that the 10th subcorpus contains only eight distinct users has a number of influences language that act as confounding variables when attempting to answer

Research Questions 1 and 2. Proper noun nicknames and pseudonyms for veteran users, for example, appear as key participants in later stages of membership, as veterans often address one another directly, or conclude their posts with a kind of informal signature. Additionally, some of the veteran members do not have bipolar disorder themselves, but enter the community to discuss a loved one. The names of loved ones, as well as their relationship with the member (*daughter, husband*) are represented as more common in veteran discourse than would be the case with a different, or larger, set of veteran users.

Unit of analysis

The case study centred on texts authored within ten artificially defined ‘stages of membership’. Despite the fact that the corpus could be symbolically restructured in a number of ways, almost all of the analysis centred on The Membership Stage Structure, where posts formed the unit of analysis, and where distinctions between individual participants were not made. One dimension lacking from the study, therefore, is a focus on specific users as they progress through the community. Approaches centred on individual user trajectories (such as those proposed in Chancellor, Mitra, & De Choudhury, 2016; Chancellor, Lin, & De Choudhury, 2016; MacLean et al., 2015), make it possible to identify events within individual medical timelines. This can be more easily mapped to clinical outcomes for individual healthcare consumers.

8.3.3. Theoretical and methodological challenges

The final part of this chapter addresses obstacles posed by the theoretical and methodological choices made throughout the thesis. These problems range in severity. Problems such as parser accuracy, for example, are problems that will likely be ameliorated to some extent by advances in NLP. Other problems, such as the issue of how to computationally model register, while having a strong theoretical foundation, have not been fully explored or detailed in existing research. The most serious problems are those where established theory or practice inhibits explanation of a phenomenon. The established categories with which tokens are annotated by a dependency grammar, for example, may not be accurate or appropriate for CL tasks.

Another example is the limitation of SFL as a theory to account for the meaning-making practices of individual people.

Parsing

A major set of challenges encountered during the thesis is those related to automatic parsing. First, there is the issue of accuracy. I used the off-the-shelf **CoreNLP** parser models (v. 3.6.0) to annotate the corpus. The models are not designed for CMC or OSG language; the parser was trained and evaluated on formal, well-edited text types such as news journalism. As such, parses are often incorrect, especially when language use deviates from what may be expected in formal, written registers/text-types. Sentence-initial vocatives and salutations, for example, are often mis-annotated as Subjects. In imperative clauses, leftmost words or constituents (either Adjuncts or Predicators) are also often analysed as Subjects, leading to inflated counts for indicative Moods. Ultimately, however, future developments in NLP will likely see improvement on many of these tasks. Moreover, the issue can already be solved by the (albeit resource-intensive) step of training a parser model that better fits the data under investigation.

The more serious challenge posed by parsing is the extent to which useful functional information can be extracted from parser output (typically, constituency and/or dependency grammar annotations). First is the constituency grammar in use—a Head Driven Phrase Structure Grammar (HPSG). Given that this grammar evolved from the formal, non-semantic, generativist tradition, it is sensible to question the practice of using its annotations to gain functional insight into texts at all. Indeed, labels such as NP and VP do not correspond particularly well to a particular component of semantics. That said, some parts of the constituency representation can be exploited for functional purposes, especially those parts corresponding to interpersonal meaning, made via choices of MOOD and MODALITY. As shown in Chapter 6, queries can be written to match Indicative and Mood Type by looking for the existence, order and hierarchy of NP/VP that descend from clause-level constituents. These can then be mapped to interpersonal semantic dimension of Speech Function. At the lower ranks of word, group and phrase, there is also great

deal of overlap between phrase structure grammar and SFL. The notion of *headness* is also valuable for locating Events and Things within verbal and nominal groups, respectively. Points of overlap between Chomskyan and Hallidayan traditions, however, are rarely pointed out, due to mutual incompatibility of the frameworks in terms of their overall conceptualisations of what language is, where it came from, and what it does.

Dependency parsing and discourse-semantics

Dependency grammars have become the de facto standard grammar for automatic parsers (de Marneffe et al., 2014). The Universal Dependencies project aims to standardise dependency representations further, providing a set of labels and definitions that can be applied cross-linguistically (Nivre, 2015). These labels are now used by many parsers, including **CoreNLP**. A great strength of dependencies is in providing simple access to meaningful functional information (De Marneffe & Manning, 2008). As seen in the previous chapter, extraction and counting of dependents in `nsubj` and `dobj` types, for example, is often enough to summarise the topic of a text or corpus. Notably, however, dependency grammar is rarely used for qualitative linguistic research into English texts. A major part of the reason for this is that the grammar is too superficial for close-reading of texts. By design, the Universal Dependencies aim to maximise the cross-linguistic interpretability of labels while minimising the number of unique dependency types, so that the parser output is useful to those without training in linguistics (de Marneffe et al., 2014). The grammar is also intended to be applied as a single layer, so that each token can only be annotated as having one dependency role. This contrasts with SFL, which makes use of a layer of (partially redundant, overlapping) annotations for each metafunction. As such, dependencies struggle to elucidate subtle distinctions that run across metafunctional lines, such as the distinction between what Halliday and Matthiessen (2004) refer to as the experiential Subject (*Agent*, within the ergative model of TRANSITIVITY) and the Grammatical Subject.³⁴ To give an example of the problem (using a classic example from the IFG: Halliday & Matthiessen, 2004, pp. 53–58) in the Universal Dependencies, despite an overall orientation toward

experiential meanings, participant heads are given different labels based on the voice of a clause, though their experiential role remains the same (Figure 8.2).

<i>the duke</i>	<i>gave</i>	<i>my auntie</i>	<i>this teapot</i>
nsubj		iobj	dobj
<i>my auntie</i>	<i>was given</i>	<i>this teapot</i>	<i>by the duke</i>
nsubjpass		dobj	agent

Figure 8.2: Annotation of participant heads in agnate active/passive clauses using the Universal Dependencies specifications

Another shortcoming is that Universal Dependency grammar does not distinguish between the Goal in a material process and the Range in a process–Range configuration (Halliday & Matthiessen, 2004): *shower* is analysed as the **dobj** in both *I cleaned a shower* and *I took a shower*. The process being undertaken in the second example is not one of *taking*, but of *showering*. This makes it very difficult to accurately recover experiential configurations of processes and participants.

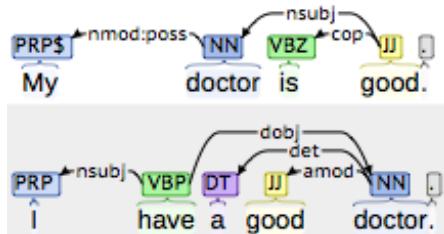


Figure 8.3: Dependency relationships

Label definitions also switch between syntactic and semantic criteria, usually to ensure that central labels such as **nsubj** have salient lexical dependents, and for the purpose of congruence with analyses of other languages (de Marneffe et al., 2014). One important example is the **root** node, which generally marks the main verb, unless the main verb is copula, in which case **root** is shifted to the head of the grammatical Complement. A second example is that **nsubj** is shifted to the grammatical Complement in existential clauses, to avoid marking the experientially empty *there* (Halliday & Matthiessen, 2004) as the most prominent participant in the clause. Potential issues arising from these decisions can be seen in Figure 8.3. In the

first example, *doctor* is problematically analysed as the dependent subject of *good*; in the second example, *good* becomes dependent on *doctor*; in the third example, *doctor* is analysed as **nsubj** despite not being a grammatical Subject (de Marneffe et al., 2014). Moreover, the criteria are variously syntactic and semantic in and of themselves. The **nsubj** definition is syntactic, while the **dobj** definition is semantic (see Nivre, 2015):

- nsubj**: A nominal subject is a nominal phrase which is the syntactic subject of a clause.
dobj: The direct object of a verb is the noun phrase that denotes the entity acted upon.

These inconsistencies are designed to make it possible to quickly extract salient tokens from texts, and to skip words with little experiential meaning in isolation, such as *is* and *there*. For more nuanced functional–semantic work, however, such inconsistencies become stumbling blocks that need to be accounted for during corpus querying with verbose sets of conditional rules. While the ergative model of TRANSITIVITY as outlined by Halliday and Matthiessen (2004) provides a potentially useful means of measuring the extent to which participants are construed as Agents in key processes, the grammar was difficult to operationalise here due to ambiguities in the constituency and dependency grammars. Moreover, the case study demonstrates that critical experiential meanings are made though apparently mundane lemmata such as *be* and *have*. The analysis of the ways in which attributions and ascriptions of bipolar disorder vary longitudinally, for example, underscores the importance of common words and delicate grammatical distinctions in both the construal of experience and the production and reproduction of a normative community ideology. This issue demonstrates the need for dedicated parsing at the stratum of discourse–semantics—a future possibility described briefly in Chapter 10.

Limitations in available systemic-functional resources

For the TRANSITIVITY analysis in Chapter 7, Events were identified by searching for words in grammatical positions that correspond with Events in the SFG. Simple wordlists were then used to group the Events into Process Types. Such an approach does not account for the fact that the same verb can realise multiple process types (*I feel sick/I feel leather*—recall Section 3.3). Distinguishing between Process Types is

a notoriously difficult task, with poor inter-rater reliability even for trained human annotators (Gwilliams & Fontaine, 2015; O'Donnell et al., 2009). Adding to the difficulty of this task is that two competing descriptions of Process Type exist within SFL, generally known as the Cardiff and Sydney Grammars respectively (Costetchi, 2013). Wordlists were drawn from the Process Type Database (Neale, 2002), which uses the Cardiff Grammar, while the theoretical framework used was the Hallidayan Sydney Grammar (as described in e.g. Eggins, 2004; Halliday & Matthiessen, 2004). As such, behavioural processes were not identified or differentiated from the Process Types that subsume them in the Cardiff Grammar. More complete computational resources for Process Type identification as per the Sydney Grammar, such as recent work outlined by Matthiessen (2014), remain to be put under version control, and packaged and distributed in a useful computational form.

The limits of lexicogrammatical querying

A key limitation of the developed methodology in general is that searching lexicogrammatical information with the aim of learning about discourse-semantics can ultimately only provide a partial account of meaning-making. This account is also biased toward text with a greater degree of congruence between wording and meaning: an interaction rich in politeness marking, or a highly nominalised piece of science writing, pose additional challenges for researchers interested in semantic phenomena. For the analysis of the Bipolar Forum, the major strategy used to quantify semantic information was to query progressively more delicate components of the lexicogrammar. At a number of points during the case study, however, it was not possible to further extend the delicacy of queries in order to disambiguate the meaning or function of a clause. The best example of this issue is in the analysis of *I would + Adjunct* declaratives (see Section 6.2). Here, the same lexicogrammatical feature set can index very different meanings and functions, based on co-text, context and the experiential likelihood of the construal. Automated lexicogrammatical querying, therefore, cannot easily distinguish between the longitudinal changes in the typical function of the construction. This is a known property of modalised declaratives and proposals more generally: in order to make a command

discretionary for the addressee, the speaker must shift to the indicative Mood to allow modalisation. Doing so creates ambiguity between proposal and proposition (Halliday & Matthiessen, 2004): *I would occassionally go to sleep for as much as 24 hours* describes past behaviour, but in another context, could grammatically function as advice. It must then be disambiguated, either through human interpretation of the semantics of the clause (*Is this plausible as advice?*), or by considering the neighbouring moves within the interaction (it would, for example, be unusual to dispense advice in the middle of a recounting of a past event).

Using existing tools to automatically resolve this kind of ambiguity is difficult. The query could be extended in delicacy and complexity to capture co-occurring features such as a second person pronoun in the Complement, which would increase the likelihood that the intended function is advice (*I would go back to your doc*). This approach is unlikely to be accurate, scalable, or reusable, however. Other approaches, such as supervised automatic classification, involve manually labelling a sample of instances to use as training data. This is more scalable, but can be a resource-intensive and computationally complex undertaking. For now, this kind of ambiguity in English lexicogrammar demonstrates the continuing necessity of manual classification of corpus instances. Looking forward, however, it seems that the most logical way to address the issue is to build tools that annotate the semantic stratum directly. This possibility is briefly described in Chapter 10.

Because semantic information could not be directly accessed through the developed methodology, the linguistic sites of change identified over the course of the case study only form a small fraction of the total set of discourse-semantic meanings at risk over the membership course. The extent to which the findings capture the total meaning potential of the community is also constrained by limitations of scope: only a handful of linguistic phenomena, chosen mostly from very frequent features in the lexicogrammar, were analysed in detail. Other parts of the meaning-making process may have therefore escaped attention for any number of reasons. First, some less common meanings at risk may have been buried within very large lists of results. Second, some patterns in meaning are realised through combinations of lexicogrammatical features that were not searched during the case study: though it

is possible to calculate keyness for entire nominal groups, or for key Thing/Event pairs, these strategies were not pursued. It also needs to be kept in mind that some aspects of meaning were filtered out of the texts before analysis had begun, during the corpus creation process. Two examples are emoticons and hyperlinks, both of which are capable of doing important pragmatic work (Koteyko & Hunt, 2015; Schandorf, 2013).

Rank shift and grammatical metaphor

Another encountered issue was rank shift—that is, the realisation of a meaning at an incongruent stratum (Halliday & Matthiessen, 2004). If a researcher is interested in the ways in which *pdocs* are appraised in the text, for example, he/she may begin by finding all adjectival modifiers within a nominal group headed by *pdoc* (as in Example 1, below). Similar kinds of appraisal, however, can also take place at clause level (Example 2), through embedded clauses (Example 3) or at the level of clause complex (Example 4).

- | | |
|---|---|
| 1. My really lovely <i>pdoc</i> helped me out. | 3. My <i>pdoc</i> , being really lovely, helped me out. |
| 2. My <i>pdoc</i> is really lovely and helped me out. | 4. My <i>pdoc</i> helped me out. She's really lovely. |

In each case, more complex kinds of interrogation are required in order to map *really lovely* to *my pdoc*. Moreover, even the most nuanced querying cannot lead to certainty in statements regarding the ways meanings are typically made, as potential realisations of the appraisal, and thus, potential lexicogrammatical queries, are limitless. As a result, in this investigation, and most CADS more generally, the discourse being analysed is for the most part only that which is realised congruently. Given the high frequency of incongruent realisation in general, this is indeed a serious concern. Furthermore, according to SFL, meanings made through rank shift or incongruence are deliberate speaker choices, with significant interpersonal, experiential and textual motivations. If they are systematically not located by interrogation queries, certain kinds of meanings may go consistently unseen. This limitation becomes more serious again when we consider the role-relationships under investigation in the Bipolar Forum. Unequal role relationships result in the need for newer members

to *disperse* the MOOD system when issuing commands: demands for information are dressed as offerings of information:

I would appreciate any advice

I wonder if this is normal!?

This point is also made by Slade et al. (2008), in the context of the ways in which healthcare consumers may relate symptoms and illnesses to themselves:

When a person describes a disease or its symptoms, they can describe it in a number of ways; as part of the goings-on in their external world (i.e. material or behavioural process) as part of the goings-on in the person's internal world (i.e. mental process), as something that they own (i.e. relational possessive process) or as a characteristic or something that can be related to some other thing (i.e. relational process) (2008, p. 290).

The case study charted attribution of *bipolar* via relational processes, with the afflicted occupying one participant role and the disorder occupying another. Other kinds of ascriptions (e.g. *my bipolar, this condition that I have*), were ignored. While nothing stops the researcher from searching for multiple kinds of realisations of a given meaning, iterative developing and checking of query accuracy can be a time-consuming process. At the same time, it can lead to unmanageably large sets of disparate results. Moreover, these results cannot be compressed without oversimplifying how language works. When investigating the *diagnose* process (Section 7.2.1), it was found that the most common modifiers are temporal. The nominalised form, *diagnosis*, however, more commonly selects modifiers of veracity. How, then, can we combine these two results, and quantify all modification of the semantic unit of diagnosis (which may be either process or participant)? One solution would be to turn adverbial modifiers into their adjectival variant (*accurately* → *accurate*), and then to merge the results. This quickly runs aground, however: because *diagnose* is more often realised (congruently) as a process, the frequencies may now be biased toward the congruent. To add another complication, the circumstances selected by the *diagnose* process are not static, but vary by membership stage. Therefore, while it is possible to say that diagnosis is increasingly framed with respect to its veracity (because veteran members modify the process that way, and because they increasingly use the participant form, which co-occurs with veracity modifiers generally), it is not yet possible to describe the semantic unit of diagnosis accurately within a single, two-dimensional representation.

A final, related issue is that grammatical categories operationalised as search criteria, even within a functional grammar, do not necessarily correspond with their sociological significance. Van Leeuwen (1996) reminds us, for instance, that there is a difference between the sociological and grammatical conceptualisations of agency. First, not all agentive constructions imply agency in the sociological sense. Concordancing of first person Agents in first posts shows many examples of the disparity (*I can never make decisions and i worry all the time about everything; I just wish i could feel one way and not so many different ones if that makes any sense; once they gave me a small electric shock i screamed and i couldn't go on with the test*). Second, agency in the sociological sense is not exclusively construed through grammatical agency: possessives and preposition phrases, for example, can accomplish in the agency sociological sense (*my decision was to discontinue lithium*).

The influence of genre

When comparing the results of Chapter 5 with those of Chapters 6 and 7, it is clear that genre, while playing a central role in setting the probabilities for lexicogrammatical and discourse-semantic features of genre stages, is not well-captured by the developed corpus linguistic methods. When new users initiate a ‘first post thread’, for example, the organisation and content of the post may conform to a generic structure related to what has been found in investigations of other OSGs. As in Varga and Paulus (2014), posts frequently provide a narrative medical history. Encounters with formal healthcare institutions, and the diagnosis event, are then presented, in line with previous studies of bipolar forums (e.g. Vayreda & Antaki, 2009). As found by Horne and Wiggins (2009), users may then go on to describe a current problem. In the Bipolar Forum, these problems are diverse, including symptoms, side-effects, relationship issues, work issues, or problems with health professionals and/or institutions. Finally, users commonly make a demand for information or support on other members. Given the low social status of newcomers, these requests are realised by any of a range of agnate linguistic structures: they may be overt or relatively implicit, with statements hinting at a need for further information or social support often standing in for more direct requests for advice. An element

of recursivity in the generic structure allows users to contextualise and formulate multiple problems that require responses from others.

The influence of genre on lexicogrammar can be seen at a number of points throughout Chapters 5–7. In terms of shallow features, the first few posts to the Forum were found to be in many respects more formal (longer words; higher proportion of nouns; higher lexical density) than those that follow. Lexically, *diagnose* as a process is very key in first posts, appearing within the ORIENTATION stage, as many new members frame their entry in terms of a recent or possible diagnosis. *Anyone* features as a prominent participant in the REQUEST stage as a means of soliciting contributions from others (*Can anyone explain this to me?*); *thank* and *appreciate* become key processes in second and third posts as users follow up to the responses generated by their first. Though the metadata embedded in the parsed corpus contains information regarding the thread, and the position of the post within it, how to use this data to either investigate or control for generic influence remains unknown.

Multimodality

Though the original HTML is preserved, making it possible to view any search result in something very similar to its original context, corpus interrogation typically involves counting so many strings of text that contextualised reading of each is infeasible. When analysing plain text without accompanying metadata, a researcher may miss key semiotic features that shape the way meanings are made. Throughout the case study, though some member quotes were presented multimodally where possible, no systematic analysis of the influence of these features was performed. Given that join date, post-count, and a visualisation of post-count as a green ‘progress bar’, are the most obvious ways in which membership stage is communicated, it is certainly a limitation that the thesis did not provide an account of how these features may work to reinforce and/or negotiate speaker roles and identities.

It should also be noted that the well-known criticism of CL methods as obscuring (multimodal) context are not inherent to the approach, but are in fact the result of limitations in software design. `corpkit`, for example, can show arbitrary metadata

features for a sentence alongside its representation within a concordance. Recent work in multimodal SFL (e.g. Hiippala, 2015, 2016; O'Halloran, E., Podlasov, & Tan, 2013) has demonstrated the possibility of extending such methods to allow display of the original, multimodal texts themselves. In this approach, XML documents store page layout, audiovisual information and text in such a way that the content remains searchable using CL methods. Looking forward, it would be feasible to link parsed versions of texts to their copies within the multimodal corpus, so that researchers could switch between monomodal and multimodal views of corpus interrogations. Searches could also be restricted to texts co-occurring with a given multimodal feature, as `corplib` allows with other metadata features.

SFL, the individual, identity and the mind

The final part of the research design requiring critical reflection is the use of SFL as a theory that models ontogenetic language change. As explained in Section 3.3, SFL encompasses both a grammar (in the form of the SFG) and conceptualisation of the relationship between language, text, language users, and context. In many respects, the grammar provided by SFL—the SFG—provides an exemplary framework for dividing language use into metafunctions, and for relating metafunctions to semantics and wordings in a systematic way. It is particularly useful in a CL context, where the investigation must proceed in some logical order (here, along metafunctional lines, and toward delicacy on the instantiative cline), and where searching must be lexicogrammatical, even when a primary interest is discourse-semantics. As a theory of language, however, SFL³⁵ is centred on modelling *language as social action* (Halliday, 1978): it explains how language accomplishes meaning, in the form of the negotiation of relationships (Tenor), the construal of experience (Field) and reflexive self organisation (Mode).

At issue, first, is that the Tenor dimension of register does not model the identity of individual speakers within an interaction. Rather, the Tenor of discourse is the *sets of role-relationships*, as interpreted from social semiotics (Halliday & Hasan, 1989). In systemic terminology, it is therefore not accurate to say that a Forum user's register shifts over the membership course. Instead, a language user, while gradually

shifting from newcomer to veteran status, simply enters into interactions within the Forum, and during each, makes linguistic choices that reflect his/her relative position within the social hierarchies of the interaction and the broader context of the board.

Because SFL does little modelling of the individual or his/her identity, it also makes little effort to engage with the cognitive processes that bring about speech—exceptions here include Crocker, Demberg, and Teich (2016) and Degaetano-Ortlieb et al. (2014). In fact, Matthiessen (1998) advances an argument that *the mind* is, for the most part, a linguistic construct grounded not in cognitive science, but simply in lay perception of the world:

But ‘what is the mind’—or more explicitly, what kind of experience is construed as ‘the mind’, and how is it located relative to related categories of experience? The most general answer is that ‘the mind’ is a linguistic construct—a category of our experience that we construe for ourselves by means of our everyday lexicogrammar (1998, p. 327).

Ultimately, therefore, SFL can be characterised as an historical–material interpretation of semiotic systems, with the individual and his/her mind understood as little more than the set of Tenor roles he/she has filled within the corpus. Identity as conceptualised in most of linguistics and beyond, as a distinct variable in the text construction process, or a component of context (see Zimmerman, 1998) is essentially absent from the systemic interpretation, where identity is for the most part simply a way of distinguishing between the various voices in an exchange.

This theoretical orientation toward the material and the semiotic (Thompson & Collins, 2001) places hard limits on what can be said about individuals’ language use within the community. The first constraint is on what can be contained within an account of the consumer journey. In SFL, the only parts of the consumer journey that are available for analysis are those where the consumer engages in an act of entextualisation—a recorded encounter within some social space. The inner or unrecorded paths within the journey—that portion comprised of a consumer’s unexpressed expectations, fears and beliefs—is inaccessible. The second constraint is on connecting language use to thought, and therefore, to key notions in HC research such as psycho-education or health literacy. While other theories conceptualise OSG use as a kind of talk therapy (Kaufman & Whitehead, 2016), and

while SFL can be fruitfully used to locate and explain the socio-semiotic processes of therapeutic talk, there is little room within the theory to argue that language change patterns with cognitive change, and therefore, that the observed patterns may in and of themselves be examples of positive or negative health outcomes. Accordingly, in future research intending to relate speakers' linguistic practices to the attainment of health outcomes, it may be useful to better distinguish between the SFG and SFL. More specifically, a researcher could adopt the grammar as a means of distinguishing between kinds of meaning and their realisation in the words and wordings of texts, but undertake the task of connecting language use to cognition through an alternative (psycholinguistic) lens.

8.4. Summary

In this and the previous chapters, I discussed an investigation of the Bipolar Forum as a linguistic corpus. Change over the course of membership was found to be both interpersonal and experiential. Users constantly renegotiate their status within the community through MOOD and MODALITY choices. Experientially, users demonstrate shifts in the kinds of participants and processes most commonly construed, but also the way these participants and processes are discursively framed. A reflection on the theory and methodology of the study highlighted key methodological challenges in extracting meaningful information from annotated corpora, and key theoretical challenges in the use of SFL as a theory of ontogenetic linguistic change.

In the next chapter, I outline the implications of the thesis and its associated tools and methods for CL, SFL and HC research.

9. Implications of the thesis

In the previous chapter, I addressed the first two research questions, concerning lexicogrammatical and semantic change in the Bipolar Forum. In this chapter, I address the third and fourth research questions. First, I outline the implications of the case study by research area, reflecting on what the case study means for CL (and CADS in particular), for SFL theory, and to the body of literature centred on healthcare communication (HC), both in online and offline contexts. With these implications in mind, I then address the final research question, which concerns the development of novel tools and methods for CL.

9.1. Corpus linguistics

The methods employed in the case study fall, for the most part, within the domain of CL. In other ways, however, the methods departed from what is typical of CL: the use of symbolic subcorpus structures and traversal of parser output, for example, are tasks more common in computational linguistics. This departure from ‘mainstream’ CL methods was necessitated by the specific nature of the research questions and dataset: interest was not in determining how the linguistic features of the Forum differ from other communities, or from language use generally. Rather, the case study sought to elucidate the internal structure of the community, and the effect of this structure on users’ linguistic choices. As explained in Chapter 4, further important key aims were to develop tools and strategies for increasing the transparency and reproducibility of CL results, and the ability to easily apply developed methods to new datasets for future work, either for the purpose of extending the analysis of the Bipolar Forum, or applying the methods to entirely new domains.

In contrast to the approach taken here, much contemporary CL is still deeply influenced by the methodological parameters of the Brown Corpus investigations of the early 1960s. In that era, corpora were expensive to construct, annotate, store and transmit; methods for automatic value-adding and querying were essentially unavailable. Today, corpora are still typically understood as being static, often monolithic collections of texts, with only the best-known/funded corpora receiving periodic updates (e.g. the BNC, COCA). Many of these corpora are accessible to the researcher only via specific query interfaces, constraining the kinds of things that can be searched for, and thus ultimately, the kinds of research questions that can be answered. As many of these interfaces are graphical (including graphical web-based tools), rather than programmatic, iterative and recursive searching of the corpus is generally not possible. Again, this constrains what is methodologically possible.

Over the past decades, technological advances (from e.g. computer science, computational linguistics, and NLP) have changed both what is feasible and what is possible for CL. A major contribution of the case study is in bringing some of these technological advances into the purview of CL research practice—this is a step forward not just for CL, but for the NLP community, who can see wider engagement with core tasks in their research area (De Marneffe & Manning, 2008). In the sections below, I outline key tools and tasks that have received little attention within CL, and then describe their utility in the context of the case study and developed tools.

9.1.1. Corpus structure and metadata

A major limitation of existing CL tools (especially of the graphical kind) has been the ability to work with the inherent structure of collections of texts used in corpora. Commonly, tools produce two dimensional results: a list of phenomena (e.g. words matching a query) and some kind of score for each (frequency, keyness, collocate strength, etc.). If a researcher is interested in discrete subcorpora, he/she may have to load each into the tool, run the same query, export results, and collate them in a separate tool. This is time-consuming, and increases the likelihood of making mistakes.

Tools that do have some awareness of subcorpora generally do this by treating each file or folder as a subcorpus. Therefore, a corpus must be duplicated and restructured in order to investigate different kinds of structures (for example, change over time and change over membership length). An issue within this approach, however, is that texts within a corpus may belong to multiple categories. In the case of the Bipolar Forum, for example, each post has a poster, a post count, a timestamp, and exists within a thread. These are metadata features that can be easily encoded within a corpus, and to which a tool can be sensitive during interrogation. `corpkit` solves this issue by transferring XML-formatted metadata from plain text into the parsed representation. These features can act as filters, where texts can be included/excluded based on some combination of metadata values, or can dynamically be treated by the tool as the subcorpora that comprise the corpus. Using the syntax of `corpkit`'s interpreter (Figure 9.1), it is trivial to investigate different dimensions within the same dataset. The first outputs results for the ten stages of membership; the second outputs annual subcorpora.

```
1 > search corpus for functions matching roles.process \
2 ...      with subcorpora as postgroup
3 > search corpus for functions matching roles.process \
4 ...      with subcorpora as year
```

Figure 9.1: Using the `corpkit` interpreter to switch between symbolic corpus structures

9.1.2. Natural language processing tool use

Tools for common computational linguistic tasks (tokenisation, lemmatisation, POS tagging, parsing, and coreference resolution for example) have become faster, more widely available, and considerably more accurate in recent years, thanks to advances in statistical machine learning/NLP (Manning & Schütze, 1999). These tools make it possible to search for more complex or abstract features of corpora than can be extracted from plain text. Of these tasks, only tokenisation has been fully embraced within CL, for the simple reason that word frequency counting is not possible without it. Lemmatisation, when performed, is often simply based on

wordlists of lemma forms and possible inflections (as is possible in *AntConc*). Such a method is unreliable, as POS information is needed to correctly return a base form. Without this information, a number of serious errors are likely to result (*human being* → *human be*). POS tagging and parsing are even more uncommon. As mentioned in Section 3.2.3, a key cause of this is a still-lingering historical skepticism toward the imposition of theory on corpus data. Corpus data, under this view, should instead inform theories of language (e.g. Sinclair, 2004).³⁶ This conviction ignores the fact that different kinds of research aims rely to a greater or lesser extent on corpus annotation (Anthony et al., 2013). Indeed, grammars can be (and have successfully been) derived from unannotated corpus data (e.g. Hunston & Francis, 2000). Discursively oriented tasks however, have different needs, centred on discovering how social actors are represented, or how interactants make interpersonal demands upon one another (Gee, 2013). With very large datasets, or with datasets in which the phenomenon of interest occurs thousands of times or more (Zinn & McDonald, 2015), researchers must find a way to effectively count and code the syntagmatic behaviour of the construction under investigation. This means choosing between the combination of parsing and traversal of parsed data structures, or leaving the data unparsed and then sampling from it, in order to create a manageably sized set of results for manual classification. The latter approach is oftentimes far from ideal, of course, as it involves both removal of instances of a phenomenon from consideration, and resource-intensive and error-prone manual work. A further advance of these kinds of annotation, discussed in more detail below, is that they can in fact eliminate theoretically problematic corpus interrogation practices, such as the use of stopword lists and reference corpora.

In the case study, all interrogation of the data relied on the output of the **Stanford CoreNLP** suite, which performs sentence splitting, tokenisation, POS tagging, lemmatisation, and constituency/dependency parsing.³⁷ These processes made possible the automated extraction of lexicogrammatical features that realise the discourse-semantics of corpus texts. Almost all search queries used in the generation of findings involved constituency and dependency parse traversal, in order to match, as closely as possible, concepts from the SFG (including but not limited to Subject,

```

1 > search corpus for function matching roles.event \
2 ...     and lemma matching processes.relational \
3 ...     and dependent-pos matching 'PRP' \
4 ...     and not dependent-word matching 'it' \
5 ...     and dependent-word matching 'bp|bipolar' \
6 ...     showing lemma \
7 ...     with subcorpora as postgroup

```

Figure 9.2: Complex query formulation using the `corpkit` interpreter

Finite, Modal and Polarity within the MOOD system and participant/Thing, process/Event, Agent, circumstance and Epithet/Classifier within TRANSITIVITY). These go far beyond what is possible using the mainstream practices of CL, such as regular-expression based searching of plain text, or of POS tagged or chunked data.

One good example of the utility of such methods is the analysis of *being/having bipolar*. To calculate the proportion of attributions that are made by *be*, *have* and *other relational processes*, lemmatisation, POS tags and parsing must all be used. The search query is complex, with five specifications at minimum. First, the query must match only verbs filling Event, rather than auxiliary roles. Second, the lemma form of the match must be in a list of possible relational processes (*be*, *have*, *seem*, *sound*, etc.). Third, it must have a dependent with a pronominal POS tag. Fourth, this pronoun should not be *it* (*It was the bipolar* is a false positive). Fifth, there should be another dependent, matching the unjargonised or jargonised terms for bipolar disorder. Two additional specifications are then required, in order to determine how the result should be returned (ideally, in lemma form), and what metadata feature should be used to delimit subcorpora (in this case, the post group, so that we can analyse the Membership Stage Structure). This very complex query can not be carried out using previously available tools. It can be easily expressed in the language of `corpkit`'s interpreter, however (Figure 9.2).

Incorporating programming

Another key technological affordance relevant to contemporary CL is an increased access to high-level programming languages, as well as comprehensive online documentation for learning and troubleshooting. Python, the main language used for the

case study and tool development, is one of a number of sensible choices for linguistic work. It is well-known for its readable, English-like syntax, and the ease with which it can be learned (Radenski, 2006). Python is high-level enough to facilitate working efficiently with linguistic data at word-level and above, and as such, is the language in which a large number of linguistic tools (including **NLTK**, **pattern**, **spaCy** and **UAM Corpus Tool**) are written.

A key advantage of programmatic approaches to CL more generally is that they allow iterative and recursive data exploration. Most obviously, the case study employs this method to iterate over multiple different symbolic corpus structures, representing change over time, or stages of membership. More powerful still is to apply this process to uncovering specific linguistic configurations within the corpus. Searching for salient processes, for example, revealed the emphasis new Forum users place on the *diagnose* Event. It was then possible to count the kinds of participants and circumstances involved in diagnosis-as-Event, and how these shift over the course of membership. The API example in Figure 9.3 locates the most key processes in the first subcorpus, and then searches for the participants in each.

```

1 # a place to store recursive results
2 part_in_proc = []
3 # search corpus for Events
4 proc = corpus.interrogate({F: roles.event}, show=L)
5 # calculate keyness and sort
6 proc = proc.edit('k', SELF, sort_by='total')
7 # get top 5 results
8 key_in_new = process.results.iloc[0].sort_values()[:5]
9 # recursively search for participants in this Event
10 for keyword in list(key_in_new):
11     parts = corpus.interrogate({GF: roles.process,
12                                 GL: keyword,
13                                 F: roles.participant},
14                                 show=L)
15 # store result
16 part_in_proc[keyword] = parts.results

```

Figure 9.3: Recursive investigation via the **corpkit** API

In addition to linguistic tools, programmatic approaches to CL also make it possible to interface with more general tools for result manipulation, statistical analysis and visualisation. Because these kinds of tasks are common within many sciences,

as well as in industry, available tools are generally fast, stable, powerful and actively maintained. In terms of future corpus tool design, Anthony et al. (2013) has highlighted the importance of modularity, and the ability to interact with these powerful third-party libraries. In the analysis of the *diagnose* process, for example, SciPy (Jones et al., 2001) was used to perform linear regression analysis, in order to sort participants and circumstances into those becoming more and less frequent over time, uncovering differences in the ways in which diagnosis was construed by new and veteran users. At the same time, the approach ensured that shifts in relative frequency were statistically significant. Dedicated corpus query interfaces generally limit the extent to which a researcher can engage with these kinds of tools. Continuing the earlier mock investigation of *being/having* bipolar disorder, we can see how these kinds of tools are useful in turning the absolute frequency results into a figure that shows longitudinal change in Forum users' linguistic preferences. Figure 9.4 provides an example of result merging and relative frequency calculation via `pandas` (McKinney, 2010), sorting via `scipy` (Jones et al., 2001) and visualisation via `matplotlib` (Hunter, 2007):

```

1  # pandas: merge result columns
2  > edit result by merging entries not matching '[be, have]' as 'Other'
3  # pandas: make relative frequencies for each subcorpus
4  > calculate edited as percentage of self
5  # scipy: calculate trend lines and sort
6  > sort edited by increase
7  # matplotlib: visualise output
8  > plot edited as line chart \
9  ...     with title 'Ascriptions of bipolar disorder' \
10 ...     with x_label as 'Membership stage' \
11 ...     and y_label as 'Percentage of all ascriptions'
```

Figure 9.4: Editing, sorting and visualising results with the `corpkit` interpreter

Indigenous reference corpora

In mainstream CL, word frequency counts in reference corpora are commonly used to identify keywords—that is, words that are unusually frequent in target corpus compared to the reference corpus according to some statistical measure (see Section 3.2). The aim of keywording, generally, is to locate words that are somehow salient

within the corpus (e.g. words that index common topics of conversation, or the interactants within corpus texts). As discussed in Chapter 3, the use of ‘balanced’ reference corpora as a means of generating keywords is inherently problematic from a theoretical perspective. When using such ‘balanced’ corpora, it is not possible to determine the extent to which contents of the corpus are representative of language use as a whole, nor is it possible to know what constitutes an appropriate weighting of different text types contained within. More accurately, there is no such thing as a ‘balanced’, ‘general’ corpus in the first place. When a person does CL, what is being investigated is the frequency of linguistic phenomena in a particular collection of instances of language use, or texts. What this is being compared to—the reference corpus—is not a representation of the probabilities inherent to the system of language, but simply a second (often larger) collection of instances. No matter the size and contents of the reference corpus, it can never contain the general probabilities for features in a language, because these probabilities are not encoded in the system until language has interacted with a situational context (i.e. a register) during the process of instantiation.

An associated problem with the use of reference corpora is that the registers they contain are often inappropriate for answering particular research questions. In the case of the Bipolar Forum, it makes little sense to compare the corpus to an arbitrary collection of British English texts (including short stories, speeches, etc.), unless the research question concerns the difference between language in the Forum and language in the BNC. More useful in almost every case would be to compare language in the Forum to that of other forums, or to language use in an offline bipolar disorder community. In systemic terms, most often, the ideal reference corpus should often differ in a single register variable (Field, Tenor or Mode) from that of the target corpus. For these reasons, the case study did not involve the use of a reference corpus of general English text. Instead, two alternative strategies were presented for extraction of salient lexicogrammatical features from the Bipolar Forum. First, keywording was performed on subsections of the corpus, using the entirety of the corpus as the reference material. This made it possible to locate salient words at different stages of membership, at different points in time, and in the language

of members who would go on to become veteran contributors, when compared to those who dropped out soon after joining. The second strategy is to use parsing to isolate words based on grammatical position, and to calculate the relative frequency of each. This method proved surprisingly accurate: simply counting the relative frequencies of participant heads was able to closely approximate a keyword list. This approach has the added usefulness of being able to quickly distinguish between grammatical roles, so that salient participants and processes could be considered separately, and so that experiential components could be considered apart from, for example, MOOD and MODALITY choices. This approach appears to be superior to keywording performed without restriction to grammatical role: the fact that participants outnumber processes in most registers of English, for example, means that participants will tend to dominate frequency lists, despite the centrality of processes to experiential meaning. This problem is exacerbated when lemmatisation is not performed, as verb inflections will be counted individually.

The proposed approach has some notable drawbacks. The most obvious is that texts must first be parsed, which can be a computationally intensive process, and which can be unsuitable for certain text types, for which parser models are not available. Manipulating parsed datasets is also more complicated than plain text files (though the developed software works toward the aim of increasing the ease with which parsed data can be queried). Finally, as mentioned earlier, parsing has occasionally been construed by CL practitioners as a (problematic) theoretical imposition on texts. In these cases, however, preference should be given to a theoretical imposition that is well-constructed and empirically informed, over reliance on arbitrary stopword lists and general corpora that even by proponents' accounts (e.g. Baker, 2012) are only impossible theoretical ideals.

Improving normalised frequency calculation

Corpus approaches to register (including much of Biber's work, e.g. Biber & Conrad, 2001; Biber, 2012, 1) have aimed to create quantitative sketches of a corpus or its subcorpora by counting relative frequencies of lexicogrammatical features (passives, interrogatives, *wh*- pronouns, etc.). These relative frequencies in these studies are

typically calculated by dividing the target feature by some constant, such as *per million words*. As others (e.g. Evert, 2005) have pointed out, however, the use of a constant denominator is often inappropriate. As shown in Chapter 7, *support* emerges as a key process in veteran discourse, as veterans explain to newcomers that the community is a site for exchange of both health information and interpersonal care. If we search for *support* as a process in the Membership Stage Structure, we will find a decreasing frequency per million words, however. The reason for this is that new users write more formally, with a slightly higher lexical density (see Figure 5.1). As such, clauses contain fewer processes, and more participants. Thus, we can easily misinterpret the quantitative results to understand *support* as less and less common, when in reality, it is a key difference in the processes selected at early and late stages of membership. The change in lexical density can in this way also be easily overlooked. A key advantage of the approach taken in the case study is that normalised frequency generation can be trivially and consistently applied: modalised clauses are divided by the number of clauses; *would/will* are divided by the number of modals, and so forth.

Collapsing the corpus/computational distinction

For the case study of this thesis, corpus linguistic tasks such as interrogation and keywording were performed with the aid of computational linguistic tools—namely, the `Stanford CoreNLP` suite. In light of the affordances of this approach, it is useful to stress at this point that corpus and computational linguistics are in many senses isomorphic, existing on a continuum of ‘digital linguistics’. The key difference is that CL tends to orient toward interpretation of datasets as a goal, while computational linguistics is generally concerned with the development of automated tools and workflows for processing (arbitrary) texts. As such, programming is central to computational linguistics, but peripheral within CL; conversely, solid theoretical linguistic grounding is generally a requirement within CL, but increasingly seen as unnecessary in computational linguistics (with the shift from rule-based to probabilistic, machine learning approaches to grammar).

Blurring the line between the two fields, as was done for the case study presented here, has distinct advantages. First, it allows corpus linguists to have a greater say in the kinds of language processing tools being built. Though tools developed in NLP may be explicitly aimed at non-computational researchers (De Marneffe & Manning, 2008), the current distinction between the two disciplines means that CL practitioners have no channel through which their needs can be articulated. At the same time, computational linguists are afforded a greater access to relevant theory and downstream applications. One example of this symbiotic relationship is in computational recent work on topic modelling—a means of classifying the ‘topic’ of documents through analysing the co-occurrence of lexical items (Blei, Ng, & Jordan, 2003). Despite widespread use elsewhere (e.g. history/digital humanities—see Blevins, 2010; Brauer, Dymitrow, & Fridlund, 2014; Yang, Torget, & Mihalcea, 2011), topic modelling has had poor take-up within CL. A key reason for this is that it willfully ignores key theoretical assumptions about how language makes meaning (Boyd-Graber & Blei, 2009). Under this approach, each lexical item in a text is given equal weight. An arbitrary list of high-frequency stopwords is generally used during pre-processing to remove tokens that are assumed to carry little topic meaning. Such approaches, do not engage with the fact that the grammatical position of words plays an important role in determining their centrality to experiential meanings (Halliday & Matthiessen, 2004). Collaboration on this front has led to topic modeling algorithms that outperform theory-free approaches, while satisfying theoretical demands for data-driven design (Rubino & McDonald, forthcoming).

9.1.3. Corpus assisted discourse studies

CL methods are used for a diverse range of research interests, ranging from lexicography to communities of practice. In being centred on investigating the relationship between lexicogrammar and discourse-semantics, the case study here is related mostly to CADS. Below, implications are outlined for this body of literature.

Exploiting parser output for discourse features

The most central contribution of the case study to CADS is in demonstrating the utility of extracting discursively significant phenomena from lexicogrammatical parses. Key to the process is in shifting between constituency and dependency parses: constituency grammars emphasise group/phrase structure and linear ordering of tokens, while dependency grammar emphasises argument types and functional roles. As such, it seems that MOOD (and presumably THEME) phenomena are best searched for using constituency parses, while TRANSITIVITY phenomena relate more closely to dependency parser output (Costetchi, 2013). The approach developed for automatically linking lexicogrammar to discourse is so far, a modest one, centred on the use of wordlists that collapse processes into types, and participants along taxonomic lines. This allowed frequency counting of different kinds of participants (*The Self, Other Members, Health Professionals, and Friends/Family*). From these categories, combined with systemic functional Process Type wordlists, it was possible to show how various participant taxonomies (see Martin, 1992) rise to prominence at later stages of membership, and at different points in the Forum's history.

Despite the successes of the case study in translating lexicogrammatical parses into patterns of meaning, it is important to acknowledge that such approaches cannot yet replace what is achieved in qualitatively oriented discourse analysis. Insights gained from sustained, manual analysis of individual texts-in-context are exceptionally difficult to approach using the methods presented here. The key reason for this is that automated processing of grammatical metaphor, as well as automated analysis of lexicogrammatical choices across the clause-complex, is a field still very much in its infancy. Improvement in these tasks will likely involve significant collaboration between computational linguists and expert human coders. Moreover, as proposed in Chapter 2, manual annotation of semantic phenomena could, in an interdisciplinary project, be used to train an automatic classifier that could annotate unseen text.

Automating thematic analysis

As discussed in Chapter 3, CADS typically relies on counting of lexis, and on concordancing of lexis to determine context of use. Missing in the current generation of corpus tools, as well as the CADS literature, however, is recognition that corpus interrogation for feature frequency counting and full concordancing are in fact more or less the same task. Techniques such as n-gram counting, where researchers locate a list of n lexical items that appear adjacently, occupy the middle-ground between the two poles (see Table 9.1). Concordancing simply uses human judgement of co-text to determine whether a search result conforms to or does not conform to some pattern in the content stratum. When researchers thematically code concordance lines, they are inherently responding to lexicogrammatical choices made in the clause(s) surrounding a lexical item of interest. Therefore, as corpus query languages and annotation conventions evolve, and as more abstract grammatical features can be traversed automatically, it becomes feasible to replace manual with automatic methods of categorisation. In many current use-cases for concordancing, the pattern matching could already potentially be automated. If we want to see how The Self is appraised, for example, we can search for Values when the Token is *I* (or, less commonly, adjectival modifiers of *me* as Thing).

Method	Stratum	Rank	Further analysis
Close reading	Lexicogrammar	Text/Clause-complex	Manual
Concordancing	Lexicogrammar	Clause-complex/clause	Manual
Parse interrogation	Lexicogrammar	Clause-complex/clause	Automatic
N-gramming	Lexicogrammar	Phrase/group	Automatic
Keywording	Lexis	Word	Automatic

Table 9.1: Methods in CADS and their coverage of the cline of instantiation, ordered by rank targeted

Using `corplib`, concordances are produced by default during every query. The concordance lines can then be used to determine that queries were capturing intended phenomena, to remove false positives, and to display typical co-text for lexicogrammatical phenomena of interest. This ensures that important meanings are not missed during the reduction of texts to frequency counts. Part of the contribution

of the developed tools and methods, therefore, is a scalable approach to thematic categorisation of language at group, phrase, clause and clause-complex levels.

9.2. Systemic-functional linguistic theory

SFL was used as the dominant underlying theory of language for the analysis of language use in the Forum. As explained in Chapter 3.3, a major motivation for this was SFL’s metafunctional division, which facilitated separate analyses of interpersonal and experiential meanings in the community. Also important is that SFL provides a well-articulated connection of morphosyntax to meaning-making and the performance of social action. Alongside other recent work (e.g. Coffin, 2013; Hunston, 2013, 4–5), the case study highlights the usefulness of SFL in CL and CMC research. Contributions for SFL are twofold: first, I highlight some theoretical issues uncovered during the course of the investigation. Second, I discuss practical issues in the use of SFL alongside currently available linguistic technology for CL research interests.

9.2.1. Theoretical contributions

The case study used only a fairly superficial rendition of the SFG, constrained by what could be extracted from constituency and dependency parses of texts. As such, theoretical contributions are best understood as discussion points and unresolved questions, rather than suggested changes to systemic theory.

Ergative transitivity and corpus methods

Process Types are a notoriously difficult part of the SFG to automatically label and/or extract (Costetchi, 2013). Even within the SFL community, multiple interpretations of the Process Type system are in use (c.f. the *Cardiff Grammar*— Fawcett, 2000). The reason for this difficulty is that lexical choices for main verbs are often metaphorical: *run*, for example, can variously realise a behavioural (*to run a mile*), material (*to run for office*), or even verbal process (*to run one’s mouth*). An added problem is that many verbs exhibit a high degree of indeterminacy, necessitating grammatically and/or

semantically informed disambiguation (Gwilliams & Fontaine, 2015; O'Donnell et al., 2009). In the best case scenario, Process Types in ambiguous cases can be resolved by looking at their TRANSITIVITY structure: as an intransitive, *appear* is generally material; when transitive, it is more likely relational; many Verbiages are tensed, embedded clauses, which can help in disambiguating verbal processes (Gwilliams & Fontaine, 2015). In many other cases, however, ambiguities are far more difficult to resolve.

The case study demonstrated the usefulness of the ergative model of TRANSITIVITY outlined by Halliday (1968, 1967b, 1967a). The ergative participant function of *Agent* is particularly useful in understanding which participants are construed as having agency—that is, being able to do things and make things happen (in material processes, according to the transitive model), to think things (in mental processes), say things (in verbal processes), and to attribute and assign values (in relational processes). The Bipolar Forum users shift from a pattern of Self-as-Medium toward Self-as-Agent over the course of membership, construing themselves as collaborators in health decision-making processes, and, at the same time, as responsible for managing manic and depressive episodes in the bipolar disorder cycle. Meanwhile, the agency of health professionals decreases over time: health professionals perform diagnosis and prescribe treatments to the Self-as-Medium in early posts, but, over time, become Mediums themselves, through which events such as medication management are carried out. As Halliday and Matthiessen (2004) explain, the transitive and ergative models of TRANSITIVITY are complementary in language, and thus in its analysis. In this investigation, the increased level of generality in the ergative grammar was found to be suitable for uncovering more general patterns in TRANSITIVITY as a representation of social actors; the transitive model of TRANSITIVITY is better employed when the researcher prefers to conceptualise unfolding experiential semantics as sequences of quanta of change (Halliday & Matthiessen, 1999).

Interaction between the metafunctions

While much has been written about intra-stratal relationships, both theoretically (e.g. Hasan, 1985), and empirically (e.g. Clarke, 2012), less attention has been paid

to the relationship between the metafunctions within SFL and the SFG. As metafunctions are realised through predominantly overlapping components of lexicogrammar, it is obvious that they cannot be completely independent variables: choices made for one metafunction can affect, on some level, the possible realisations in another. It is plausible to suggest, therefore, an **interplay between the metafunctions**, where changes in one dimension result in changes within another. Put differently, the difference between experiential meanings made in the Bipolar Forum, and experiential meanings made in emergency room interactions, is best understood in relation to what distinguishes the situations in the dimensions of Mode and Tenor. In terms of Mode, reflective and time-unlimited turns open up space for construal of Events and Things only loosely related to the overarching purpose of the community, or things that may be of little importance or relevance to other Forum users. Likewise, in terms of the Tenor of discourse, the relatively equal consumer-consumer power dynamic opens up room for a broader range of appropriate topics than is generally possible within a hospital or a clinic, where the power-unequal nature of interactions between professionals and consumers in hospitals and clinics, as well as the time-critical Mode, can both also manifest experientially in a narrowed semantic field.

Another Tenor/Mode-driven cause of experiential meaning choices is the apparent absence of health professionals in the community. In not being co-present, but in being commonly encountered by most community members, health professionals can be freely discussed, with a reduced risk of loss of face for interactants. This allows users to question the appropriateness of their and others' diagnoses or treatment plan. As shown in Chapter 7, veteran users may go so far as to construe health professionals as ignorant of the needs of those living with bipolar disorder, and of the phenomenology of the illness itself. Such challenges to the institutionally prescribed power of health professionals over health consumers are unlikely to take place within formal healthcare institutions.

9.2.2. Practical contributions

The thesis also contributes to SFL in a more practical sense, better integrating systemic theory with emerging computational methods for doing linguistic research, and identifying current obstacles to this goal.

Accounting for genre

One difficulty arising in the investigation was accounting for the influence of genre: first contributions contained a high frequency of *I am bipolar* constructions, with a number conforming to the well-known formulaic self-introduction of Alcoholics Anonymous meetings (see Chapter 7). Though first posts were treated as a stage of membership (and though, of course, the first post does represent a stage of membership), first posts are subject to particularly strong generic expectations: many users provide an account that stretches back to childhood, or to the onset of symptoms of bipolar or a similar condition; users typically include the events and circumstances surrounding their diagnosis, and provide a coda appraising their current situation or past choices. These generic expectations lead to a unique quantitative profile for first posts when compared to posts made at all other stages of membership. This effect poses a difficulty for automated analysis: in this case study, the generic expectations provide an unwanted variable when attempting to chart longitudinal change.

Ergative approaches

Developments in systemic parsing are ongoing. Existing methods, however, tend to be based on post-processing of CoreNLP parser output. Moreover, annotation of the discourse-semantic stratum is still a far-off goal. A central issue in systemic parsing is that participant role labels are selected by the Process Type. Incorrect Process Type identification therefore leads to incorrect participant labelling. One potential solution is to rely on the ergative model of TRANSITIVITY. The ergative interpretation requires a smaller label set, with many components determinable based on the existence or non-existence of an object argument of the main verb.

Quantitative register modelling

The quantitative modelling of registers is potentially useful for both theoretical and computational linguistics. Theoretically, quantitative register models could be used to build taxonomies and topologies of language as a network of registers (Matthiessen, 2015a). Computationally, register models could improve lexicogrammatical parsing and parse re-ranking, or facilitate dedicated parsing of the semantic stratum.

While SFL provides a delineation of register into Field, Tenor and Mode, and identifies the linguistic systems that relate to each, there does not yet appear to be a consensus regarding which linguistic features should be contained within a computational register model, or how the model should be built, stored or shared. Regarding possible features to include, the case study revealed that shallow features only tell part of the registerial story of a corpus: the syntagmatic behaviour of frequent or key lexical or lexicogrammatical features can also index discourses and ideologies within texts. The tools developed for this study certainly present one potential approach to a standardised modelling of register as one or more two-dimensional arrays. A set of interrogations could be defined as a script or routine, run over novel corpora, and added to a common (open-access, version controlled) model. The model could then be incorporated into a number of computational linguistic tasks: arbitrary texts could then be classified based on their similarity to modelled registers; a given corpus could be compared and contrasted with others, in order to locate suitable reference corpora, and so forth.

9.3. Health discourse

Language performs a number of roles in healthcare: it may be used to describe or elicit descriptions of symptoms; it may facilitate procedures; it may be consulted for information in reference books (Matthiessen, 2013). This centrality of language to the institution of healthcare means that language use can be related to the success of treatment (Divi, Koss, Schmaltz, & Loeb, 2007). As such, linguistic analysis of HC

can lead to interventions that lower the risk of avoidable patient harm (Slade et al., 2015a).

Most existing work on HC has centred on hospitals and clinics as settings, on intra-professional or professional-consumer interactions as Tenor configurations, and on face-to-face, spoken language as Mode. These contexts are undoubtedly of immense importance: miscommunication poses a serious risk to patients, especially in high-stakes emergency room or multilingual contexts (Slade et al., 2015b; Slade, Chandler, et al., 2015). Such data, however, is expensive and time-consuming to produce, requiring large-scale collaboration between universities and healthcare settings, and the accompanying ethical protocols; it is often unpredictable in terms of quality, content and structure.

The case study presented here demonstrates that very large datasets can be efficiently built using existing sites of CMC. OSG literature reviewed in Chapter 2 has also shown the feasibility of creating health-focussed online communities with the explicit aim of generating data for applied linguistic analysis (e.g. Johnson, Safadi, & Faraj, 2015). This comes at the expense of access to the face-to-face Mode variable, and often, the potential for follow-up interviews. Ethics may be more difficult to define and implement, as national guides may not provide clear and comprehensive guidelines for working with CMC. The affordances of digital data are many, however. First, it is relatively trivial to collect enough information to form quantitatively reliable accounts of language use. Second, data is amenable to both automatic collection and automatic analysis. Third, the method is scalable—a well designed workflow can potentially be applied to a new online community by doing little more than selecting a different URL to harvest.

Different Tenor variables provide a further key affordance: not only can we understand the ways in which relationships between interactants are formed, but also, we are given access to vivid, honest construals of the non-present participants in health discourse. Because of the non-presence of health professionals in the Bipolar Forum, consumers are free to construe them in ways which could be potentially face-threatening within clinical encounters. Just as analysis of communication between nurses has provided insights into the ways doctors and patients are construed, OSGs

provide representations of healthcare professionals that cannot be obtained through analysis of clinical encounters. Similarly, opinions that may diverge from normative biomedical discourse can also be aired.

9.4. Addressing Question 3: Needed tools and methods

Tool development comprised a substantial part of the overall research design. The primary reason for this, as noted in Chapter 4, is that existing tools were insufficient for some of the techniques presented in the case study analysis. Generally speaking, tools with graphical interfaces pose hard constraints on what kinds of things can be searched for and calculated. Command line tools, on the other hand, while being more flexible, are often centred on a single task (searching a single text; performing a calculation; drawing a figure) rather than larger workflows that require quickly shifting between these tasks. Though a combination of multiple tools could replicate some of the findings presented here, constant switching between tools has serious drawbacks. First, such a method is difficult to repeat or reproduce. Second, a lack of uniformity in the ways tools work with text can lead to inaccuracies: tokenisation and lemmatisation are handled differently by different tools, and in many cases, cannot be customised by the user. Third, some tools are proprietary, and may obscure the algorithms that underlie result generation. Fourth, the time and effort involved in moving between interfaces, importing and exporting data, and keeping records of what has been done would be substantial. Finally, adopting such a method would not save future researchers time or effort.

As a result of these considerations, purpose-built tools were constructed for the case study analysis. At the same time, however, the tools were designed for more general use within any of the intersecting disciplines interested in analysis of digital text (digital humanities, CL, NLP, etc.). That is to say, the computational infrastructure was not built around the dataset, but around a more general conceptualisation of linguistic datasets as structured collections of digitised text. The aim, therefore, was to develop software that could both facilitate analysis of the Bipolar Forum and

which could allow researchers to undertake similar kinds of analysis on arbitrary datasets in the future.

9.5. Implications: a summary

The findings of the case study, as well as the methods used to generate them, have implications for CL, SFL and HC. At the level of theory, new parts of the healthcare journey have been sketched from both quantitative and qualitative linguistic perspectives. At the same time, the production of an open-source tool for creation and analysis of parsed and structured corpora expands the kinds of research questions that can be answered using corpus methods. The next chapter provides a short research agenda, a summary of the thesis, and a conclusion centred on possible integration of the tools and methods developed in this case study within outcome-driven HC research.

10. Future directions

This thesis presented an investigation of a bipolar disorder OSG, which at the time of data collection contained over eight million words of user-generated text, over 5,800 members, and approximately 9,000 unique threads. Because the main aim of the thesis was to analyse shifts in both wordings and meanings over the course of membership, SFL, a theory of language that connects lexicogrammar to discourse and/or semantics was needed. At the same time, because the research design involved accounting for *all* contributions to the Forum, rather than a sample thereof, extensive computational tools and workflows were required. Key tasks from CL were reimplemented within a new software tool called `corplib`, which provides CL practitioners with improved support for symbolic subcorpus structures and annotated and/or parsed data. The project, therefore, was very much an interdisciplinary one, spanning qualitative and quantitative, as well as theoretical and computational approaches to language research.

An unavoidable limitation of interdisciplinary work is that it cannot engage with the theory and practices of a given discipline with the level of depth possible in work focussed on a single area of study. The central contribution of the thesis, therefore cannot be, for example, a linear extension of a grammatical system network within SFL. Instead, the central contribution is to demonstrate the usefulness of taking practices and ideas from one field and putting them to work in another—in this case, using computational, corpus, and systemic linguistics to better understand consumers' communication about healthcare. Throughout the thesis, an argument has been advanced that the use of state-of-the-art theory, tools and methods, can increase knowledge within the field of healthcare communication. Coupled with future advances, the same approach will also be able to contribute the nascent field

of clinical NLP, where computational processing of digital natural language leads to the discovery of information that can improve, in one way or another, the practice of medicine. In this chapter, I briefly sketch the kinds of developments that can make such an application of discourse-oriented linguistics possible. I then provide a brief summary of the thesis, and a conclusion.

10.1. A computational approach to healthcare discourse

There is increasing recognition within healthcare institutions that state-of-the-art tools and methods from computational linguistics/NLP can be fruitfully applied to the enormous and ever-increasing amount of digital linguistic healthcare data (Velupillai et al., 2015). As with recent research into the consumer healthcare journey (Slade et al., 2015a), however, most current work has focussed on intra-professional or professional-consumer settings, and on the components of the consumer journey that take place within the hospital or clinic. A key observation made in this thesis, echoing Jones (2013), is that consumer journeys can and often do include communication with non-professionals, and often leave the confines of formal healthcare institutions. Online, more and more often, people exchange important kinds of meanings that can undoubtedly affect decision-making practices regarding their health.

Already, CMC data has been used in computational workflows for disease surveillance (Kim et al., 2013) and public mental health monitoring (Paul et al., 2016). Intra-consumer text has been mined in order to detect adverse events and their causes (Chee et al., 2011). Addiction cycles can be modelled and predicted by analysis of contributions to an addiction forum (MacLean et al., 2015); successful rhetorical strategies for counsellors can be identified by comparing linguistic features of the session to follow-up survey results (Althoff et al., 2016). Despite the enormous promise of this emerging area of research, however, current approaches have serious shortcomings in terms of their conceptualisation of what language is and how it works. Many such studies treat texts as nothing but lists or sets of tokens, despite

the fact that decades of work in functional and discourse linguistics have shown the centrality of both grammar and context to the meaning-making process.

10.1.1. Necessary developments

`corpkit`, the tool developed for the analysis of the Bipolar Forum is a step in the direction of a computational discourse analysis that accounts for the role of grammar and context, as well as and the multifunctionality of language. Already, the tool can generate a quantitative description of shallow features of a corpus with a single command (`calculate features of corpus`, in the syntax of the interpreter). Extremely delicate querying of lexicogrammar, as shown throughout Chapters 6 and 7, is also facilitated by the tool. The output of these searches could foreseeably improve dramatically on the bag-of-words approach for many discourse-oriented tasks. At the same time, however, tool development is no substitute for the expertise of qualitative researchers. Many discourse-analytic studies involve manual classification of instances of language based on functional-semantic features that no current computational method can accurately identify. Locher's (2006) use of XML tags to annotate instances of advice provision with semantic information is just one of many examples of high quality resources that could be re-used as training data for machine learning classification tasks. Such an approach could put the insights generated in qualitative and corpus-assisted discourse studies to use on exponentially larger scales, and in novel domains. The first needed development, therefore, is a conduit through which qualitative and computational social scientists can share theory, tools, and data. What this conduit may look like, however, remains for the most part unknown.

Automatic annotation of the semantic stratum

At many points in the investigation, it became apparent that existing methods for lexicogrammatical parsing could not be used to directly access discourse-semantic features of texts. Even relatively simple semantic relationships, such as the relationship between *diagnose* and *diagnosis* (the latter being a reconfiguration of the former as a participant, rather than a process) is not annotated by any currently

popular NLP system. Any project interested in discourse–semantics, however, would benefit from explicit annotation of grammatical metaphor—for many applications, lexicogrammar is only of interest because it is the most convenient entry point to meaning-making at the rank of clause and below. Semantic annotation, while an ideal solution, is a task that remains in its infancy (see Rayson, Archer, Piao, & McEnery, 2004).

The most obvious challenge is that a useful automated semantic analysis is predicated on a conceptualisation of register, and of the text as an ongoing exchange. The Speech Function of a given Mood Type, for example, depends on the overall Tenor of the text, and on the content of adjacent clauses and moves. Most current approaches to parsing, however, do not exploit or respond to registerial information in any way. From a systemic-functional standpoint, it seems plausible to suggest that a discourse-semantic layer could be generated via oscillation between parser output (that is, the n -best parses of a clause/sentence) and a quantitative model of the register of a text. The parser could both develop and refine a semantic representation of the clause, and perform re-ranking of the parses, so that the most likely lexicographical parse is one that is sensitive to genre, register, individual speakers and the text as an interactive, unfolding event.

10.2. Summary of the thesis

The primary aim of the thesis was to investigate linguistic change over the course of membership in an OSG. This was to be accomplished **with** corpus/computational linguistic tools and methods, **via** SFL as theory and SFG as grammar of language, **for** the burgeoning area of HC research. Four research questions were developed and addressed. They are summarised in the four sections below.

10.2.1. Lexicogrammar at risk over membership

MOOD and MODALITY features of language vary considerably and consistently over the course of membership in the Bipolar Forum, with a steady increase in imperatives, and modalised declarative Mood. In terms of Subject choice, first person

is displaced by second person over time. A number of TRANSITIVITY features are also at risk. At the level of lexis, jargonised participants displace lay terms found in newcomers' talk (*doctor* → *doc* → *pdoc/tdoc*). In terms of processes, *wondering* and *thanking* give way to processes of *welcoming*; *feeling* is replaced by *thinking*, and violent and emotional processes (*killing*, *suffering*) are replaced by more positive ones (*hugging*, *recommending*, *improving*). Processes selected for sustained analysis showed shifts in the participants and circumstances that they typically select. The process of *diagnosis* shifts from being modified temporally to being modified for veracity (*recently diagnosed* → *accurately diagnosed*). Over the membership course, the ways in which Forum users ascribe bipolar disorder to themselves and others also undergo longitudinal change: while most Forum contributors tend to use identifying relational processes (*I am bipolar*), in late stages of membership, there is a shift toward construal of people as possessors and Agents over bipolar disorder (*I have bipolar*).

10.2.2. Linking lexicogrammar to meaning

Shifting upward from lexicogrammar on the hierarchy of stratification allows insight into changes in the kinds of meanings being made. In the Bipolar Forum, there are longitudinal differences in both interpersonal and experiential semantics. Interpersonally, users shift from requesting information and support to providing it. Modalised declaratives are used by newcomers to provide an account of their healthcare journey leading up to the present. In veteran talk, however, the modalised declarative is more commonly a realisation of advice. Experientially, users increasingly construe themselves and other healthcare consumers as active patients, and represent the relationship between the professional and consumer as more equal in terms of who is responsible for bringing about change.

10.2.3. Linking back to research

The third research question is centred on the implications of the case study and the developed methods for corpus linguistics, SFL, and healthcare communication research. Methodologically, a number of computational linguistic tools and methods

(e.g. dynamic corpus structures, parsing, programmatic workflows) were brought into the purview of corpus linguistics and corpus-assisted discourse research for the first time. At the level of theory, the thesis generated an account of the components of the consumer healthcare journey that take place outside hospitals and clinics. Online intra-consumer interaction was shown to be a rich resource for learning about how Forum users construe bipolar disorder and the world of healthcare with which they engage.

10.2.4. Needed tools and methods

A major outcome of the thesis was the development of a new tool for CL, which extends upon the functionality of existing tools by better engagement with corpus metadata and structure, and by allowing the user to exploit automatic annotation and parsing. The tool has three user interfaces, each of which is tailored for differing levels of familiarity with computer programming and the command line. This maximises potential take-up of the tool within a wide array of fields, ranging from computer science to the digital humanities. The simplest of these—the graphical interface—requires no knowledge of programming, and only a basic knowledge of constituency, dependency, or systemic grammar. The most powerful—the Python API—is flexible enough for incorporation within the kinds of workflows outlined in the previous section, and thus contributes to the nascent fields of computational social sciences and clinical NLP. Public use of the tool has increased steadily since its initial release. Engagement with users via GitHub suggests that all three interfaces are in use, and that users span a number of fields within and outside linguistics, as well as industry.

10.3. Conclusion

This thesis identified shifts in words, wordings, discourses and meanings over the course of membership in Bipolar Forum, an online health support group. This task involved the application of a theory of language (SFL) capable of distinguishing between interpersonal and ideational meaning and their realisations within gram-

matical systems. I found that Forum users' language undergoes a number of changes over the membership course. Most obviously, their language use increasingly reflects alignment with a biomedical account of bipolar disorder, and a consumer-centred model of the healthcare journey. Meanwhile, the thesis involved extensive development of new tools and methods for identifying linguistic change. These tools are now publicly available and in use.

Looking toward the future, it is clear that developments in statistical NLP will lead to greater accuracy in a number of tasks integral to automatic computational modelling and understanding of texts. At the same time, however, computational approaches would do well to incorporate insights from functional linguistic theory, which has shown how lexis and grammar operate as a single stratum to realise meaning, and from discourse analysis, where empirically informed functional-semantic categorisation could be repurposed as training data for computational models. Within the context of healthcare, such an interdisciplinary approach to discourse could foreseeably lead not only to the generation of new knowledge and theory, but eventually to improved clinical practice and health outcomes as well.

Notes

1. Simplification of the SFG for computational purposes, however, has long been the standard approach to computational SFL (Honnibal & Curran, 2007; Matthiessen & Bateman, 1991; Costetchi, 2013).
2. For example, the myriad studies of communication on *Usenet*—the text-only Internet communication system in which many of the social elements of CMC were first popularised—have been made largely redundant by the decommissioning of Usenet servers in 2010 (e.g. Berge & Collins, 1995; Eklundh & MacDonald, 1994; Jaffe, Lee, Huang, & Oshagan, 1995). That said, researchers have noted that despite greater bandwidth and the potential for multimodal interaction, in many respects, CMC has remained surprisingly text-based.
3. Interpersonally, not all health information online targets consumers—there are also sites with information for health professionals, academics, and so forth. These, however, are not relevant to the thesis.
4. The phenomenon of translating medical texts for new audiences, though possible in text-based forums, is apparently very uncommon. No literature was found that discusses it; nor was it seen in the case study.
5. Terminologically, related terms include *acculturation*, *enculturation*, *induction*, *initiation* and *inculcation*. These terms have been used both synonymously and contrastively in socialisation literature (Duff, 2010). Following Duff (2010), they are considered more or less interchangeable here, with the first preferred.
6. Technically, CL can be done on a single text: De Beaugrande (2001), for example, conducts a CDA using corpus methods to interrogate a single journal article by Widdowson.
7. It is worth noting that *corpus linguistics* itself may be a poorly chosen term (Baker, 2010), Lee's proposal of using 'corpus-based linguistics' (2007), though perhaps technically more accurate, would likely only create more confusion.
8. There are many other types and subtypes of corpus and corpus research not relevant to this thesis. An introduction to *monolingual/parallel*, *diachronic/synchronic* and *static/dynamic corpora* has been provided by Gries (2009). Multimodal corpora are discussed by Bateman (2013) and Adolphs (2012).
9. Arguably, specialised corpora are as old as balanced general corpora, as individual subsections of general corpora are essentially specialised corpora when interrogated in isolation (Warren, 2012). Though this may technically be the case, most often, such subsections are inadequate in terms of size or representativeness to be useful by themselves (Flowerdew, 2004).
10. Context and metadata retention has the drawback of being more difficult to anonymise, however.
11. Gries (2006) reminds us that for syntactic and grammatical investigations, keywords are largely unimportant.
12. Though alternate names and subdivisions between corpus-based, corpus-driven, corpus-informed discourse research (etc.) have been proposed, this thesis will simply use CADS as an umbrella term for all.
13. In Hallidayan SFL, experiential meanings and logical meanings together comprise the experiential metafunction. In the vein of Eggins (2004), this thesis discusses only experiential meaning, and all references to the transitivity system concern only experiential function.

14. Despite their potential relevance as a means of explicating the ways in which turn-taking is operationalised in OSGs, textual meanings are not covered in significant detail here due to limitations in scope.
15. A fuller account of the differences and tests needed to differentiate mood adjuncts has been provided by Eggins (2004).
16. It is important to note that the interpretation of genre as a level of abstraction greater than that of register is a development rejected within the ‘Hallidayan’ approach to SFL (Lukin et al., 2011). The genre analysis in Chapter 5 is performed because it elucidates in an intuitive way how Forum interactions may proceed in identifiable sequences, and how the probabilities for lexicogrammatical features are different within each sequence. Theoretical issues of language and context, such as where in the hierarchy of stratification the lexicogrammatical probabilities are set, and whether or not genre stages themselves reconfigure the probabilities, are not issues covered in this thesis.
17. It is important to distinguish generic structure potential—from *genre potential*—that is, ‘all the linguistically-achieved activity types recognised as meaningful (i.e. appropriate) in a given culture’ (Eggins, 2004, p. 35). In practice, this translates to every possible configuration of Field, Tenor and Mode.
18. Halliday has addressed a perceived lack of engagement with the field of pragmatics: Pragmatics, he argues ‘has always been simply the instantial end of the semantics. We don’t need a separate discipline’ (Thompson & Collins, 2001, p. 138).
19. Similar patterns of declining forum use have been noted elsewhere in online community literature: in a study of a forum dedicated to the band *Belle and Sebastian* (Deller, 2014), interviewed Forum users suggest that both social (changing interests, lack of time) and technological (obsolescence of forums) factors contribute to a rapid decline in use.
20. Users’ expectation of privacy is more likely to be an ethical issue in modes such as online chat, where chat messages appear to vanish after new messages arrive, and where chat transcripts are not searchable online, or connected through hyperlinks to a user’s profile.
21. Under the National Statement, forum users qualify as participants, despite their not being contacted, or even being aware of the fact that their data is being analysed.
22. It is important to remember that since a user must post 560 or more times to enter the last stage of membership, veteran users in the earlier years of the Forum’s existence are bound to be less common.
23. **UAM Corpus Tool**’s creator, Mick O’Donnell, has since re-factored the tool to handle very large datasets.
24. O’Donnell is preparing to release much of the non-GUI code as open-source (personal communication, 2015).
25. More recent releases of **NLTK** have better integration with **Stanford CoreNLP**.
26. Original list provided by Mick O’Donnell. Extending these lists to other Process Types is a key future aim.
27. In this thesis, all constituency queries use **Tregex**, which has a richer syntax, and is faster.
28. The qualitative, generic analysis was in fact undertaken as a kind of pilot study, before any corpus linguistic analysis was performed. The thread selected for analysis was therefore not chosen because it exemplified quantitatively identified features. The fact that many phenomena identified during the qualitative analysis are picked up during the corpus analysis is not by design—rather, it highlights the surprising uniformity of texts within the genre of ‘first post threads’.
29. This can also be seen as an attempt to maintain active discussion within the forum more generally: veteran members contribute by choice, and are therefore personally invested in the health of the community itself.

30. MOOD structure is not well-annotated in most typed dependency grammars currently used for parsing. As such, MOOD features need to be derived from verbose constituency queries. A possible reason for this is that most parsing tasks are experientially or textually oriented. Furthermore, parsers are typically trained on corpora containing very few non-declarative clauses.
31. There was little reason to expect dramatic or consistent polarity shifts over the course of membership. The analysis is carried out and presented mostly for the purposes of completeness of the register description.
32. Searching for participants by their order in clauses is often inaccurate, due to passivisation. In the case of relational processes, however, this is not an issue, as relational processes cannot be passivised (**bipolar was suffered from by me*).
33. The recent development of Twitter parsers has been, in part, an effort to cope with the amount of ellipses in some kinds of CMC (Kong et al., 2014). Even so, omissions in Twitter are brought about under unique contextual circumstances: constraints on the number of characters allowed in a message are a motivation for elision, rather than speed, avoiding of redundancy, etc. For this reason, Twitter parsers still face problems when annotating informal text.
34. When compared to the Universal Dependencies with computational applications in mind, it is important to consider the fact that the SFG may have a level of complexity that can lower parsing accuracy or exponentially increase processing time. Due to overlap and redundancy between the metafunctions, the SFG may provide diminishing returns in terms of depth of insight (O'Donnell & Bateman, 2005).
35. I am referring to Hallidayan SFL here—Fawcett, within the Cardiff tradition, has attempted to provide a link between SFL and cognition (e.g. 1980).
36. It also needs to be borne in mind that even the tokenisation process constitutes a theoretical imposition on corpus texts.
37. Coreference resolution and named entity recognition annotations were also available as annotators, but not used.

Bibliography

- Adolphs, S. (2012). Corpora: Multimodal. In *The Encyclopedia of Applied Linguistics*. Wiley Online Library.
- Althoff, T., Clark, K., & Leskovec, J. (2016). Large-scale analysis of counseling conversations: an application of natural language processing to mental health. *Transactions of the Association for Computational Linguistics*, 4, 463–476.
- Anderson, I. M., Haddad, P. M., & Scott, J. (2012). Bipolar disorder. *British Medical Journal*, 345, 1–10.
- Anthony, L. (2006). Developing a freeware, multiplatform corpus analysis toolkit for the technical writing classroom. *Professional Communication, IEEE Transactions on*, 49(3), 275–286.
- Anthony, L., Crosthwaite, P., Kim, T., Marchand, T., Yoon, S., Cho, S.-Y., ... Lee, H.-K. (2013). A critical look at software tools in corpus linguistics. *Linguistic Research*, 30(2), 141–161.
- Archer, D. (2012). Corpus annotation: a welcome addition or an interpretation too far? *Studies in variation, contacts and change (eSeries)*.
- Attard, A., & Coulson, N. S. (2012). A thematic analysis of patient communication in Parkinson's disease online support group discussion forums. *Computers in Human Behavior*, 28(2), 500–506.
- Austin, J. L. (1975). *How to do things with words*. Oxford, UK: Oxford University Press.
- Bachmann, I. (2011). Civil partnership—"gay marriage in all but name": a corpus-driven analysis of discourses of same-sex relationships in the UK Parliament. *Corpora*, 6(1), 77–105.
- Baker, P., Gabrielatos, C., Khosravinik, M., Krzyżanowski, M., McEnery, T., & Wodak, R. (2008). A useful methodological synergy? Combining critical discourse ana-

- lysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press. *Discourse & Society*, 19(3), 273–306.
- Baker, P. (2004). Querying Keywords Questions of Difference, Frequency, and Sense in Keywords Analysis. *Journal of English Linguistics*, 32(4), 346–359.
- Baker, P. (2010). *Sociolinguistics and corpus linguistics*. Edinburgh, UK: Edinburgh University Press.
- Baker, P. (2012). Acceptable bias? Using corpus linguistics methods with critical discourse analysis. *Critical Discourse Studies*, 9(3), 247–256.
- Baker, P. (2013). Corpora and discourse analysis. In K. Hyland (Ed.), *Discourse studies reader* (pp. 11–34). New York, NY: Bloomsbury.
- Baker, P., Gabrielatos, C., & McEnery, T. (2012, October). Sketching Muslims: A Corpus Driven Analysis of Representations Around the Word ‘Muslim’ in the British Press 1998–2009. *Applied Linguistics*, 1–25.
- Baker, P., & McEnery, T. (2015). Introduction. In *Corpora and Discourse Studies* (pp. 1–19). London, UK: Springer.
- Baldry, A. (2008). What are concordances for? Getting multimodal concordances to perform neat tricks in the university teaching and testing cycle. In A. Baldry, M. Pavese, C. T. Torsello, & C. Taylor (Eds.), *From didactas to ecolingua: an ongoing research project on translation and corpus linguistics* (pp. 35–50). Trieste: Edizioni Univesita.
- Balka, E., Krueger, G., Holmes, B. J., & Stephen, J. E. (2010). Situating Internet Use: Information-Seeking Among Young Women with Breast Cancer. *Journal of Computer-Mediated Communication*, 15(3), 389–411.
- Banks, D. (2009). The position of ideology in a systemic functional model. *WORD*, 60(1), 39–63.
- Baroni, M., Bernardini, S., Ferraresi, A., & Zanchetta, E. (2009). The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3), 209–226.
- Baroni, M., Kilgarriff, A., Pomikálek, J., & Rychlý, P. (2006). WebBootCaT: a web tool for instant corpora. In *Proceeding of the EuraLex Conference* (pp. 123–132).

- Bartley, L., & Benitez-Castro, M.-A. (2016). Evaluation and Attitude towards Homosexuality in the Irish Context: A Corpus-assisted Discourse Analysis of APPRAISAL Patterns in 2008 Newspaper Articles. *Irish Journal of Applied Social Studies*, 16(1), 1–20.
- Bateman, J. A. (2013). Multimodal Corpus-Based Approaches. In C. A. Chapelle (Ed.), *The encyclopedia of applied linguistics* (pp. 1–8). Wiley Online Library.
- Bauman, R., & Briggs, C. L. (1990). Poetics and performance as critical perspectives on language and social life. *Annual review of Anthropology*, 19, 59–88.
- Beckett, G. H., Amaro-Jimenez, C., & Beckett, K. S. (2010). Students' use of asynchronous discussions for academic discourse socialization. *Distance Education*, 31(3), 315–335.
- Berge, Z., & Collins, M. (1995). Computer-mediated scholarly discussion groups. *Computers & Education*, 24(3), 183–189.
- Biber, D. (2012). Register as a predictor of linguistic variation. *Corpus Linguistics and Linguistic Theory*, 8, 9–37.
- Biber, D., & Conrad, S. (2001). Quantitative corpus-based research: Much more than bean counting. *TESOL quarterly*, 35(2), 331–336.
- Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with python*. "O'Reilly Media, Inc."
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *the Journal of Machine Learning research*, 3, 993–1022.
- Blevins, C. (2010). Topic Modeling Martha Ballard's Diary. *Online: http://historying.org/2010/04/01/topic-modeling-martha-ballards-diary*.
- Bolander, B., & Locher, M. A. (2014). Doing sociolinguistic research on computer-mediated data: A review of four methodological issues. *Discourse, Context & Media*, 3, 14–26.
- boyd, d., & Ellison, N. (2007). Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication*, 13(1), 210–230.
- boyd, d. (2007). Social network sites: Public, private, or what. *Knowledge Tree*, 13(1), 1–7.

- boyd, d., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society*, 15(5), 662–679.
- Boyd-Graber, J. L., & Blei, D. M. (2009). Syntactic topic models. In *Advances in neural information processing systems* (pp. 185–192).
- Brauer, R., Dymitrow, M., & Fridlund, M. (2014). The digital shaping of humanities research: The emergence of Topic Modeling within historical studies. In *Enacting Futures: DASTS 2014 Conference (Danish Association for Science and Technology Studies), 12–13 June 2014, Roskilde University, Denmark*.
- Brown, P., & Levinson, S. C. (1987). *Politeness: Some universals in language usage*. Cambridge, UK: Cambridge University Press.
- Burnard, L. (2016, September 5). Reference guide for the british national corpus (xml edition). Retrieved February 5, 2014, from <http://www.natcorp.ox.ac.uk/docs/URG.xml>
- Burnett, G. (2000). Information exchange in virtual communities: a typology. *Information research*, 5(4).
- Butler, C. S. (2004). Corpus studies and functional linguistic theories. *Functions of language*, 11(2), 147–186.
- Caldas-Coulthard, C. R. (1993). From discourse analysis to critical discourse analysis: the differential re-presentation of women and men speaking in written news. *Techniques of description: Spoken and written discourse*, 196–208.
- Caldas-Coulthard, C. R., & Moon, R. (2010). ‘Curvy, hunky, kinky’: Using corpora as tools for critical analysis. *Discourse & Society*, 21(2), 99–133.
- Canary, D. J., & Yum, Y.-O. (2015). Relationship Maintenance Strategies. In *The international encyclopedia of interpersonal communication* (pp. 1–9). Wiley Online Library.
- Carey, J. (1980). Paralanguage in computer mediated communication. In *Acl 1980* (pp. 67–69).
- Cassell, J., & Tversky, D. (2005). The language of online intercultural community formation. *Journal of Computer-Mediated Communication*, 10(2).

- Chancellor, S., Lin, Z., & De Choudhury, M. (2016). “this post will just get taken down”: characterizing removed pro-eating disorder social media content. In *Proceedings of the 2016 chi conference on human factors in computing systems* (pp. 1157–1162). CHI 2016. Santa Clara, California, USA: ACM.
- Chancellor, S., Mitra, T., & De Choudhury, M. (2016). Recovery amid pro-anorexia: analysis of recovery in social media. In *Proceedings of the 2016 chi conference on human factors in computing systems* (pp. 2111–2123). CHI ’16. Santa Clara, California, USA: ACM.
- Chee, B. W., Berlin, R., & Schatz, B. (2011). Predicting adverse drug events from personal health messages. In *Amia annual symposium proceedings* (Vol. 2011, pp. 217–226).
- Chen, Z., Koh, P. W., Ritter, P. L., Lorig, K., Bantum, E. O., & Saria, S. (2015). Dissecting an Online Intervention for Cancer Survivors Four Exploratory Analyses of Internet Engagement and Its Effects on Health Status and Health Behaviors. *Health Education & Behavior*, 42(1), 32–45.
- Chiluwa, I. (2012, May). Social media networks and the discourse of resistance: A sociolinguistic CDA of Biafra online discourses. *Discourse & Society*, 23(3), 217–244.
- Chomsky, N. (1965). *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.
- Chomsky, N. (1986). *Knowledge of language: Its nature, origins, and use*. Westport, CT: Greenwood Publishing Group.
- Christie, F., & Martin, J. R. (2005, November). *Genre and Institutions: Social Processes in the Workplace and School*. Continuum.
- Clancy, P. M. (1999). The socialization of affect in Japanese mother-child conversation. *Journal of Pragmatics*, 31(11), 1397–1421.
- Clarke, B. P. (2012). *Do patterns of ellipsis in text support systemic functional linguistics’ ‘context-metfunction hook-up’ hypothesis? A corpus based approach* (Unpublished PhD thesis, Cardiff University, Cardiff, UK).
- Coffin, C. (2013). Using systemic functional linguistics to explore digital technologies in educational contexts. *Text & Talk*, 33(4–5), 497–522.

- Cosentino, V., Izquierdo, J. L. C., & Cabot, J. (2015). Assessing the bus factor of git repositories. In *2015 ieee 22nd international conference on software analysis, evolution, and reengineering (saner)* (pp. 499–503). IEEE.
- Costetchi, E. (2013). A method to generate simplified Systemic Functional Parses from Dependency Parses. *DepLing 2013*, 68–77.
- Courtney Walton, S., & Rice, R. E. (2013). Mediated disclosure on Twitter: The roles of gender and identity in boundary impermeability, valence, disclosure, and stage. *Computers in Human Behavior*, 29(4), 1465–1474.
- Crawford, P., Gilbert, P., Gilbert, J., Gale, C., & Harvey, K. (2013). The Language of Compassion in Acute Mental Health Care. *Qualitative Health Research*, 719–727.
- Crocker, M. W., Demberg, V., & Teich, E. (2016). Information density and linguistic encoding (ideal). *KI-Künstliche Intelligenz*, 30(1), 77–81.
- Crowston, K., & Williams, M. (2000). Reproduced and emergent genres of communication on the World Wide Web. *The Information Society*, 16(3), 201–215.
- Cummings, M. (2010). *An introduction to the grammar of old english: a systemic functional approach*. Equinox.
- Daft, R., & Lengel, R. (1983). *Information richness. A new approach to managerial behavior and organization design*. Office of Naval Research Technical Report Series.
- Danescu-Niculescu-Mizil, C., West, R., Jurafsky, D., Leskovec, J., & Potts, C. (2013). No country for old members: user lifecycle and linguistic change in online communities. In *Proceedings of the 22nd International Conference on World Wide Web* (pp. 307–318). International World Wide Web Conferences Steering Committee.
- Davies, B. (2005). Communities of practice: Legitimacy not choice. *Journal of Sociolinguistics*, 9(4), 557–581.
- De Beaugrande, R. (2001). Interpreting the discourse of HG Widdowson: a corpus-based critical discourse analysis. *Applied Linguistics*, 22(1), 104–121.
- De Choudhury, M., Kiciman, E., Dredze, M., Coppersmith, G., & Kumar, M. (2016, May). Shifts to suicidal ideation from mental health content in social media. In *Proceedings of the 2016 chi conference on human factors in computing systems* (pp. 2098–2110). ACM – Association for Computing Machinery.

- De Marneffe, M.-C., MacCartney, B., Manning, C. D., et al. (2006). Generating typed dependency parses from phrase structure parses. In *Proceedings of lrec* (Vol. 6, 2006, pp. 449–454).
- De Marneffe, M.-C., & Manning, C. D. (2008). The stanford typed dependencies representation. In *Coling 2008: proceedings of the workshop on cross-framework and cross-domain parser evaluation* (pp. 1–8). Association for Computational Linguistics.
- De Smedt, T., & Daelemans, W. (2012). Pattern for python. *Journal of Machine Learning Research*, 13, 2031–2035.
- de Marneffe, M.-C., Dozat, T., Silveira, N., Haverinen, K., Ginter, F., Nivre, J., & Manning, C. D. (2014). Universal dependencies: a cross-linguistic typology. *Proceedings of LREC*.
- de Saussure, F. (1916). *Course in general linguistics*. New York, NY: McGraw-Hill.
- DeCapua, A., & Dunham, J. F. (1993). Strategies in the discourse of advice. *Journal of Pragmatics*, 20(6), 519–531.
- DeCapua, A., & Huber, L. (1995). ‘If I were you...’: Advice in American English. *Multilingua—Journal of Cross-Cultural and Interlanguage Communication*, 14(2), 117–132.
- Degaetano-Ortlieb, S., Fankhauser, P., Kermes, H., Lapshinova-Koltunski, E., Ordan, N., & Teich, E. (2014). Data Mining with Shallow vs. Linguistic Features to Study Diversification of Scientific Registers. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC)* (pp. 1327–1334).
- Deller, R. (2014). A Decade in the Life of Online Fan Communities. *The Ashgate Research Companion to Fan Cultures*, Surrey: Ashgate Publishing, Ltd, 237–248.
- Delpisheh, E., & An, A. (2014). Topic Modeling Using Collapsed Typed Dependency Relations. In U. M. Fayyad, G. Piatetsky-Shapiro, P. Padhraic Smyth, & R. Uthurusamy (Eds.), *Advances in Knowledge Discovery and Data Mining* (pp. 146–161). Cambridge, MA: MIT Press.
- Dennen, V. P. (2008). Pedagogical lurking: Student engagement in non-posting discussion behavior. *Computers in Human Behavior*, 24(4), 1624–1633.

- Dickerson, S. S., Reinhart, A., Boemhke, M., & Akhu-Zahaya, L. (2011). Cancer as a problem to be solved: internet use and provider communication by men with cancer. *Computers Informatics Nursing*, 29(7), 388–395.
- Divi, C., Koss, R. G., Schmaltz, S. P., & Loeb, J. M. (2007). Language proficiency and adverse events in US hospitals: a pilot study. *International journal for quality in health care*, 19(2), 60–67.
- Dresner, E., & Herring, S. C. (2010). Functions of the nonverbal in CMC: Emoticons and illocutionary force. *Communication Theory*, 20(3), 249–268.
- Dridan, R., & Oepen, S. (2012). Tokenization: returning to a long solved problem a survey, contrastive experiment, recommendations, and toolkit. In *Proceedings of the 50th annual meeting of the association for computational linguistics: short papers-volume 2* (pp. 378–382). Association for Computational Linguistics.
- Duff, P. A. (2010). Language socialization into academic discourse communities. *Annual Review of Applied Linguistics*, 30(1), 169–192.
- Eggins, S. (2004). *Introduction to systemic functional linguistics*. London, UK: Continuum.
- Eggins, S., & Slade, D. (2004). *Analysing Casual Conversation*. Sheffield, UK: Equinox.
- Eklundh, K., & MacDonald, C. (1994). The use of quoting to preserve context in electronic mail dialogues. *Professional Communication, IEEE Transactions on*, 37(4), 197–202.
- Elkin, P. L., Froehling, D., Wahner-Roedler, D., Trusko, B. E., Welsh, G., Ma, H., ... Brown, S. H. (2008). NLP-based identification of pneumonia cases from free-text radiological reports. In *American medical informatics association annual symposium* (pp. 172–176).
- Ess, C. (2007). Internet research ethics. *The Oxford handbook of Internet psychology*, 487–502.
- Evert, S. (2005). *The statistics of word cooccurrences: word pairs and collocations* (Doctoral dissertation, University of Stuttgart).
- Evert, S., & Hardie, A. (2011). Twenty-first century corpus workbench: updating a query architecture for the new millennium. In *Proceedings of the corpus linguistics 2011 conference*. Birmingham, UK: University of Birmingham.

- Eysenbach, G. (2000). Towards ethical guidelines for e-health: JMIR theme issue on eHealth ethics. *Journal of Medical Internet Research*, 2(1), e7.
- Eysenbach, G., & Till, J. E. (2001). Ethical issues in qualitative research on internet communities. *British Medical Journal*, 323(7321), 1103–1105.
- Fawcett, R. P. (1980). *Cognitive linguistics and social interaction. towards an integrated model of a systemic functional grammar and the other components of a communicating mind* (Doctoral dissertation, Exeter, UK).
- Fawcett, R. P. (2000). *A theory of syntax for systemic functional linguistics*. Amsterdam, The Netherlands: John Benjamins.
- Feng, B., Li, S., & Li, N. (2016). Is a profile worth a thousand words? How online support-seeker's profile features may influence the quality of received support messages. *Communication Research*, 43(2), 253–276.
- Feng, V. W. (2015). *RST-Style Discourse Parsing and Its Applications in Discourse Analysis* (Unpublished PhD thesis, University of Toronto).
- Fletcher, W. H. (2012). Corpus Analysis of the World Wide Web. In C. A. Chapelle (Ed.), *The Encyclopedia of Applied Linguistics* (pp. 1–9). Wiley Online Library.
- Flowerdew, L. (2004). The argument for using English specialized corpora to understand academic and professional language. *Discourse in the professions. Perspectives from corpus linguistics*, 11–33.
- Fox, G. (1993). A comparison of “policespeak” and “normalspeak”: a preliminary study. *Techniques of Description: Spoken and Written Discourse, A Festschrift for Malcolm Coulthard*, London: Routledge, 183–95.
- Fox, G., Hoey, M., & Sinclair, J. M. (1993). *Techniques of description: Spoken and written discourse*. New York, NY: Routledge.
- Fox, S. (2014). The social life of health information. Retrieved June 12, 2016, from <http://www.pewresearch.org/fact-tank/2014/01/15/the-social-life-of-health-information/>
- Fox, S., & Duggan, M. (2013). Health online 2013. Retrieved from <http://pewinternet.org/Reports/2013/Health-online.aspx>
- Francis, W. N., & Kucera, H. (1979). Brown corpus manual. *Brown University*.

- Frisch, A.-L., Camerini, L., Diviani, N., & Schulz, P. J. (2012). Defining and measuring health literacy: how can we profit from other literacy domains? *Health Promotion International*, 27(1), 117–126.
- Füller, J., Jawecki, G., & Mühlbacher, H. (2007). Innovation creation by online basketball communities. *Journal of Business Research*, 60(1), 60–71.
- Galegher, J., Sproull, L., & Kiesler, S. (1998). Legitimacy, authority, and community in electronic support groups. *Written communication*, 15(4), 493–530.
- Gallagher, S. E., & Savage, T. (2015). “What is, Becomes What is Right”: A Conceptual Framework of Newcomer Legitimacy for Online Discussion Communities. *Journal of Computer-Mediated Communication*, 20(4), 400–416.
- Gee, J. P. (2004). Discourse analysis: What makes it critical. *An introduction to critical discourse analysis in education*, 19–50.
- Gee, J. P. (2007). *Social linguistics and literacies: Ideology in discourses*. London, UK: Routledge.
- Gee, J. P. (2013). *An introduction to discourse analysis: Theory and method*. New York, NY: Routledge.
- Giacomini, M. K., & Cook, D. J. (2000). Users' guides to the medical literature: XXIII. Qualitative research in health care A. Are the results of the study valid? *Jama*, 284(3), 357–362.
- Giesbrecht, E., & Evert, S. (2009). Is Part-of-Speech Tagging a Solved Task? An Evaluation of POS Taggers for the German Web as Corpus. In I. Alegria, I. Leturia, & S. Sharoff (Eds.), *Proceedings of the 5th web as corpus workshop* (pp. 27–36). San Sebastian, Spain.
- Gitelman, L. (2013). *Raw Data Is an Oxymoron*. Cambridge, MA: MIT Press.
- Goffman, E. (1959). *The Presentation of Self in Everyday Life*. New York: Doubleday Anchor.
- Goldsmith, D. J. (2000). Soliciting advice: The role of sequential placement in mitigating face threat. *Communications Monographs*, 67(1), 1–19.
- Goldsmith, D. J. (2004). *Communicating social support*. Cambridge, UK: Cambridge University Press.

- Gries, S. T. (2006). Exploring variability within and between corpora: some methodological considerations. *Corpora*, 1(2), 109–151.
- Gries, S. T. (2009). What is Corpus Linguistics? *Language and Linguistics Compass*, 3(5), 1225–1241.
- Gries, S. T. (2013, March). 50-something years of work on collocations. *International Journal of Corpus Linguistics*, 18(1), 137–165.
- Groom, N., Charles, M., & John, S. (2015). Introduction: corpora, grammar, and discourse analysis: recent trends, current challenges. In *Corpora, grammar, and discourse analysis* (Vol. 73, pp. 1–20). Amsterdam, The Netherlands: John Benjamins.
- Gwilliams, L., & Fontaine, L. (2015). Indeterminacy in process type classification. *Functional Linguistics*, 2(1), 1–19.
- Hadlington, L. (2015, October). Cognitive failures in daily life: Exploring the link with Internet addiction and problematic mobile phone use. *Computers in Human Behavior*, 51, 75–81.
- Halliday, M. A. K. (1966a). Some notes on ‘deep’ grammar. *Journal of Linguistics*, 2(1), 57–67.
- Halliday, M. A. K. (1978). *Language as social semiotic*. London, UK: Edward Arnold.
- Halliday, M. A. K. (1991). Corpus studies and probabilistic grammar. In K. Aijmer, & B. Altenberg (Eds.), *English Corpus Linguistics: studies in honour of Jan Svartvik* (pp. 30–43). New York, NY: Longman.
- Halliday, M. A. K. (1993a). Towards a language-based theory of learning. *Linguistics and Education*, 5(2), 93–116.
- Halliday, M. A. K. (2004). Introduction: how big is a language? On the power of language. *The Language of Science*, 49–101.
- Halliday, M. A. K., & Hasan, R. (1989). *Language, context, and text: Aspects of language in a social-semiotic perspective*. Victoria, Australia: Deakin University.
- Halliday, M. A. K., & Matthiessen, C. M. I. M. (1999). *Construing experience through meaning: a language-based approach to cognition*. London, UK: Cassell.
- Halliday, M. A. K., & Matthiessen, C. M. I. M. (2004). *An Introduction to Functional Grammar*. New York, NY: Routledge.

- Halliday, M. (1966b). The concept of rank: a reply. *Journal of Linguistics*, 2(01), 110–118.
- Halliday, M. (1967a). Notes on transitivity and theme in English: Part 2. *Journal of linguistics*, 3(02), 199–244.
- Halliday, M. (1967b). Notes on transitivity and theme in English Part I. *Journal of linguistics*, 3(01), 37–81.
- Halliday, M. (1968). Notes on transitivity and theme in English: Part 3. *Journal of linguistics*, 4(2), 179–215.
- Halliday, M. A. (1993b). Towards a language-based theory of learning. *Linguistics and education*, 5(2), 93–116.
- Han, J. Y., Kim, J.-H., Yoon, H. J., Shim, M., McTavish, F. M., & Gustafson, D. H. (2012). Social and psychological determinants of levels of engagement with an online breast cancer support group: posters, lurkers, and nonusers. *Journal of Health Communication*, 17(3), 356–371.
- Hardey, M. (1999). Doctor in the house: the Internet as a source of lay health knowledge and the challenge to expertise. *Sociology of Health & Illness*, 21(6), 820–835.
- Hardt-Mautner, G. (1995). “Only Connect”: Critical Discourse Analysis and Corpus Linguistics. Lancaster University. UCREL Lancaster.
- Harrison, S., & Barlow, J. (2009). Politeness strategies and advice-giving in an online arthritis workshop. *Journal of Politeness Research*, 5, 93–111.
- Harter, L. M., & Krone, K. J. (2001). Exploring the emergent identities of future physicians: Toward an understanding of the ideological socialization of osteopathic medical students. *Southern Journal of Communication*, 67(1), 66–83.
- Harvey, K., & Brown, B. (2012). Health Communication and Psychological Distress: Exploring the Language of Self-harm. *Canadian Modern Language Review/La Revue canadienne des langues vivantes*, 68(3), 316–340.
- Harvey, K. (2012, September). Disclosures of depression: Using corpus linguistics methods to examine young people’s online health concerns. *International Journal of Corpus Linguistics*, 17(3), 349–379.

- Harvey, K., Brown, B., Crawford, P., Macfarlane, A., & McPherson, A. (2007). "Am I normal?" Teenagers, sexual health and the internet. *Social science & medicine*, 65(4), 771–781.
- Hasan, R. (1985). The structure of a text. *Language, context, and text: Aspects of language in a social-semiotic perspective*, 52–69.
- Hasan, R., & Perrett, G. (1994). Learning to function with the other tongue: A systemic functional perspective on second language teaching. *Perspectives on pedagogical grammar*, 179–226.
- Hawkins, J. A. (1992). Syntactic weight versus information structure in word order variation. In *Informationsstruktur und Grammatik* (pp. 196–219). Berlin, Germany: Springer.
- Heritage, J., & Sefi, S. (1992). Dilemmas of advice: Aspects of the delivery and reception of advice in interactions between health visitors and first-time mothers. *Talk at work: Interaction in institutional settings*, 359–419.
- Herring, S. (2001). Computer-mediated discourse. In D. Schiffrin, D. Tannen, & H. E. Hamilton (Eds.), *The handbook of discourse analysis* (pp. 612–634). Malden, MA: Blackwell.
- Herring, S. (2000). Gender differences in CMC: Findings and implications. *Computer Professionals for Social Responsibility Journal*, 18(1).
- Herring, S. (2007). A faceted classification scheme for computer-mediated discourse. *Language@Internet*, 4(1).
- Herring, S. (1996a). *Computer-mediated communication: Linguistic, social, and cross-cultural perspectives*. Amsterdam, The Netherlands: John Benjamins.
- Herring, S. (1996b). Two variants of an electronic message schema. *PRAGMATICS AND BEYOND NEW SERIES*, 81–108.
- Herring, S. C. (2011). Discourse in Web 2.0: Familiar, reconfigured, and emergent. In Tannen, D, & Trester, A. M. (Eds.), *Discourse 2.0: Language and new media* (pp. 1–25). Washington, DC: Georgetown University Press.
- Hewson, C. (2015). Ethics Issues in Digital Methods Research. In S. Roberts, H. Snee, C. Hine, Y. Morey, & H. Watson (Eds.), *Digital Methods for Social Science*:

- An Interdisciplinary Guide to Research Innovation* (pp. 206–221). New York, NY: Palgrave Macmillan.
- Heyvaert, L. (2003). Nominalization as grammatical metaphor: On the need for a radically systemic and metafunctional approach. In A. M. Simon-Vandenbergen, L. Ravelli, & M. Taverniers (Eds.), *Grammatical metaphor: views from systemic functional linguistics* (pp. 65–99). Amsterdam, The Netherlands: John Benjamins.
- Hiippala, T. (2015). *The structure of multimodal documents: an empirical approach*. New York: Routledge.
- Hiippala, T. (2016, August 11). Semi-automated annotation of page-based documents within the Genre and Multimodality framework. In *Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*. Berlin, Germany: Association for Computational Linguistics.
- Hoch, D. B., Norris, D., Lester, J. E., & Marcus, A. D. (1999). Information exchange in an epilepsy forum on the World Wide Web. *Seizure*, 8(1), 30–34.
- Hoey, M. (2005). *Lexical priming: a new theory of words and language*. London, UK: Psychology Press.
- Hoffman-Goetz, L., Donelle, L., & Thomson, M. D. (2009). Clinical guidelines about diabetes and the accuracy of peer information in an unmoderated online health forum for retired persons. *Informatics for Health and Social Care*, 34(2), 91–99.
- Honnibal, M., & Curran, J. R. (2007). Creating a systemic functional grammar corpus from the Penn treebank. In *Proceedings of the Workshop on Deep Linguistic Processing* (pp. 89–96).
- Horne, J., & Wiggins, S. (2009). Doing being ‘on the edge’: managing the dilemma of being authentically suicidal in an online forum. *Sociology of Health & Illness*, 31(2), 170–184.
- Hsiao, I. H. (2012). A Corpus Approach to Discourse Analysis of Newspaper Restaurant Reviews: A Preliminary Analysis. *Studies in Literature & Language*, 5(3), 95–100.
- Huddleston, R. (1988). Constituency, multi-functionality and grammaticalization in Halliday’s Functional Grammar. *Journal of Linguistics*, 24(01), 137–174.

- Hudson, J. M., & Bruckman, A. (2004). "Go away": participant objections to being studied and the ethics of chatroom research. *The Information Society*, 20(2), 127–139.
- Hudson, T. (1990). The discourse of advice giving in English: 'I wouldn't feed until spring no matter what you do'. *Language & Communication*, 10(4), 285–297.
- Hunston, S. (2006). Corpus linguistics. *Linguistics*, 7(2), 215–244.
- Hunston, S. (2013). Systemic functional linguistics, corpus linguistics, and the ideology of science. *Text & Talk*, 33, 617–640.
- Hunston, S., & Francis, G. (2000). *Pattern grammar: A corpus-driven approach to the lexical grammar of English*. Cambridge, MA: MIT Press.
- Hunter, J. D. (2007, May). Matplotlib: a 2d graphics environment. *Computing in Science Engineering*, 9(3), 90–95.
- Hymes, D. (1972). On communicative competence. In J. Pride, & J. Holmes (Eds.), *Sociolinguistics* (pp. 269–293). Harmondsworth: Penguin Books.
- Jaffe, J., Lee, Y., Huang, L., & Oshagan, H. (1995). Gender, pseudonyms, and CMC: Masking identities and baring souls. In *Proceedings of the 45th annual conference of the international communication association*. Albuquerque, New Mexico, USA.
- Jakubíček, M., Kilgarriff, A., Kovář, V., Rychlý, P., & Suchomel, V. (2013). The TenTen corpus family. In *7th International Corpus Linguistics Conference CL* (pp. 125–127).
- Jaworska, S. (2016). Using a Corpus-Assisted Discourse Studies (CADS) Approach to Investigate Constructions of Identities in Media Reporting Surrounding Mega Sports Events: The Case of the London Olympics 2012. In I. R. Lamond, & L. Platt (Eds.), *Critical Event Studies* (pp. 149–174). New York, NY: Palgrave Macmillan.
- Jaworska, S., & Krishnamurthy, R. (2012, July). On the F word: A corpus-based analysis of the media representation of feminism in British and German press discourse, 1990–2009. *Discourse & Society*, 23(4), 401–431.
- Johnson, S. L., Safadi, H., & Faraj, S. (2015). The Emergence of Online Community Leadership. *Information Systems Research*, 26(1), 165–187.
- Jones, E., Oliphant, T., & Peterson, P. (2001). SciPy: open source scientific tools for Python.

- Jones, R. (2013). *Health and risk communication: An applied linguistic perspective*. New York, NY: Routledge.
- Jorm, A., Barney, L., Christensen, H., Highet, N., Kelly, C., & Kitchener, B. (2006). Research on mental health literacy: what we know and what we still need to know. *Australian and New Zealand Journal of Psychiatry*, 40(1), 3–5.
- Kaufman, S., & Whitehead, K. A. (2016). Producing, ratifying, and resisting support in an online support forum. *Health*.
- Kiesler, S., Siegel, J., & McGuire, T. (1984). Social psychological aspects of computer-mediated communication. *American Psychologist*, 39(10), 1123–1134.
- Kilgarriff, A. (2007). Googleology is bad science. *Computational linguistics*, 33(1), 147–151.
- Kilgarriff, A. (2013). *WebBootCat usage 2010–13*.
- Kilgarriff, A., & Renau, I. (2013). esTenTen, a Vast Web Corpus of Peninsular and American Spanish. *Procedia - Social and Behavioral Sciences*, 95, 12–19.
- Kim, E.-K., Seok, J. H., Oh, J. S., Lee, H. W., & Kim, K. H. (2013). Use of hangeul twitter to track and predict human influenza infection. *PloS one*, 8(7), e69305.
- Kim, E., Han, J. Y., Moon, T. J., Shaw, B., Shah, D. V., McTavish, F. M., & Gustafson, D. H. (2012). The process and effect of supportive message expression and reception in online breast cancer support groups. *Psycho-Oncology*, 21(5), 531–540.
- Kim, M.-S., & Raja, N. S. (1991). Verbal Aggression and Self-Disclosure on Computer Bulletin Boards. In *Paper presented at the 41st annual meeting of the international communication association*. Chicago, IL.
- King, S. A. (1996). Researching Internet communities: Proposed ethical guidelines for the reporting of results. *The Information Society*, 12(2), 119–128.
- Kong, L., Schneider, N., Swayamdipta, S., Bhatia, A., Dyer, C., & Smith, N. A. (2014). A dependency parser for tweets. *Proceedings of Conference on Empirical Methods In Natural Language Processing*, 1001–1012.
- Koteyko, N. (2010). Mining the Internet for linguistic and social data: an analysis of ‘carbon compounds’ in web feeds. *Discourse & Society*, 21(6), 655–674.

- Koteyko, N. (2014). *Language and politics in post-Soviet Russia: A corpus assisted approach*. UK: Palgrave Macmillan.
- Koteyko, N. (2015). Corpus-assisted analysis of Internet-based discourses: from patterns to rhetoric. In J. Ridolfo, & W. Hart-Davidson (Eds.), *Rhetoric and digital humanities* (p. 184). Chicago, IL: University of Chicago Press.
- Koteyko, N., & Hunt, D. (2015). Performing health identities on social media: An online observation of Facebook profiles. *Discourse, Context & Media*, 59–67.
- Koteyko, N., Jaspal, R., & Nerlich, B. (2013). Climate change and ‘climategate’ in online reader comments: a mixed methods study. *The Geographical Journal*, 179(1), 74–86.
- Kouper, I. (2010). The pragmatics of peer advice in a LiveJournal community. *Language@Internet*, 7.
- Labov, W., & Waletzky, J. (1967). Narrative analysis: Oral versions of personal experience. In J. Jehlm (Ed.), *Essays on the verbal and visual arts* (pp. 12–44). Seattle, WA: University of Washington Press.
- Lam, W. S. E. (2008). Language socialization in online communities. In N. H. Hornberger (Ed.), *Encyclopedia of language and education* (pp. 2859–2869). London, UK: Springer.
- Lander, J. (2014). Building community in online discussion: A case study of moderator strategies. *Linguistics and Education*, 107–120.
- Landert, D., & Jucker, A. H. (2011). Private and public in mass media communication: From letters to the editor to online commentaries. *Journal of Pragmatics*, 43(5), 1422–1434.
- Lange, P. G. (2007). Publicly private and privately public: Social networking on YouTube. *Journal of Computer-Mediated Communication*, 13(1), 361–380.
- Larsen-Freeman, D. (2000). *Techniques and Principles in Language Teaching*. Oxford, UK: Oxford University Press.
- Latour, B., & Canea, M. (2010). Tarde’s idea of quantification. In *The Social After Gabriel Tarde: Debates and Assessments* (pp. 145–162). London, UK: Routledge.

- Le, T., & Wang, X. (2009). Systematic Functional Linguistics and Critical Discourse Analysis. In *Critical discourse analysis: an interdisciplinary perspective* (pp. 27–36). Hauppauge, NY: Nova Science.
- Lee, D. (2007). Corpora and discourse analysis: New ways of doing old things. *Advances in Discourse Studies*, 86–99.
- Lee, N.-J., Shah, D. V., & McLeod, J. M. (2013). Processes of Political Socialization A Communication Mediation Approach to Youth Civic Engagement. *Communication Research*, 40(5), 669–697.
- Lee, S., Park, D.-H., & Han, I. (2014). New members' online socialization in online communities: The effects of content quality and feedback on new members' content-sharing intentions. *Computers in Human Behavior*, 30, 344–354.
- Leech, G. (1992a). 100 million words of english: the british national corpus (bnc). *Language Research*, 28(1), 1–13.
- Leech, G. (1992b). Corpora and theories of linguistic performance. *Directions in corpus linguistics*, 105–122.
- Leech, G. (1997). Introducing corpus annotation. In R. Garside, G. Leech, & A. McEnery (Eds.), *Corpus annotation: linguistic information from computer text corpora* (pp. 1–18). New York, NY: Longman.
- Leech, G. (2006). New resources, or just better old ones? The Holy Grail of representativeness. *Language and Computers*, 59(1), 133–149.
- Levy, R., & Andrew, G. (2006). Tregex and tsurgeon: tools for querying and manipulating tree data structures. In *Proceedings of the fifth international conference on language resources and evaluation* (pp. 2231–2234).
- Lindholm, M. (2012). *Identity construction in a pwnage video on YouTube* (Unpublished Bachelor thesis, University of Jyväskylä, Jyväskylä, Finland).
- Locher, M. A. (2006). *Advice online: Advice-giving in an American Internet health column*. Amsterdam, The Netherlands: John Benjamins.
- Locher, M. A. (2010, February). Health Internet sites: a linguistic perspective on health advice columns. *Social Semiotics*, 20(1), 43–59.
- Lukač, M. (2011, December). Down to the bone: A corpus-based critical discourse analysis of pro-eating disorder blogs. *Jezikoslovje*, 12(2), 187–209.

- Lukin, A., Moore, A. R., Herke, M., Wegener, R., & Wu, C. (2011). Halliday's model of register revisited and explored. *Linguistics and the Human Sciences*, 4(2), 187–213.
- MacLean, D., Gupta, S., Lembke, A., Manning, C., & Heer, J. (2015). Forum77: An analysis of an online health forum dedicated to addiction recovery. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing* (pp. 1511–1526). ACM.
- Maddox, T. M., & Matheny, M. A. (2015). Natural Language Processing and the Promise of Big Data Small Step Forward, but Many Miles to Go. *Circulation: Cardiovascular Quality and Outcomes*, 8(5), 463–465.
- Manchaiah, V. K., Stephens, D., Andersson, G., Rönnberg, J., & Lunner, T. (2013). Use of the 'patient journey' model in the internet-based pre-fitting counseling of a person with hearing disability: study protocol for a randomized controlled trial. *Trials*, 14, 1–7.
- Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing*. Cambridge, MA: MIT Press.
- Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., & McClosky, D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. In *Acl 2014* (pp. 55–60).
- Markham, A., & Buchanan, E. (2012). Ethical decision-making and Internet research: Recommendations from the AoIR Ethics Working Committee (version 2.0). Retrieved from <http://www.aoir.org/reports/ethics2.pdf>.
- Martin, J. R., & White, P. R. R. (2005). *The language of evaluation: appraisal in English*. New York, NY: Palgrave Macmillan.
- Martin, J. R. (1992). *English text: System and structure*. Amsterdam, The Netherlands: John Benjamins.
- Martin, J. R. (2006). Genre, ideology and intertextuality: a systemic functional perspective. *Linguistics & the Human Sciences*, 2(2).
- Martin, J. R. (2013, March). Genre-based literacy programs: contextualising the SLATE project. *Linguistics and the Human Sciences*, 7(1–3).
- Martin, J. R., & Wodak, R. (2003). *Re/reading the past: Critical and functional perspectives on time and value*. Amsterdam, The Netherlands: John Benjamins.

- Martin, J. R., & Rose, D. (2003). *Working with discourse: Meaning beyond the clause*. London, UK: Continuum.
- Matthiessen, C. M. I. M. (2014). Extending the description of process type within the system of transitivity in delicacy based on Levinian verb classes. *Functions of Language*, 21(2), 139–175.
- Matthiessen, C. M. (1995). *Lexicogrammatical cartography: English systems*. Tokyo, Japan: International Language Science.
- Matthiessen, C. M. (2013). Applying systemic functional linguistics in healthcare contexts. *Text & Talk*, 33(4–5), 437–466.
- Matthiessen, C. M. (2015a). Modeling context and register: the long-term project of registerial cartography. *Letras*, (50), 15–90.
- Matthiessen, C. M. (2015b). Register in the round: registerial cartography. *Functional Linguistics*, 2(1), 1–48.
- Matthiessen, C., & Bateman, J. A. (1991). *Text generation and systemic-functional linguistics: experiences from English and Japanese*. London, UK: Pinter.
- Matthiessen, C., Teruya, K., & Lam, M. (2010). *Key terms in systemic functional linguistics*. London, UK: Continuum.
- Matthiessen, C. (1998). Construing processes of consciousness: from the commonsense model to the uncommonsense model of cognitive science. *Reading science. Critical and functional perspectives on discourses of science* New York: Routledge, 329–354.
- Mauranen, A. (2003). “But heres a flawed argument”: Socialisation into and through Metadiscourse. *Language and Computers*, 46(1), 19–34.
- Mautner, G. (2005). Time to get wired: Using web-based corpora in critical discourse analysis. *Discourse & Society*, 16(6), 809–828.
- Mayfield, E., Laws, M. B., Wilson, I. B., & Rosé, C. P. (2014). Automating annotation of information-giving for analysis of clinical conversation. *Journal of the American Medical Informatics Association*, 21, e122–e128.
- McEnery, T., & Wilson, A. (1996). *Corpus linguistics*. Edinburgh, UK: Edinburgh University Press.

- McEnery, T., Xiao, R., & Tono, Y. (2006). *Corpus-based language studies: An advanced resource book*. London, UK: Taylor & Francis.
- McKinney, W. (2010). Data structures for statistical computing in python. In S. van der Walt, & J. Millman (Eds.), *Proceedings of the 9th python in science conference* (pp. 51–56).
- Meyer, C. F. (2002). *English corpus linguistics: An introduction*. Cambridge, UK: Cambridge University Press.
- Miller, T. A., Bethard, S., Dligach, D., Pradhan, S., Lin, C., & Savova, G. K. (2013). Discovering narrative containers in clinical text. In *Acl 2013* (pp. 18–26).
- Minocha, A., Hyderabad, I., Reddy, S., & Kilgarriff, A. (2013). Feed Corpus: An Ever Growing Up-To-Date Corpus. In S. Evert, E. Stemle, & P. Rayson (Eds.), *Proceedings of the 8th web as corpus workshop (wac-8)*.
- Mo, P. K., & Coulson, N. S. (2013). Are online support groups always beneficial? A qualitative exploration of the empowering and disempowering processes of participation within HIV/AIDS-related online support groups. *International Journal of Nursing Studies*, 983–993.
- Morzy, M. (2012). Analysis and Mining of Online Communities of Internet Forum Users. In *Data Mining: Foundations and Intelligent Paradigms* (pp. 225–263). Berlin, Germany: Springer.
- Mulderrig, J. (2012, November). The hegemony of inclusion: A corpus-based critical discourse analysis of deixis in education policy. *Discourse & Society*, 23(6), 701–728.
- Mulveen, R., & Hepworth, J. (2006). An interpretative phenomenological analysis of participation in a pro-anorexia internet site and its relationship with disordered eating. *Journal of Health Psychology*, 11(2), 283–296.
- Myers, D. (1987). “Anonymity is part of the magic”: Individual manipulation of computer-mediated communication contexts. *Qualitative Sociology*, 10(3), 251–266.
- Nambisan, P. (2011). Information seeking and social support in online health communities: impact on patients’ perceived empathy. *Journal of the American Medical Informatics Association*, 18(3), 298–304.

- National Health and Medical Research Council. (2015). *National Statement on Ethical Conduct in Human Research 2007 (Updated May 2015)*. Australian Research Council. Canberra, Australia.
- Neale, A. C. (2002). *More Delicate TRANSITIVITY: Extending the PROCESS TYPE for English to include full semantic classifications* (Unpublished PhD thesis, Cardiff University, Cardiff, UK).
- Nivre, J. (2015). Towards a universal grammar for natural language processing. In *Computational Linguistics and Intelligent Text Processing* (pp. 3–16). Springer.
- Ochs, E. (1991). Socialization through language and interaction: A theoretical introduction. *Issues in Applied Linguistics*, 2(2), 143–147.
- Ochs, E., Schegloff, E. A., & Thompson, S. A. (1996). Introduction. In E. Ochs, E. A. Schegloff, & S. A. Thompson (Eds.), *Interaction and grammar* (Vol. 13, pp. 1–51). Cambridge University Press.
- O'Donnell, M. (2014, February). [sys-func] request for sources. Retrieved February 5, 2014, from <http://listserv.uts.edu.au/pipermail/sys-func/2014-February/000791.html>
- O'Donnell, M., & Bateman, J. A. (2005). SFL in computational contexts: a contemporary history. *Continuing Discourse on Language: A functional perspective*, 343–382.
- O'Donnell, M., Zappavigna, M., & Whitelaw, C. (2009). A survey of process type classification over difficult cases. In C. Jones, & E. Ventola (Eds.), *New Developments in the Study of Ideational Meaning: From Language to Multimodality* (pp. 47–64). London, UK: Continuum Publishers.
- O'Halloran, K. L., E., M. K. L., Podlasov, A., & Tan, S. (2013). Multimodal digital semiotics: the interaction of language with other resources. *Text & Talk*, 33(4–5), 665–690.
- O'Leary, D. E. (2015). Twitter Mining for Discovery, Prediction and Causality: Applications and Methodologies. *Intelligent Systems in Accounting, Finance and Management*, 22(3), 227–247.
- Parks, M., & Floyd, K. (1996). Making friends in cyberspace. *Journal of Computer-Mediated Communication*, 1(4), 80–97.

- Partington, A. (2004). Corpora and discourse, a most congruous beast. In A. Partington, J. Morley, & L. Haarman (Eds.), *Corpora and discourse* (Vol. 1, pp. 11–20). Bern, Switzerland: Peter Lang.
- Partington, A. (2008a). From Wodehouse to the White House: a corpus-assisted study of play, fantasy and dramatic incongruity in comic writing and laughter-talk. *Lodz Papers in Pragmatics*, 4(2), 189–213.
- Partington, A. (2008b). Teasing at the White House: a corpus-assisted study of face work in performing and responding to teases. *Text & Talk*, 28(6), 771–792.
- Partington, A. (2010). Modern Diachronic Corpus-Assisted Discourse Studies (MDCADS) on UK newspapers: an overview of the project. *Corpora*, 5(2), 83–108.
- Partington, A. (2011). “Double-speak” at the White House: A corpus-assisted study of bisociation in conversational laughter-talk. *International Journal of Humor Research*, 24(4), 371–398.
- Partington, A. (2013). Corpus Analysis of Political Language. In C. A. Chapelle (Ed.), *The encyclopedia of applied linguistics* (pp. 1–8). Wiley Online Library.
- Paul, M. J., Sarker, A., Brownstein, J. S., Nikfarjam, A., Scotch, M., Smith, K. L., & Gonzalez, G. (2016). Social media mining for public health monitoring and surveillance. In *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing* (Vol. 21, p. 468).
- Paulus, T. M., & Varga, M. A. (2015). ‘Please know that you are not alone with your pain’: Responses to newcomer posts in an online grief support forum. *Death studies*, 39(10), 633–640.
- Pederson, S., & Smithson, J. (2010). Supporting or stressing out? A study of membership, activity and interactions in an online parenting community. In R. Taiwo (Ed.), *Handbook of research on discourse behavior and digital communication: Language structures and social interaction* (pp. 88–103). Hershey, PA: IGI Global.
- Pérez, F., & Granger, B. E. (2007). Ipython: a system for interactive scientific computing. *Computing in Science & Engineering*, 9(3), 21–29.
- Pfeil, U., Svangstu, K., Ang, C. S., & Zaphiris, P. (2011). Social roles in an online support community for older people. *Intl. Journal of Human–Computer Interaction*, 27(4), 323–347.

- Piotti, S. (2014). *Exploring corporate rhetoric in English: Hedging in company annual reports: A corpus-assisted analysis*. Milan, Italy: EDUCatt Università Cattolica.
- Polat, N. (2011). Nature and Content of L2 Socialization Patterns and Attainment of a Turkish Accent by Kurds. *Critical Inquiry in Language Studies*, 8(3), 261–288.
- Postmes, T., Spears, R., & Lea, M. (2000). The formation of group norms in computer-mediated communication. *Human communication research*, 26(3), 341–371.
- Preece, J., & Maloney-Krichmar, D. (2005). Online communities: Design, theory, and practice. *Journal of Computer-Mediated Communication*, 10(4).
- Preece, J., Nonnecke, B., & Andrews, D. (2004, March). The top five reasons for lurking: improving community experiences for everyone. *Computers in Human Behavior*, 20(2), 201–223.
- Prentice, S. (2010, July). Using automated semantic tagging in Critical Discourse Analysis: A case study on Scottish independence from a Scottish nationalist perspective. *Discourse & Society*, 21(4), 405–437.
- Ptaszynski, M., Rzepka, R., Araki, K., & Momouchi, Y. (2012). Annotating Syntactic Information on 5.5 Billion Word Corpus of Japanese Blogs. In *Proceedings of The 18th Annual Meeting of The Association for Natural Language Processing (NLP-2012)* (pp. 385–388).
- Pudlinski, C. (1998). Giving advice on a consumer-run warm line: Implicit and dilemmatic practices. *Communication Studies*, 49(4), 322–341.
- Pullinger, D. J. (1986). Chit-chat to electronic journals: Computer conferencing supports scientific communication. *Professional Communication, IEEE Transactions on*, (1), 23–29.
- Quirk, R. (1960). Towards a description of English usage. *Transactions of the philosophical society*, 59(1), 40–61.
- Radenski, A. (2006). Python First: A lab-based digital introduction to computer science. In *ACM SIGCSE Bulletin* (Vol. 38, pp. 197–201). ACM.
- Raghavan, P. (2014). *Medical event timeline generation from clinical narratives*.
- Rayson, P. (2012). Corpus Analysis of Key Words. In C. A. Chapelle (Ed.), *The Encyclopedia of Applied Linguistics* (pp. 1–7). Wiley Online Library.

- Rayson, P., Archer, D., Piao, S., & McEnery, A. (2004). The ucrel semantic analysis system. In *Proceedings of the beyond named entity recognition semantic labelling for nlp tasks workshop* (pp. 7–12). Lisbon, Portugal.
- Reyes, A. (2011). Strategies of legitimization in political discourse: From words to actions. *Discourse & Society*, 22(6), 781–807.
- Roberts, G. L., & Bavelas, J. (1996). The Communicative Dictionary: A Collaborative Theory of Meaning. In *Beyond the symbol model: Reflections on the representational nature of language* (pp. 135–160). Albany: SUNY.
- Roberts, L. D. (2015). Ethical issues in conducting qualitative research in online communities. *Qualitative Research in Psychology*, 12(3), 314–325.
- Rohdenburg, G. (1996). Cognitive complexity and increased grammatical explicitness in English. *Cognitive Linguistics*, 7(2), 149–182.
- Ryder, M., & Wilson, B. (1996). Affordances and Constraints of the Internet for Learning and Instruction. In *Paper presented to a joint session of the association for educational communications technology*. Indianapolis, IN.
- Salahshour, N. (in press). Liquid Metaphors as Positive Evaluations: A Corpus-Assisted Discourse Analysis of the Representation of Migrants in a Daily New Zealand Newspaper. *Discourse, Context & Media*.
- Salama, A. H. (2011, May). Ideological collocation and the recontextualization of Wahhabi-Saudi Islam post-9/11: A synergy of corpus linguistics and critical discourse analysis. *Discourse & Society*, 22(3), 315–342.
- Sandaunet, A.-G. (2008). The challenge of fitting in: non-participation and withdrawal from an online self-help group for breast cancer patients. *Sociology of health & illness*, 30(1), 131–144.
- Schandorf, M. (2013). Mediated gesture: Paralinguistic communication and phatic text. *Convergence: The International Journal of Research into New Media Technologies*, 19(3), 319–344.
- Schieffelin, B. B., & Ochs, E. (1986). Language socialization. *Annual review of anthropology*, 15(1), 163–191.

- Schröter, M., & Storjohann, P. (2015). Patterns of discourse semantics: A corpus-assisted study of financial crisis in British newspaper discourse in 2009. *Pragmatics and Society*, 6(1), 43–66.
- Sharf, B. F. (1997). Communicating breast cancer on-line: support and empowerment on the Internet. *Women & health*, 26(1), 65–84.
- Shen, K. N., & Khalifa, M. (2013). Effects of technical and social design on virtual community identification: a comparison approach. *Behaviour & Information Technology*, 32(10), 986–997.
- Shumaker, S. A., & Brownell, A. (1984). Toward a theory of social support: Closing conceptual gaps. *Journal of social issues*, 40(4), 11–36.
- Sillence, E. (2013). Giving and receiving peer advice in an online breast cancer support group. *Cyberpsychology, Behavior, and Social Networking*, 16(6), 480–485.
- Sillence, E., Hardy, C., & Briggs, P. (2013). Why don't we trust health websites that help us help each other? An analysis of online peer-to-peer healthcare. In *Proceedings of the 5th annual acm web science conference* (pp. 396–404). New York, NY: ACM.
- Sillence, E., & Mo, P. K. (2012). Communicating health decisions: An analysis of messages posted to online prostate cancer forums. *Health Expectations*, 244–253.
- Simon-Vandenbergen, A. M., Ravelli, L., & Taverniers, M. (2003). *Grammatical metaphor: views from systemic functional linguistics*. Amsterdam, The Netherlands: John Benjamins.
- Sinclair, J. (2001). Preface. In M. Ghadessy, A. Henry, & R. L. Roseberry (Eds.), *Small corpus studies and elt* (pp. vii–xv). Amsterdam, The Netherlands: John Benjamins.
- Sinclair, J. (2004). *Trust the text: Language, corpus and discourse*. London, UK: Routledge.
- Sinclair, J. M. (1997). Corpus evidence in language description. In *Teaching and language corpora* (pp. 27–39).
- Slade, D., Chandler, E., Pun, J., Lam, M., Matthiessen, C., Williams, G., ... Tang, S. Y. H. (2015). Effective healthcare worker-patient communication in Hong

- Kong accident and emergency departments. *Hong Kong Journal of Emergency Medicine*, 22(2), 69–83.
- Slade, D., Manidis, M., McGregor, J., Scheeres, H., Chandler, E., Stein-Parbury, J., ... Matthiessen, C. M. (2015a). *Communicating in hospital emergency departments*. London, UK: Springer.
- Slade, D., Manidis, M., McGregor, J., Scheeres, H., Chandler, E., Stein-Parbury, J., ... Matthiessen, C. M. (2015b). The Role of Communication in Safe and Effective Health Care. In *Communicating in Hospital Emergency Departments* (pp. 1–23). London, UK: Springer.
- Slade, D., Scheeres, H., Manidis, M., Iedema, R., Dunston, R., Stein-Parbury, J., ... McGregor, J. (2008). Emergency communication: the discursive challenges facing emergency clinicians and patients in hospital emergency departments. *Discourse & Communication*, 2(3), 271–298.
- Smithson, J., Jones, R. B., & Ashurst, E. (2012). Developing an online learning community for mental health professionals and service users: a discursive analysis. *BMC Medical Education*, 12(1), 2–10.
- Smithson, J., Sharkey, S., Hewis, E., Jones, R. B., Emmens, T., Ford, T., & Owens, C. (2011a). Membership and boundary maintenance on an online self-harm forum. *Qualitative Health Research*, 21(11), 1567–1575.
- Smithson, J., Sharkey, S., Hewis, E., Jones, R., Emmens, T., Ford, T., & Owens, C. (2011b). Problem presentation and responses on an online forum for young people who self-harm. *Discourse Studies*, 13(4), 487–501.
- St Louis, C., & Zorlu, G. (2012). Can Twitter predict disease outbreaks? *British Medical Journal*, 344, 1–3.
- Stevens, G., O'Donnell, V. L., & Williams, L. (2015). Public domain or private data? Developing an ethical approach to social media research in an inter-disciplinary project. *Educational Research and Evaluation*, 21(2), 154–167.
- Stewart, M. A. (1995). Effective physician-patient communication and health outcomes: a review. *Canadian Medical Association Journal*, 152(9), 1423–1433.
- Stommel, W., & Koole, T. (2010). The online support group as a community: A micro-analysis of the interaction with a new member. *Discourse Studies*, 12(3), 357–378.

- Stommel, W., & Meijman, F. (2011). The use of conversation analysis to study social accessibility of an online support group on eating disorders. *Global health promotion*, 18(2), 18–26.
- Stubbs, M. (2004). Language Corpora. In A. Davies, & C. Elder (Eds.), *The handbook of applied linguistics* (pp. 106–132). Wiley Online Library.
- Stubbs, M. (2001). Texts, corpora, and problems of interpretation: A response to Widdowson. *Applied Linguistics*, 22(2), 149–172.
- Sullivan, C. F. (2003). Gendered cybersupport: A thematic analysis of two online cancer support groups. *Journal of health psychology*, 8(1), 83–104.
- Swan, M., Lau, A., & Bromberg, H. (2010). Sharing Health Data for Better Outcomes on PatientsLikeMe. *Journal of Medical Internet Research*, 12(2), 12–16.
- Tanis, M., & Postmes, T. (2007). Two faces of anonymity: Paradoxical effects of cues to identity in CMC. *Computers in Human Behavior*, 23(2), 955–970.
- Taverniers, M. (2002). *Systemic-Functional Linguistics and the Notion of Grammatical Metaphor* (Doctoral dissertation, PhD Thesis. Dept. of English, University of Ghent).
- Teich, E., Degaetano-Ortlieb, S., Fankhauser, P., Kermes, H., & Lapshinova-Koltunski, E. (2015). The linguistic construal of disciplinarity: A data-mining approach using register features. *Journal of the Association for Information Science and Technology*, 1668–1678.
- Thompson, G., & Collins, H. (2001). Interview with mak halliday, cardiff, july 1998. *DELTA*, 17(1), 131–153.
- Thompson, G., & Hunston, S. (2014). System and corpus: Two traditions with a common ground. *Equinox Books*, 1–14.
- Thompson, J., Bissell, P., Cooper, C., Armitage, C. J., & Barber, R. (2012). Credibility and the ‘professionalized’ lay expert: Reflections on the dilemmas and opportunities of public involvement in health research. *Health*, 16(6), 602–618.
- Thorne, S. L. (2008). Computer-mediated communication. *Encyclopedia of language and education*, 4, 325–336.
- Tognini-Bonelli, E. (2001). *Corpus linguistics at work*. Amsterdam, The Netherlands: John Benjamins.

- Van Dijk, T. A. (2004). Text and context of parliamentary debates. *Cross-cultural perspectives on parliamentary discourse*, 339–372.
- Van Kleek, M., Smith, D., Shadbolt, N. R., Murray-Rust, D., & Guy, A. (2015). Self cur-
ation, social partitioning, escaping from prejudice and harassment: the many
dimensions of lying online. In *Proceedings of the 24th International Conference on
World Wide Web Companion* (pp. 371–372). International World Wide Web Con-
ferences Steering Committee.
- Van Leeuwen, T. (1996). The representation of social actors. In C. R. Caldas-Coulthard,
& M. Coulthard (Eds.), *Texts and practices: Readings in critical discourse analysis*
(pp. 32–70). London.
- Van Leeuwen, T. (2007). Legitimation in discourse and communication. *Discourse &
Communication*, 1(1), 91–112.
- Varga, M. A., & Paulus, T. M. (2014). Grieving online: Newcomers' constructions of
grief in an online support group. *Death studies*, 38(7), 443–449.
- Vayreda, A., & Antaki, C. (2009). Social support and unsolicited advice in a bipolar
disorder online forum. *Qualitative Health Research*, 19(7), 931–942.
- Veel, R. (1997). Learning how to mean—scientifically speaking: Apprenticeship into
scientific discourse in the secondary school. *Genre and institutions: Social pro-
cesses in the workplace and school*, 161–195.
- Velupillai, S., Mowery, D., South, B. R., Kvist, M., & Dalianis, H. (2015). Recent
advances in clinical natural language processing in support of semantic analysis.
Yearbook of medical informatics, 10(1), 183.
- Virtanen, T. (2009, October). Discourse linguistics meets corpus linguistics: the-
oretical and methodological issues in the troubled relationship. *Language &
Computers*, 69(1), 49–65.
- Walther, J. B. (1996). Computer-mediated communication: Impersonal, interper-
sonal, and hyperpersonal interaction. *Communication research*, 23(1), 3–43.
- Walther, J. B. (2002). Research ethics in Internet-enabled research: Human subjects
issues and methodological myopia. *Ethics and information technology*, 4(3), 205–
216.

- Wang, Y.-C., Kraut, R., & Levine, J. M. (2012). To stay or leave?: the relationship of emotional and informational support to commitment in online health support groups. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work* (pp. 833–842). ACM.
- Warren, M. (2012). Corpora: Specialized. In C. A. Chapelle (Ed.), *The Encyclopedia of Applied Linguistics* (pp. 1191–1200). Wiley Online Library.
- Wasfy, J. H., Singal, G., O'Brien, C., Blumenthal, D. M., Kennedy, K. F., Strom, J. B., ... Yeh, R. W. (2015). Enhancing the prediction of 30-day readmission after percutaneous coronary intervention using data extracted by querying of the electronic health record. *Circulation: Cardiovascular Quality and Outcomes*, 8(5), 477–485.
- Weber, H. (2011). Missed cues: How disputes can socialize virtual newcomers. *Language@Internet*, 8.
- West, L. E. (2010). *Facework on Facebook: How it legitimizes community membership and enables linguistic socialization through intertextuality* (Doctoral dissertation, Georgetown University).
- Widdowson, H. G. (1991). The description and prescription of language. *Georgetown University round table on languages and linguistics, 1991*, 11–24.
- Widdowson, H. G. (2000). On the limitations of linguistics applied. *Applied linguistics*, 21(1), 3–25.
- Widdowson, H. G. (2008). *Text, Context, Pretext: Critical Issues in Discourse Analysis*. Malden, MA: Blackwell Publishing.
- Williams, P., & Weninger, C. (2013). Applying Goffman's assumptions about communication to a new media environment. In *The Present and Future of Symbolic Interactionism. Proceedings of the International Symposium, Pisa 2010* (Vol. 2, p. 47).
- Wilson, P. M., Kendall, S., & Brooks, F. (2007). The Expert Patients Programme: a paradox of patient empowerment and medical dominance. *Health & social care in the community*, 15(5), 426–438.
- Wolf, Z. R. (1989). Learning the professional jargon of nursing during change of shift report. *Holistic nursing practice*, 4(1), 78–83.

- Woodward-Kron, R. (2016). International medical graduates and the discursive patterns of patient-centred communication. *Language Learning in Higher Education*, 6(1), 253–273.
- Wu, S. S. (2013). *Is CMC the new FTF: a study exploring the nature of computer mediated communication on Facebook* (Doctoral dissertation, California Polytechnic State University).
- Yan, L. L., & Tan, Y. (2015). Good Intentions, Bad Outcomes: The Effects of Mismatches in Social Support and Health Outcomes in an Online Weight Loss Community. *Kelley School of Business Research Paper No. 16-7*, 1–32.
- Yang, T.-I., Torget, A. J., & Mihalcea, R. (2011). Topic modeling on historical newspapers. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities* (pp. 96–104). Association for Computational Linguistics.
- Yao, T., Zheng, Q., & Fan, X. (2015). The Impact of Online Social Support on Patients' Quality of Life and the Moderating Role of Social Exclusion. *Journal of Service Research*, 18, 369–383.
- Yesha, R., & Gangopadhyay, A. (2015). A method for analyzing health behavior in online forums. In *Proceedings of the 6th ACM Conference on Bioinformatics, Computational Biology and Health Informatics* (pp. 615–621). ACM.
- Zappavigna, M. (2011). Ambient affiliation: A linguistic perspective on Twitter. *New Media & Society*, 13(5), 788–806.
- Zappavigna, M. (2012). *Discourse of Twitter and social media: How we use language to create affiliation on the web*. Sydney, Australia: Bloomsbury.
- Zappavigna, M. (2013). Enacting identity in microblogging through ambient affiliation. *Discourse & Communication*, 209–228.
- Zhang, W., & Storck, J. (2001). Peripheral members in online communities. In *Proceedings of AMCIS 2001 the Americas Conference on Information Systems*. Boston, MA.
- Ziebland, S., Chapple, A., Dumelow, C., Evans, J., Prinjha, S., & Rozmovits, L. (2004). How the internet affects patients' experience of cancer: a qualitative study. *British Medical Journal*, 328(7439), 564.

- Zimmer, M. (2010). “But the data is already public”: on the ethics of research in Facebook. *Ethics and information technology*, 12(4), 313–325.
- Zimmerman, D. H. (1998). Identity, context and interaction. In C. Antaki (Ed.). S. Widdicombe (Ed.), *Identities in talk* (pp. 87–106). Sage Publications Ltd.
- Zinn, J. O., & McDonald, D. (2015). Changing Discourses of Risk and Health Risk. In *Medicine, Risk, Discourse and Power* (pp. 207–240). New York, NY: Routledge.

Glossary

A number of theoretical definitions provided here echo those used in SFL generally. Other systemic-functional terminology has not been glossed, simply because concise definitions are already available. Readers are directed to Matthiessen, Teruya, & Lam, 2010 for an overview of key terms.

Bipolar Disorder

Bipolar Disorder is a mental disorder characterised by oscillation between manic and depressive states. The thesis is not concerned with the phenomenology of the illness itself, making a more specific definition unnecessary here.

Consumer

Consumer is used to refer to people who utilise healthcare services, ranging from consultations with healthcare professionals to reading information about health online. The term is used in preference to *patient*, which refers more specifically to those who are undergoing ongoing treatment in formal institutions (hospitals, psychiatric hospitals, clinics, etc.), and which construes the consumer as a (passive) recipient of medical services, rather than a collaborator in the process of treatment. *Patient* is also less accurate when referring to the members of the forum, as some members are not pursuing treatment through formal medical channels, but are nonetheless consuming healthcare information

Consumer-centredness

Synonymous with *patient-centredness*, though *consumer* is used here for consistency. *Consumer-centredness* refers to a model of medical treatment that

seeks to increase the agency of healthcare consumers in their own journeys through healthcare institutions. Under this model, consumers' attitudes, beliefs and emotional needs are prioritised, and consumers are encouraged to collaborate in decision-making processes. Another key aim is the fostering of an ongoing dialogue between professional and consumer.

Consumer-centredness can increase consumer satisfaction, adherence to treatment plans, and lead to better health outcomes. Noting the importance of successful interpersonal exchange between healthcare professional and healthcare consumer, Matthiessen (2013) prefers *relationship-centredness*.

Corpus

A collection of linguistic text, which is almost always large and digitised.

Corpus linguistics (CL)

In this thesis, *CL* primarily denotes Corpus Linguistics—a branch of linguistics centred on quantitative analysis of digitised texts. In related and overlapping literature, *CL* may denote *computational linguistics*. Because there is no clear line between corpus and computational linguistics, however, and because the methods used in this thesis in many respects blur boundaries between the two, references to *CL* can be understood as acknowledging both corpus and computational theory and methods. When disambiguation is necessary, the full names are used.

Discourse-semantics

The stratum of function and meaning in language. It encompasses (interpersonal) pragmatics and ideation, as well as text organisation, which is not treated in detail in this thesis.

Extensible Markup Language (XML)

An HTML-like markup language designed to store arbitrary information in a human-readable and machine-readable format.

Forum

Forum is the main term used to denote the site of the study, a large online Bipolar Disorder message board. *Forum*, *board*, *community*, and *group* are used interchangeably, just as they are by forum members.

forum

Without an initial capital, *forum*/*forums* refers to online message boards—text-based *modes* of CMC where registered users can create and reply to threads. The site of the investigation, the *Bipolar Forum*, is a fairly prototypical example.

HyperText Markup Language (HTML)

The main markup language for creating webpages. The Forum was downloaded as HTML. Texts and metadata were extracted from the HTML using `lxml`.

Lexicogrammar

The stratum of words and wordings in language. A single system, with grammar at the broad end and lexis as the delicate end.

Member

Somebody who has signed up and/or contributed to an online community. Most often, in this thesis, this refers to those who contribute to the Bipolar Forum. *Members*, *users*, *contributors* and *participants* are used interchangeably.

mode

Uncapitalised, *mode* refers to a constellation of medium factors that together realise a culturally recognised variety of CMC. A web forum is therefore a mode of CMC, as is an instant messenger, a wiki talk page, a Skype video call, etc.

Mode

With an initial capital, *Mode* refers to the systemic-functional register dimension, which broadly corresponds to *the role played by language in the text*.

Online support group (OSG)

Online support groups are online spaces in which people can exchange information about health related topics, and to offer social support.

Post

A single message within a thread. Can be used as a verb to describe the process of authoring/transmitting a message. *Contribution* is sometimes used.

Systemic Functional Grammar (SFG)

A detailed description of the discourse-semantics, lexicogrammar and phonology of a language that draws upon systemic-functional categories and the fits within a systemic-functional conceptualisation of language and context. Such grammars are functional-semantic, in that the functions and meanings made by language are used as the basis for grammatical distinctions. In the case of this thesis, SFG refers to the grammar of English developed chiefly by Halliday, as presented in Halliday and Matthiessen 2004 and elsewhere. The phonological/graphological stratum is not considered in this work .

theme

Uncapitalised, *theme* is used to mean a salient, recurring kind of meaning being made in a text (as in thematic analysis).

Theme

With an initial capital, *Theme* refers to the leftmost group in a clause, as per the systemic functional grammar.

THEME

In small caps, THEME denotes the system incorporating *Theme*, as per the systemic functional grammar.

Thread

A ‘discussion’ within the Forum, initiated by a single user. The forum presents a list of threads, organised by date of last post.

List of abbreviations

API	Application programming interface
CA	Conversation analysis
CADS	Corpus assisted discourse studies
CDA	Critical discourse analysis
CL	Corpus Linguistics
CMC	Computer mediated communication
CWB	Corpus Workbench
DSM	Diagnostic and Statistical Manual of Mental Disorders
GUI	Graphical user interface
HC	Healthcare communication
HTML	HyperText Markup Language
ICD	International Classification of Diseases
NLP	Natural language processing
OSG	Online support group
POS	Part-of-speech
SFG	Systemic Functional Grammar
SFL	Systemic functional linguistics
XML	Extensible Markup Language

Appendices

A. Jess' first post

Below is the thread analysed in Chapter 5. It is presented multimodally, in a form similar to its original appearance, but with some identifying details obscured.

It appears you have not yet Signed Up with our community. To Sign Up for free, please click here....



SOBER SPRING

ADDICTION AND RECOVERY

ROBERT F.
BOLLENDORF

Bipolar Disorder Message Board

[HealthBoards](#) > [Mental Health Board](#) > [Bipolar Disorder](#) > am i bipolar and what should i do?

am i bipolar and what should i do?



Subscribe To Bipolar Disorder

LinkBack ▾ ▾ ? Thread Tools ▾ ▾ ? Search this Thread ▾ ▾ ?

#1

05-29-2011, 07:13 AM

jessff1989787
Newbie
(female)

Join Date: May 2011
Location: england staffordshire
Posts: 4



am i bipolar and what should i do?

hi im amy new to this site, umm... well im currently 20 years old and have been diagnosed with depression from a young age but in the last 5 years or so i have been feeling very odd having some extreme highs which include loss of appetite concentration using drugs and alcohol spending sprees and also sex, i can not control this when i get impulses like the above it is impossible though i have tried. I also suffer with depression which is quite severe most of the time i find it hard to get out of bed or to even be able to connect with anyone including my partner who lives with me, i am hurting him so much but i dont feel like i can do anytjing about it

i have asked my doctor to test me to see if i am bipolar as my antidepressants do not work even though they have been changed a million times!! he said no that he wont test me and i also asked for counselling and he also declined that, at the moment i feel that im loosing control of everything and its getting worse, i want to change my doctor and have been telling my partner i would but im scared of finding out that i am bipolar, but i really feel like either an extreme high or an extreme low is on the way and im quite scared i dont know what to do
please help

5 Foods That Destroy Testosterone (Avoid)



Cut down a bit of stomach fat every day by never eating these 5 foods

[Never Eat>>](#)

V-Taper Solution



The following 2 users give hugs of support to:
jessff1989787
jacq42 (05-29-2011),fairygirl32 (05-29-2011)

Sponsors

5 Foods That Destroy Testosterone (Avoid)



Cut down a bit of stomach fat every day by never eating these 5 foods

[Never Eat>>](#)

V-Taper Solution

Sign Up Today!

Ask our community of thousands of members your health questions, and learn from others experiences. Join the conversation!

I WANT MY FREE ACCOUNT

05-29-2011, 09:15 AM #2

Luvsoccer
Newbie
(female)

Join Date: May 2011
Location: Arkansas
Posts: 6

Re: am i bipolar and what should i do?

It sounds like you might have bipolar to me. You need to change Drs. One with more knowledge apparently. The reason the antidepressants are not Working is because If you are bi polar and they put U on an antidepressant alone it can make things worse... I know from experience. Don't be scared it's treatable. Find you a dr that can make a correct Diagnosis and go from there. Good luck!

The Following User Says Thank You to Luvsoccer For This Useful Post:
jessff1989787 (05-30-2011)

Treato

See what patients are saying about:

Bipolar Disorder

more info

Top Issues Medications

Click on any issue to view related posts

Issue	Percentage
Depression	High
Mania	Medium
Anxiety	Medium
Mental Disorder	Low
Pain	Low

T&C view all

05-29-2011, 09:41 AM #3

jacq42
Member
(female)

Join Date: Dec 2009
Location: uk
Posts: 79

Re: am i bipolar and what should i do?

Hi first thing I would like to suggest is changing your doctor(obviously your choice),my background is very similar to yours and I was being given antidepressants for clinical depression for around 18yrs,and believe me I've tried them all. It wasn't until I went to see my usual GP and they had another doctor there that it was noticed that I have bipolar. I could not believe what I was hearing,along with what I was describing (identical to what you said in your post) and the dr was saying it made perfect sense,so now I'm awaiting my new medication etc.

Why on earth is your doctor not helping you is beyond me hence the suggestion for a second opinion from a doctor you've never seen before. I think a lot of gps to be honest,really don't have such a vast amount of knowledge of mental illness.

My new doctor won't medicate me or anything until I've seen the psychiatrist again due to other health issues.

Sorry to ramble on but you sound so similar to me and it does get so frustrating,just remember everyones different and its your choice if you want to seek help from a different doctor,hope you keep well and I have helped you a little

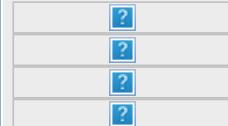
Jacq

The Following User Says Thank You to jacq42 For This Useful Post:
jessff1989787 (05-30-2011)

05-29-2011, 09:43 AM #4

ghelpmelivelife
Senior Member
(female)

Join Date: May 2011
Location: North Palm Beach
Florida
Posts: 104



Re: am i bipolar and what should i do?

I would strongly suggest you do change doctors! Him declining to test you for bi-polar when it is a major concern for you and the therapy as well sounds not good of him.

I hope you do find another doctor and you get the help you need!
good luck 😊



The Following User Says Thank You to ghelpmelivelife For This Useful Post:
jessff1989787 (05-30-2011)

05-29-2011, 11:01 AM

#5

fairygirl32
Senior Member
(female)

Join Date: Sep 2010
Location: Virginia
Posts: 241



Re: am i bipolar and what should i do?

I think you should get a second opinion ASAP!!!! I can't believe that your dr. won't even consider therapy for you. I know he is your dr. but YOU know your body better than he/she does and it is telling you something is NOT right and it needs attention NOW so PLEASE take some action. We all on this board care about each other so PLEASE keep us posted. Until next time TAKE CARE!!!!😊



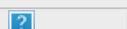
The Following User Says Thank You to fairygirl32 For This Useful Post:
jessff1989787 (05-30-2011)

05-29-2011, 12:45 PM

#6

girlegracing7
Junior Member
(female)

Join Date: Jan 2009
Location: Womelsdorf, Pa
Posts: 16



Re: am i bipolar and what should i do?

Hello and welcome! I would seriously suggest finding a new doctor. do not be scared to find out if you are bipolar. If you find out that you are then the doctor can get you proper help and medication. You have come to a great place for support! i wish you the best of luck and please keep us updated! it will be ok and you can get through this it maybe hard at first but eventually it does get better!



The following user gives a hug of support to girlegracing7:
fairygirl32 (05-30-2011)



The Following User Says Thank You to girlegracing7 For This Useful Post:
jessff1989787 (05-30-2011)

05-29-2011, 01:42 PM

#7

Emz45
Inactive
(female)

Join Date: Mar 2008
Posts: 5,063



Re: am i bipolar and what should i do?

Hi, welcome to the boards, hopefully we can help you out and be a support system for you. First off you're not A bipolar, *!* we're not things, it's a condition. From what you say, if sounds very likely that you might have Bipolar disorder. I would go to a psychiatrist for testing and find out. You don't have to have your docs permission to do this. If you're already seeing a psychiatrist and that's who's doing all the denying, then find a different one, because he's not doing his job, nor is he considering your best interest. That's all that you can really do in the beginning, find out what's what. We aren't docs here and can't diagnose you. But I think it would definitely be smart to go and get a diagnoses.

Take care, and please keep in touch, let us know how you're doing, okay?

Kat



The Following User Says Thank You to Emz45 For This Useful Post:

fairygirl32 (05-30-2011)

[?]

Similar Threads

[?]

Thread	Thread Starter	Board	Replies	Last Post
is_it_bipolar?	punkrocker89	Bipolar Disorder	9	09-24-2007 09:55 PM
bipolar/panic disorder	e0a54	Bipolar Disorder	5	07-10-2006 02:02 PM
My Daughters Bipolar...	joessy	Bipolar Disorder	2	06-06-2006 08:13 PM
Help! Confrontation w/ bipolar daughter	liz49	Family & Friends of the Mentally Ill	11	05-15-2006 12:35 AM
Girlfriend's Father is Bipolar	Panna_06	Family & Friends of the Mentally Ill	8	04-20-2006 05:38 PM
BiPolar Vs Borderline	BorderChild	Bipolar Disorder	5	04-20-2006 06:52 AM
BiPolar I vs BiPolar II	ThornyRose	Bipolar Disorder	2	04-05-2006 06:29 AM
type two bipolar....	skpgh152	Bipolar Disorder	13	08-18-2005 01:58 PM
Attention Newbies! Please Read: Bipolar Disorder Symptom Primer	reesie	Bipolar Disorder	0	06-06-2005 08:51 AM

Search HealthBoards

Board Search

« [could use some direction!](#) | [Lamictal](#) »

Posting Rules

[?]

You **may not** post new threads
You **may not** post replies
You **may not** post attachments
You **may not** edit your posts

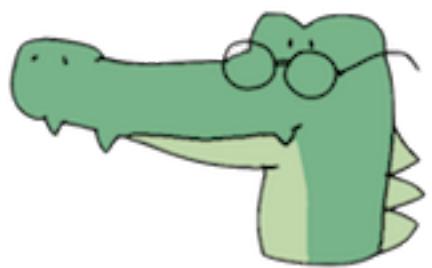
[BB code](#) is **On**
[Smilies](#) are **On**
[\[IMG\]](#) code is **Off**
HTML code is **Off**
Trackbacks are **Off**
Pingbacks are **Off**
Refbacks are **Off**

[Forum Rules](#)

B. corpkit documentation

Overleaf is a manual for the `corpkit` API, automatically generated at the time of printing from the source code and online documentation. The HTML version is available at *ReadTheDocs* (<http://corpkit.readthedocs.org>). An **Epub** version can be built by visiting <http://readthedocs.org/projects/corpkit/downloads/epub/latest/>. The manual contains a short introduction, followed by a user guide and an API reference.

The source code for the module is not reproduced here. It can be found hosted on GitHub. Both source code and documentation are still evolving, and, as they are open-source, may be modified by other authors in the future.



corpkit documentation

Release 2.3.8

Daniel McDonald

Dec 05, 2016

1 Creating projects and building corpora	5
1.1 Creating a new project	5
1.2 Adding a corpus	6
1.3 Creating a Corpus object	6
1.4 Pre-processing the data	6
1.5 Manipulating a parsed corpus	7
1.6 Counting key features	7
2 Interrogating corpora	9
2.1 Introduction	9
2.2 Search types	11
2.3 Grammatical searching	11
2.4 Excluding results	12
2.5 What to show	12
2.6 Working with trees	13
2.7 Tree <i>show</i> values	13
2.8 Working with dependencies	14
2.9 Working with metadata	14
2.10 Working with coreferences	14
2.11 Multiprocessing	15
2.12 N-grams	15
2.13 Collocation	15
2.14 Saving interrogations	15
2.15 Exporting interrogations	16
2.16 Other options	16
3 Concordancing	17
3.1 Generating a concordance	17
3.2 Displaying concordance lines	18
3.3 Working with concordance lines	18
3.4 The <i>calculate</i> method	18
4 Editing results	21
4.1 Keeping or deleting results and subcorpora	21
4.2 Editing result names	22
4.3 Spelling normalisation	22
4.4 Generating relative frequencies	22
4.5 Keywording	23
4.6 Sorting	23
4.7 Calculating trends, P values	24
4.8 Saving results	24
4.9 Exporting results	24
4.10 Next step	24
5 Visualising results	25

5.1	Basics	25
5.2	Plot type	26
5.3	Plot style	27
5.4	Figure and font size	27
5.5	Title and labels	27
5.6	Subplots	27
5.7	TeX	28
5.8	Legend	28
5.9	Colours	28
5.10	Saving figures	28
5.11	Other options	28
5.12	Multiplotting	29
6	Using language models	31
6.1	Customising models	31
6.2	Compare subcorpora	32
6.3	Advanced stuff	32
7	Managing projects	33
7.1	Loading saved data	33
7.2	Managing multiple corpora	33
7.3	Using the GUI	35
8	Overview	37
8.1	Objects	37
8.2	Commands	38
8.3	Prompt features	39
9	Setup	41
9.1	Dependencies	41
9.2	Accessing	41
9.3	The prompt	41
10	Making projects and corpora	43
10.1	Adding a corpus	43
10.2	Parsing a corpus	43
10.3	Tokenising, POS tagging and lemmatising	43
10.4	Working with metadata	43
11	Interrogating corpora	45
11.1	Search examples	45
11.2	Working with metadata	46
11.3	Sampling a corpus	46
12	Concordancing	47
12.1	Customising appearance	47
12.2	Sorting	47
12.3	Colouring	47
12.4	Editing	48
12.5	Recalculating results from concordance lines	48
12.6	Working with metadata	48
13	Annotating your corpus	49
13.1	Tagging sentences	49
13.2	Creating fields and values	49
13.3	Removing annotations	50
14	Editing results	51
14.1	The edit command	51
14.2	Doing basic statistics	51

14.3	Sorting results	51
15	Plotting	53
16	Settings and management	55
16.1	Managing data	55
16.2	Toggles and settings	55
16.3	Switching to IPython	55
16.4	Running scripts	56
17	Corpus classes	57
17.1	<i>Corpus</i>	57
17.2	<i>Corpora</i>	62
17.3	<i>Subcorpus</i>	63
17.4	<i>File</i>	63
17.5	<i>Datalist</i>	64
18	Interrogation classes	65
18.1	<i>Interrogation</i>	65
18.2	<i>Interrodict</i>	71
18.3	<i>Concordance</i>	73
19	Functions	75
19.1	<i>as_regex</i>	75
19.2	<i>load</i>	75
19.3	<i>load_all_results</i>	75
19.4	<i>new_project</i>	76
20	Wordlists	77
20.1	Closed class word types	77
20.2	Systemic functional process types	77
20.3	Stopwords	77
20.4	Systemic/dependency label conversion	77
20.5	BNC reference corpus	77
20.6	Spelling conversion	78

corpkit is a Python-based tool for doing more sophisticated corpus linguistics. It exists as a graphical interface, a Python API, and a natural language interpreter. The API and interpreter are documented here.

With *corpkit*, you can create parsed, structured and metadata-annotated corpora, and then search them for complex lexicogrammatical patterns. Search results can be quickly edited, sorted and visualised, saved and loaded within projects, or exported to formats that can be handled by other tools. In fact, you can easily work with any dataset in **CONLL U** format, including the freely available, multilingual [Universal Dependencies Treebanks](#).

Concordancing is extended to allow the user to query and display grammatical features alongside tokens. Key-wording can be restricted to certain word classes or positions within the clause. If your corpus contains multiple documents or subcorpora, you can identify keywords in each, compared to the corpus as a whole.

corpkit leverages [Stanford CoreNLP](#), [NLTK](#) and [pattern](#) for the linguistic heavy lifting, and [pandas](#) and [matplotlib](#) for storing, editing and visualising interrogation results. Multiprocessing is available via [joblib](#), and Python 2 and 3 are both supported.

API example

Here's a basic workflow, using a corpus of news articles published between 1987 and 2014, structured like this:

```
./data/NYT:
+--1987
|   |--NYT-1987-01-01-01.txt
|   |--NYT-1987-01-02-01.txt
|   ...
|
+--1988
|   |--NYT-1988-01-01-01.txt
|   |--NYT-1988-01-02-01.txt
|   ...
...
```

Below, this corpus is made into a *Corpus* object, parsed with *Stanford CoreNLP*, and interrogated for a lexicogrammatical feature. Absolute frequencies are turned into relative frequencies, and results sorted by trajectory. The edited data is then plotted.

```
>>> from corpkit import *
>>> from corpkit.dictionaries import processes

### parse corpus of NYT articles containing annual subcorpora
>>> unparsed = Corpus('data/NYT')
>>> parsed = unparsed.parse()

### query: nominal nsubjps that have verbal process as governor lemma
>>> crit = {F: r'^nsubj$', 
...           GL: processes.verbal.lemmata,
...           P: r'^V'}

### interrogate corpus, outputting lemma forms
>>> sayers = parsed.interrogate(crit, show=L)
>>> sayers.quickview(10)

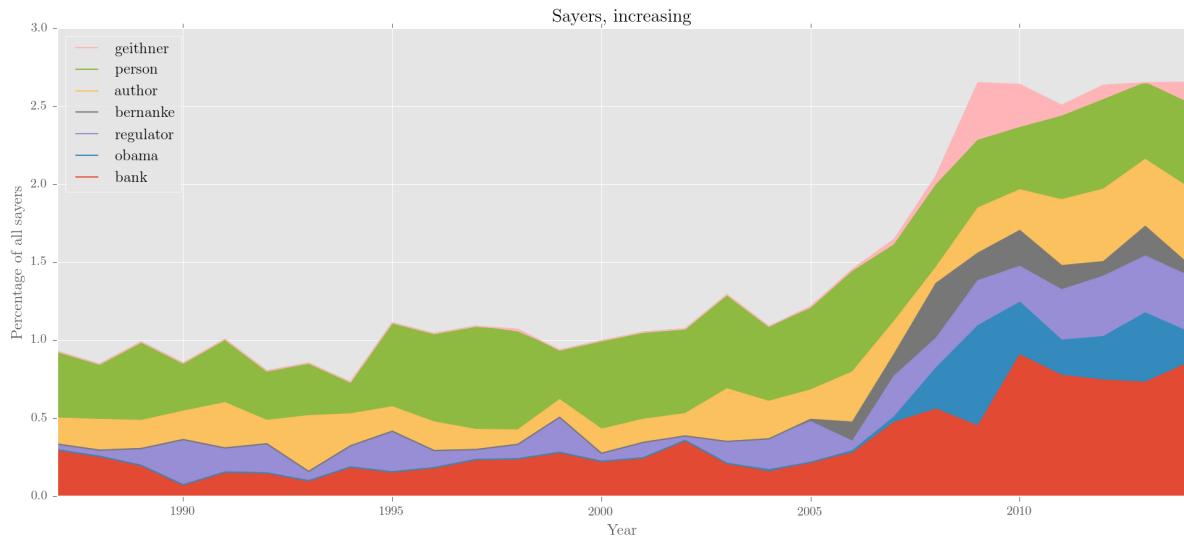
0: official      (n=4348)
1: expert        (n=2057)
2: analyst       (n=1369)
3: report         (n=1103)
4: company        (n=1070)
5: which          (n=1043)
6: researcher     (n=987)
7: study          (n=901)
8: critic          (n=826)
9: person          (n=802)

### get relative frequency and sort by increasing
>>> rel_say = sayers.edit('%', SELF, sort_by='increase')

### plot via matplotlib, using tex if possible
```

```
>>> rel_say.visualise('Sayers, increasing', kind='area',
...                      y_label='Percentage of all sayers')
```

Output:



Installation

Via pip:

```
$ pip install corpkit
```

via Git:

```
$ git clone https://www.github.com/interrogator/corpkit
$ cd corpkit
$ python setup.py install
```

Parsing and interrogation of parse trees will also require *Stanford CoreNLP*. *corpkit* can download and install it for you automatically.

Graphical interface

Much of *corpkit*'s command line functionality is also available in the *corpkit GUI*. After installation, it can be started from the command line with:

```
$ python -m corpkit.gui
```

If you're working on a project from within Python, you can open it graphically with:

```
>>> from corpkit import gui
>>> gui()
```

Alternatively, the GUI is available (alongside documentation) as a standalone OSX app [here](#).

Interpreter

corpkit also has its own interpreter, a bit like the [Corpus Workbench](#). You can open it with:

```
$ corpkit
# or, alternatively:
$ python -m corpkit.env
```

And then start working with natural language commands:

```
> set junglebook as corpus
> parse junglebook with outname as jb
> set jb as corpus
> search corpus for governor-lemma matching processes:verbal showing pos and lemma
> calculate result as percentage of self
> plot result as line chart with title as 'Example figure'
```

From the interpreter, you can enter ipython, jupyter notebook or gui to switch between interfaces, preserving the local namespace and data where possible.

Information about the syntax is available at the [Overview](#).

CREATING PROJECTS AND BUILDING CORPORA

Doing corpus linguistics involves building and interrogating corpora, and exploring interrogation results. `corpkit` helps with all of these things. This page will explain how to create a new project and build a corpus.

- *Creating a new project*
- *Adding a corpus*
- *Creating a Corpus object*
- *Pre-processing the data*
- *Manipulating a parsed corpus*
- *Counting key features*

1.1 Creating a new project

The simplest way to begin using `corpkit` is to import it and to create a new project. Projects are simply folders containing subfolders where corpora, saved results, images and dictionaries will be stored. The simplest way is to do it is to use the `new_project` command in *bash*, passing in the name you'd like for the project as the only argument:

```
$ new_project psyc
# move there:
$ cd psyc
# now, enter python and begin ...
```

Or, from Python:

```
>>> import corpkit
>>> corpkit.new_project('psyc')
### move there:
>>> import os
>>> os.chdir('psyc')
>>> os.listdir('.')

['data',
 'dictionaries',
 'exported',
 'images',
 'logs',
 'saved_concordances',
 'saved_interrogations']
```

1.2 Adding a corpus

Now that we have a project, we need to add some plain-text data to the *data* folder. At the very least, this is simply a text file. Better than this is a folder containing a number of text files. Best, however, is a folder containing subfolders, with each subfolder containing one or more text files. These subfolders represent subcorpora.

You can add your corpus to the *data* folder from the command line, or using Finder/Explorer if you prefer.

```
$ cp -R /Users/me/Documents/transcripts ./data
```

Or, in *Python*, using *shutil*:

```
>>> import shutil  
>>> shutil.copytree('/Users/me/Documents/transcripts', './data')
```

If you've been using *bash* so far, this is the moment when you'd enter *Python* and `import corpkit`.

1.3 Creating a Corpus object

Once we have a corpus of text files, we need to turn it into a *Corpus* object.

```
>>> from corpkit import Corpus  
### you can leave out the 'data' if it's in there  
>>> unparsed = Corpus('data/transcripts')  
>>> unparsed  
<corpkit.corpus.Corpora instance: transcripts; 13 subcorpora>
```

1.4 Pre-processing the data

A *Corpus* object can only be interrogated if tokenisation or parsing has been performed. For this, `corpkit.corpus.Corpora` objects have `tokenise()` and `parse()` methods. Tokenising is faster, simpler, and will work for more languages. As shown below, you can also elect to POS tag and lemmatise the data:

```
> corpus = unparsed.tokenise(postags=True, lemmatisation=True)  
# switch either to false to disable---but lemmatisation requires pos
```

Parsing relies on Stanford CoreNLP's parser, and therefore, you must have the parser and Java installed. `corpkit` will look around in your PATH for the parser, but you can also pass in its location manually with (e.g.) `corenlppath='users/you/corenlp'`. If it can't be found, you'll be asked if you want to download and install it automatically. Parsing has sensible defaults, and can be run with:

```
>>> corpus = unparsed.parse()
```

Note: Remember that parsing is a computationally intensive task, and can take a long time!

`corpkit` can also work with speaker IDs. If lines in your file contain capitalised alphanumeric names, followed by a colon (as per the example below), these IDs can be stripped out and turned into metadata features in the parsed dataset.

```
JOHN: Why did they change the signs above all the bins?  
SPEAKER23: I know why. But I'm not telling.
```

To use this option, use the `speaker_segmentation` keyword argument:

```
>>> corpus = unparsed.parse(speaker_segmentation=True)
```

Tokenising or parsing creates a corpus that is structurally identical to the original, but with annotations in *CONLL-U* formatted files in place of the original .txt files. When parsing, there are also methods for multiprocessing, memory allocation and so on:

<code>parse()</code> argument	Type	Purpose
<code>corenlpPath</code>	<code>str</code>	Path to CoreNLP
<code>operations</code>	<code>str</code>	List of annotations
<code>copula_head</code>	<code>bool</code>	Make copula head of dependency parse
<code>speaker_segmentation</code>	<code>bool</code>	Do speaker segmentation
<code>memory_mb</code>	<code>int</code>	Amount of memory to allocate
<code>multiprocess</code>	<code>int/bool</code>	Process in n parallel jobs
<code>outname</code>	<code>str</code>	Custom name for parsed corpus

You can run parsing operations from the command line:

```
$ parse mycorpus --multiprocess 4 --outname MyData
```

1.5 Manipulating a parsed corpus

Once you have a parsed corpus, you're ready to analyse it. `corpkit.corpus.Corporus` objects can be navigated in a number of ways. *CoreNLP XML* is used to navigate the internal structure of *CONLL-U* files within the corpus.

```
>>> corpus[:3]                                     # access first three subcorpora
>>> corpus.subcorpora.chapter1                   # access subcorpus called chapter1
>>> f = corpus[5][20]                            # access 21st file in 6th subcorpus
>>> f.document.sentences[0].parse_string        # get parse tree for first sentence
>>> f.document.sentences.tokens[0].word          # get first word
```

1.6 Counting key features

Before constructing your own queries, you may want to use some predefined attributes for counting key features in the corpus.

```
>>> corpus.features
```

Output:

S	Characters	Tokens	Words	Closed class	Open class	Clauses	Sentences	Unmod.
→declarative	Passives	Mental processes	Relational processes	Mod.	declarative			
→Interrogative	Verbal processes	Imperative	Open interrogative	Closed	interrogative			
01	4380658	1258606	1092113	643779	614827	277103	68267	
→35981	16842		11570		11082		3691	5012
→	2962	615		787			813	
02	3185042	922243	800046	471883	450360	209448	51575	
→26149	10324		8952		8407		3103	3407
→	2578	540		547			461	
03	3157277	917822	795517	471578	446244	209990	51860	
→26383	9711		9163		8590		3438	3392
→	2572	583		556			452	
04	3261922	948272	820193	486065	462207	216739	53995	
→27073	9697		9553		9037		3770	3702
→	2665	652		669			530	
05	3164919	921098	796430	473446	447652	210165	52227	
→26137	9543		8958		8663		3622	3523
→	2738	633		571			467	
06	3187420	928350	797652	480843	447507	209895	52171	
→25096	8917		9011		8820		3913	3637
→	2722	686		553			480	
07	3080956	900110	771319	466254	433856	202868	50071	
→24077	8618		8616		8547		3623	3343
→	2676	615		515			434	

08	3356241	972652	833135	502913	469739	218382	52637		
→25285		9921		9230		9562		3963	3497 ↴
→	2831		692		603		442		
09	2908221	840803	725108	434839	405964	191851	47050		
→21807		8354		8413		8720		3876	3147 ↴
→	2582		675		554		455		
10	2868652	815101	708918	421403	393698	185677	43474		
→20763		8640		8067		8947		4333	3181 ↴
→	2727		584		596		424		

This can take a while, as it counts a number of complex features. Once it's done, however, it saves automatically, so you don't need to do it again. There are also `postags`, `wordclasses` and `lexicon` attributes, which behave similarly:

```
>>> corpus.posttags  
>>> corpus.wordclasses  
>>> corpus.lexicon
```

These results can be useful when generating relative frequencies later on. Right now, however, you're probably interested in searching the corpus yourself, however. Hit *Next* to learn about that.

INTERROGATING CORPORA

Once you've built a corpus, you can search it for linguistic phenomena. This is done with the `interrogate()` method.

- *Introduction*
- *Search types*
- *Grammatical searching*
- *Excluding results*
- *What to show*
- *Working with trees*
- *Tree show values*
- *Working with dependencies*
- *Working with metadata*
- *Working with coreferences*
- *Multiprocessing*
- *N-grams*
- *Collocation*
- *Saving interrogations*
- *Exporting interrogations*
- *Other options*

2.1 Introduction

Interrogations can be performed on any `corpkit.corpus.Corpus` object, but also, on `corpkit.corpus.Subcorpus` objects, `corpkit.corpus.File` objects and `corpkit.corpus.Datalist` objects (slices of Corpus objects). You can search plaintext corpora, tokenised corpora or fully parsed corpora using the same method. We'll focus on parsed corpora in this guide.

```
>>> from corpkit import *
### words matching 'woman', 'women', 'man', 'men'
>>> query = {W: r'/(^wo)m.n/'}
### interrogate corpus
>>> corpus.interrogate(query)
### interrogate parts of corpus
>>> corpus[2:4].interrogate(query)
>>> corpus.files[:10].interrogate(query)
```

```
### if you have a subcorpus called 'abstract':
>>> corpus.subcorpora.abstract.interrogate(query)
```

Corpus interrogations will output a `corpkit.interrogation.Interrogation` object, which stores a DataFrame of results, a Series of totals, a dict of values used in the query, and, optionally, a set of concordance lines. Let's search for proper nouns in *The Great Gatsby* and see what we get:

```
>>> corp = Corpus('gatsby-parsed')
## turn on concordancing:
>>> propnoun = corp.interrogate({P: '^NNP'}, do_concordancing=True)
>>> propnoun.results

      gatsby  tom  daisy  mr.  wilson  jordan  new  baker  york  miss
chapter1     12    32     29     4       0      2    10     21      6    19
chapter2      1    30      6     8      26      0     6      0      6      0
chapter3     28      0      1     8       0      22      5      6      5      1
chapter4     38    10     15    25      1      9      5      8      4      7
chapter5     36      3     26     4       0      0      1      1      1      1
chapter6     37    21     19    11      0      1      4      0      3      4
chapter7     63    87     60     9      27     35      9      2      5      1
chapter8     21      3     19     1      19      1      0      1      0      0
chapter9     27      5      9    14      4      3      4      1      4      1

>>> propnoun.totals

chapter1    232
chapter2    252
chapter3    171
chapter4    428
chapter5    128
chapter6    219
chapter7    438
chapter8    139
chapter9    208
dtype: int64

>>> propnoun.query

{'case_sensitive': False,
 'corpus': 'gatsby-parsed',
 'dep_type': 'collapsed-ccprocessed-dependencies',
 'do_concordancing': True,
 'exclude': False,
 'excludemode': 'any',
 'files_as_subcorpora': True,
 'gramsize': 1,
 ...}

>>> propnoun.concordance # (sample)

54 chapter1          They had spent a year in france      for no particular reason and_
  ↪then d
55 chapter1  n't believe it I had no sight into daisy      's heart but i felt that tom_
  ↪would
56 chapter1  into Daisy 's heart but I felt that tom      would drift on forever seeking_
  ↪a li
57 chapter1          This was a permanent move said daisy      over the telephone but i did n
  ↪'t be
58 chapter1  windy evening I drove over to East egg      to see two old friends whom i_
  ↪scarc
59 chapter1  warm windy evening I drove over to east      egg to see two old friends whom_
  ↪i s
60 chapter1  d a cheerful red and white Georgian colonial
61 chapter1  pen to the warm windy afternoon and tom      mansion overlooking the bay
  ↪stan
62 chapter1  to the warm windy afternoon and Tom buchanan  buchanan in riding clothes was_
  ↪with
```

Cool, eh? We'll focus on what to do with these attributes later. Right now, we need to learn how to generate them.

2.2 Search types

Parsed corpora contain many different kinds of things we might like to search. There are word forms, lemma forms, POS tags, word classes, indices, and constituency and (three different) dependency grammar annotations. For this reason, the search query is a dict object passed to the `interrogate()` method, whose keys specify what to search, and whose values specify a query. The simplest ones are given in the table below.

Note: Single capital letter variables in code examples represent lowercase strings (`W = 'w'`). These variables are made available by doing `from corpkit import *`. They are used here for readability.

Search	Gloss
W	Word
L	Lemma
F	Function
P	POS tag
X	Word class
E	NER tag
A	Distance from root
I	Index in sentence
S	Sentence index
R	Coref representative

Because it comes first, and because it's always needed, you can pass it in like an argument, rather than a keyword argument.

```
### get variants of the verb 'be'
>>> corpus.interrogate({L: 'be'})
### get words in 'nsubj' position
>>> corpus.interrogate({F: 'nsubj'})
```

Multiple key/value pairs can be supplied. By default, all must match for the result to be counted, though this can be changed with `searchmode=ANY` or `searchmode=ALL`:

```
>>> goverb = {P: r'^v', L: r'^go'}
### get all variants of 'go' as verb
>>> corpus.interrogate(goverb, searchmode=ALL)
### get all verbs and any word starting with 'go':
>>> corpus.interrogate(govert, searchmode=ANY)
```

2.3 Grammatical searching

In the examples above, we match attributes of tokens. The great thing about parsed data, is that we can search for relationships between words. So, other possible search keys are:

Search	Gloss
G	Governor
D	Dependent
H	Coreference head
T	Syntax tree
A1	Token 1 place to left
Z1	Token 1 place to right

```
>>> q = {G: r'^b'}
### return any token with governor word starting with 'b'
>>> corpus.interrogate(q)
```

Governor, *Dependent* and *Left/Right* can be combined with the earlier table, allowing a large array of search types:

	Match	Governor	Dependent	Coref head	Left/right
Word	W	G	D	H	A1/Z1
Lemma	L	GL	DL	HL	A1L/Z1L
Function	F	GF	DF	HF	A1F/Z1F
POS tag	P	GP	DP	HP	A1P/Z1P
Word class	X	GX	DX	HX	A1X/Z1X
Distance from root	A	GA	DA	HA	A1A/Z1A
Index	I	GI	DI	HI	A1I/Z1I
Sentence index	S	GS	DS	HS	A1S/Z1S

Syntax tree searching can't be combined with other options. We'll return to them in a minute, however.

2.4 Excluding results

You may also wish to exclude particular phenomena from the results. The `exclude` argument takes a dict in the same form a search. By default, if any key/value pair in the `exclude` argument matches, it will be excluded. This is controlled by `excludemode=ANY` or `excludemode=ALL`.

```
>>> from corpkit.dictionaries import wordlists
### get any noun, but exclude closed class words
>>> corpus.interrogate({P: r'^n'}, exclude={W: wordlists.closedclass})
### when there's only one search criterion, you can also write:
>>> corpus.interrogate(P, r'^n', exclude={W: wordlists.closedclass})
```

In many cases, rather than using `exclude`, you could also remove results later, during editing.

2.5 What to show

Up till now, all searches have simply returned words. The final major argument of the `interrogate` method is `show`, which dictates what is returned from a search. Words are the default value. You can use any of the search values as a `show` value. `show` can be either a single string or a list of strings. If a list is provided, each value is returned with forward slashes as delimiters.

```
>>> example = corpus.interrogate({W: r'fr?iends?'}, show=[W, L, P])
>>> list(example.results)
['friend/friend/nn', 'friends/friend/nns', 'fiend/fiend/nn', 'fiends/fiend/nns', ... ]
```

Unigrams are generated by default. To get n-grams, pass in an `n` value as `gramsize`:

```
>>> example = corpus.interrogate({W: r'wom[ae]n'}, show=N, gramsize=2)
>>> list(example.results)
['a/woman', 'the/woman', 'the/women', 'women/are', ... ]
```

So, this leaves us with a huge array of possible things to show, all of which can be combined if need be:

	Match	Governor	Dependent	Coref Head	1L position	1R position
Word	W	G	D	H	A1	Z1
Lemma	L	GL	DL	HL	A1L	Z1L
Function	F	GF	DF	HF	A1F	Z1F
POS tag	P	GP	DP	HP	A1P	Z1P
Word class	X	GX	DX	HX	A1X	Z1X
Distance from root	A	GA	DA	HA	A1A	Z1R
Index	I	GI	DI	HI	A1I	Z1I
Sentence index	S	GS	DS	HS	A1S	Z1S

One further extra `show` value is '`c`' (count), which simply counts occurrences of a phenomenon. Rather than returning a DataFrame of results, it will result in a single Series. It cannot be combined with other values.

2.6 Working with trees

If you have elected to search trees, by default, searching will be done with Java, using Tregex. If you don't have Java, or if you pass in `tgrep=True`, searching will the more limited Tgrep2 syntax. Here, we'll concentrate on Tregex.

Tregex is a language for searching syntax trees like this one:

To write a Tregex query, you specify *words and/or tags* you want to match, in combination with *operators* that link them together. First, let's understand the Tregex syntax.

To match any adjective, you can simply write:

```
JJ
```

with `JJ` representing adjective as per the [Penn Treebank tagset](#). If you want to get NPs containing adjectives, you might use:

```
NP < JJ
```

where `<` means *with a child/immediately below*. These operators can be reversed: If we wanted to show the adjectives within NPs only, we could use:

```
JJ > NP
```

It's good to remember that **the output will always be the left-most part of your query**.

If you only want to match Subject NPs, you can use bracketting, and the `$` operator, which means *sister/directly to the left/right of*:

```
JJ > (NP $ VP)
```

In this way, you build more complex queries, which can extent all the way from a sentence's *root* to particular tokens. The query below, for example, finds adjectives modifying *book*:

```
JJ > (NP <<# /book/)
```

Notice that here, we have a different kind of operator. The `<<` operator means that the node on the right does not need to be a child, but can be a descendant. the `#` means *head*—that is, in SFL, it matches the *Thing* in a Nominal Group.

If we wanted to also match *magazine* or *newspaper*, there are a few different approaches. One way would be to use `|` as an operator meaning *or*:

```
JJ > (NP ( <<# /book/ | <<# /magazine/ | <<# /newspaper/ ))
```

This can be cumbersome, however. Instead, we could use a regular expression:

```
JJ > (NP <<# /^ (book|newspaper|magnitude) s*$ /)
```

Though it is beyond the scope of this guide to teach Regular Expressions, it is important to note that Regular Expressions are extremely powerful ways of searching text, and are invaluable for any linguist interested in digital datasets.

Detailed documentation for Tregex usage (with more complex queries and operators) can be found [here](#).

2.7 Tree *show* values

Though you can use the same Tregex query for tree searches, the output changes depending on what you select as the `show` value. For the following sentence:

```
These are prosperous times.
```

you could write a query:

```
r'JJ < __'
```

Which would return:

Show	Gloss	Output
W	Word	<i>prosperous</i>
T	Tree	(JJ <i>prosperous</i>)
p	POS tag	<i>JJ</i>
C	Count	<i>I</i> (added to total)

2.8 Working with dependencies

When working with dependencies, you can use any of the long list of search and *show* values. It's possible to construct very elaborate queries:

```
>>> from corpkit.dictionaries import process_types, roles
### nominal nsubj with verbal process as governor
>>> crit = {F: r'^nsubj$', ...
...     GL: processes.verbal.lemmata,
...     GF: roles.event,
...     P: r'^N'}
### interrogate corpus, outputting the nsubj lemma
>>> sayers = parsed.interrogate(crit, show=L)
```

2.9 Working with metadata

If you've used speaker segmentation and/or metadata addition when building your corpus, you can tell the *interrogate()* method to use these values as subcorpora, or restrict searches to particular values. The code below will limit searches to sentences spoken by Jason and Martin, or exclude them from the search:

```
>>> corpus.interrogate(query, just_metadata={'speaker': ['JASON', 'MARTIN']})
>>> corpus.interrogate(query, skip_metadata={'speaker': ['JASON', 'MARTIN']})
```

If you wanted to compare Jason and Martin's contributions in the corpus as a whole, you could treat them as subcorpora:

```
>>> corpus.interrogate(query, subcorpora='speaker',
...                      just_metadata={'speaker': ['JASON', 'MARTIN']})
```

The method above, however, will make an interrogation with two subcorpora, 'JASON' AND MARTIN. You can pass a list in as the *subcorpora* keyword argument to generate a multiindex:

```
>>> corpus.interrogate(query, subcorpora=['folder', 'speaker'],
...                      just_metadata={'speaker': ['JASON', 'MARTIN']})
```

2.10 Working with coreferences

One major challenge in corpus linguistics is the fact that pronouns stand in for other words. Parsing provides coreference resolution, which maps pronouns to the things they denote. You can enable this kind of parsing by specifying the *dcoref* annotator:

```
>>> corpus = Corpus('example.txt')
>>> ops = 'tokenize,ssplit,pos,lemma,parse,ner,dcoref'
>>> parsed = corpus.interrogate(operations=ops)
## print a plaintext representation of the parsed corpus
>>> print(parsed.plain)
```

```
0. Clinton supported the independence of Kosovo
1. He authorized the use of force.
```

If you have done this, you can use `coref=True` while interrogating to allow coreferent forms to be counted alongside query matches. For example, if you wanted to find all the processes Clinton is engaged in, you could do:

```
>>> from corpkit.dictionaries import roles
>>> query = {W: 'clinton', GF: roles.process}
>>> res = parsed.interrogate(query, show=L, coref=True)
>>> res.results.columns
```

This matches both *Clinton* and *he*, and thus gives us:

```
['support', 'authorize']
```

2.11 Multiprocessing

Interrogating the corpus can be slow. To speed it up, you can pass an integer as the `multiprocess` keyword argument, which tells the `interrogate()` method how many processes to create.

```
>>> corpus.interrogate({T: r'__ > MD'}, multiprocess=4)
```

Note: Too many parallel processes may slow your computer down. If you pass in `multiprocessing=True`, the number of processes will equal the number of cores on your machine. This is usually a fairly sensible number.

2.12 N-grams

N-gramming can be generated by making `gramsize > 1`:

```
>>> corpus.interrogate({W: 'father'}, show='L', gramsize=3)
```

2.13 Collocation

Collocations can be shown by making using `window`:

```
>>> corpus.interrogate({W: 'father'}, show='L', window=6)
```

2.14 Saving interrogations

```
>>> interro.save('savename')
```

Interrogation savenames will be prefaced with the name of the corpus interrogated.

You can also quicksave interrogations:

```
>>> corpus.interrogate(T, r'/NN.?.+', save='savename')
```

2.15 Exporting interrogations

If you want to quickly export a result to CSV, LaTeX, etc., you can use Pandas' DataFrame methods:

```
>>> print(nouns.results.to_csv())
>>> print(nouns.results.to_latex())
```

2.16 Other options

`interrogate()` takes a number of other arguments, each of which is documented in the API documentation.

If you're done interrogating, you can head to the page on [Editing results](#) to learn how to transform raw frequency counts into something more meaningful. Or, hit *Next* to learn about concordancing.

CHAPTER THREE

CONCORDANCING

Concordancing is the task of getting an aligned list of *keywords in context*. Here's a very basic example, using *Industrial Society and Its Future* as a corpus:

```
>>> tech = corpus.concordance({W: r'techn*'})  
>>> tech.format(n=10, columns=[L, M, R])  
  
0    The continued development of technology      will worsen the situation  
1  vernments but the economic and technological basis of the present society  
2    They want to make him study technical subjects become an executive o  
3  program to acquire some petty technical skill then come to work on tim  
4  rom nature are consequences of technological progress  
5  n them and modern agricultural technology has made it possible for the e  
6          -LRB- Also technology exacerbates the effects of cro  
7  changes very rapidly owing to technological change  
8  they enthusiastically support technological progress and economic growth  
9  e rapid drastic changes in the technology and the economy of a society w
```

3.1 Generating a concordance

When using *corpkit*, any interrogation is also optionally a concordance. If you use the `do_concordancing` keyword argument, your interrogation will have a `concordance` attribute containing concordance lines. Like interrogation results, concordances are stored as *Pandas DataFrames*. `maxconc` controls the number of lines produced.

```
>>> withconc = corp.interrogate({L: ['man', 'woman', 'person']},
...                               show=[W,P],
...                               do_concordancing=True,
...                               maxconc=500)  
  
0  T Asian/JJ a/DT disabled/JJ person/nn   or/cc a/dt woman/nn origin
1  led/JJ person/NN or/CC a/DT woman/nn     originally/rb had/vbd no/d
2  woman/NN or/CC disabled/JJ person/nn     but/cc a/dt minority/nn of
3  n/JJ immigrant/JJ abused/JJ woman/nn     or/cc disabled/jj person/n
4  ing/VBG weak/JJ -LRB--/LRB- women/nns  -rrb--/rrb- defeated/vbn -
```

If you like, you can use `only_format_match=True` to keep the left and right context simple:

```
>>> withconc = corp.interrogate({L: ['man', 'woman', 'person']},
...                               show=[W,P],
...                               only_format_match=True,
...                               do_concordancing=True,
...                               maxconc=500)  
  
0  African an Asian a disabled person/nn   or a woman originally had
1  sian a disabled person or a woman/nn     originally had no derogato
2  nt abused woman or disabled person/nn   but a minority of activist
3  ller Asian immigrant abused woman/nn   or disabled person but a m
4  n image of being weak -LRB- women/nns  -rrb- defeated -lrb- ameri
```

If you don't want or need the interrogation data, you can use the `concordance()` method:

```
>>> conc = corpus.concordance(T, r'JJ.*? > (NP <<# /man/)' )
```

3.2 Displaying concordance lines

How concordance lines will be displayed really depends on your interpreter and environment. For the most part, though, you'll want to use the `format()` method.

```
>>> lines.format(kind='s',
...                 n=100,
...                 window=50,
...                 columns=[L, M, R])
```

`kind='c'/'l'/'s'` allows you to print as CSV, LaTeX, or simple string. `n` controls the number of results shown. `window` controls how much context to show in the left and right columns. `columns` accepts a list of column names to show.

Pandas' `set_option` can be used to customise some visualisation defaults.

3.3 Working with concordance lines

You can edit concordance lines using the `edit()` method. You can use this method to keep or remove entries or subcorpora matching regular expressions or lists. Keep in mind that because concordance lines are DataFrames, you can use Pandas' dedicated methods for working with text data.

```
### get just uk variants of words with variant spellings
>>> from corpkit.dictionaries import usa_convert
>>> concs = result.concordance.edit(just_entries=usa_convert.keys())
```

Concordance objects can be saved just like any other `corpkit` object:

```
>>> concs.save('adj_modifying_man')
```

You can also easily turn them into CSV data, or into LaTeX:

```
### pandas methods
>>> concs.to_csv()
>>> concs.to_latex()

### corpkit method: csv and latex
>>> concs.format('c', window=20, n=10)
>>> concs.format('l', window=20, n=10)
```

3.4 The `calculate` method

You might have begun to notice that interrogating and concordancing aren't really very different tasks. If we drop the left and right context, and move the data around, we have all the data we get from an interrogation.

For this reason, you can use the `calculate()` method to generate an `corpus.interrogation`. `Interrogation` object containing a frequency count of the middle column of the concordance as the `results` attribute.

Therefore, one method for ensuring accuracy is to:

1. Run an interrogation, using `do_concordance=True`
2. Remove false positives from the concordance result using `edit()`
3. Use the `calculate()` method to regenerate the overall frequencies

4. Edit, visualise or export the data

If you'd like to randomise the order of your results, you can use `lines.shuffle()`

EDITING RESULTS

Corpus interrogation is the task of getting frequency counts for a lexicogrammatical phenomenon in a corpus. Simple absolute frequencies, however, are of limited use. The `edit()` method allows us to do complex things with our results, including:

- *Keeping or deleting results and subcorpora*
- *Editing result names*
- *Spelling normalisation*
- *Generating relative frequencies*
- *Keywording*
- *Sorting*
- *Calculating trends, P values*
- *Saving results*
- *Exporting results*
- *Next step*

Each of these will be covered in the sections below. Keep in mind that because results are stored as DataFrames, you can also use Pandas/Numpy/Scipy to manipulate your data in ways not covered here.

4.1 Keeping or deleting results and subcorpora

One of the simplest kinds of editing is removing or keeping results or subcorpora. This is done using keyword arguments: `skip_subcorpora`, `just_subcorpora`, `skip_entries`, `just_entries`. The value for each can be:

1. A string (treated as a regular expression to match)
2. A list (a list of words to match)
3. An integer (treated as an index to match)

```
>>> criteria = r'ing$'  
>>> result.edit(just_entries=criteria)
```

```
>>> criteria = ['everything', 'nothing', 'anything']  
>>> result.edit(skip_entries=criteria)
```

```
>>> result.edit(just_subcorpora=['Chapter_10', 'Chapter_11'])
```

You can also span subcorpora, using a tuple of `(first_subcorpus, second_subcorpus)`. This works for numerical and non-numerical subcorpus names:

```
>>> just_span = result.edit(span_subcorpora=(3, 10))
```

4.2 Editing result names

You can use the `replace_names` keyword argument to edit the text of each result. If you pass in a string, it is treated as a regular expression to delete from every result:

```
>>> ingdel = result.edit(replace_names=r'ing$')
```

You can also pass in a dict with the structure of `{newname: criteria}`:

```
>>> rep = {'-ing words': r'ing$', '-ed words': r'ed$'}
>>> replaced = result.edit(replace_names=rep)
```

If you wanted to see how commonly words start with a particular letter, you could do something creative:

```
>>> from string import lowercase
>>> crit = {k.upper() + ' words': r'(?i)^%s.*' % k for k in lowercase}
>>> firstletter = result.edit(replace_names=crit, sort_by='total')
```

4.3 Spelling normalisation

When results are single words, you can normalise to UK/US spelling:

```
>>> spelled = result.edit(spelling='UK')
```

You can also perform this step when interrogating a corpus.

4.4 Generating relative frequencies

Because subcorpora often vary in size, it is very common to want to create relative frequency versions of results. The best way to do this is to pass in an `operation` and a `denominator`. The `operation` is simply a string denoting a mathematical operation: '+', '-', '*', '/', '%'. The last two of these can be used to get relative frequencies and percentage.

Denominator is what the result will be divided by. Quite often, you can use the string '`'self'`'. This means, after all other editing (deleting entries, subcorpora, etc.), use the totals of the result being edited as the denominator. When doing no other editing operations, the two lines below are equivalent:

```
>>> rel = result.edit('%', 'self')
>>> rel = result.edit('%', result.totals)
```

The best denominator, however, may not simply be the totals for the results being edited. You may instead want to relativise by the total number of words:

```
>>> rel = result.edit('%', corpus.features.Words)
```

Or by some other result you have generated:

```
>>> words_with_oo = corpus.interrogate(W, 'oo')
>>> rel = result.edit('%', words_with_oo.totals)
```

There is a more complex kind of relative frequency making, where a `.results` attribute is used as the denominator. In the example below, we calculate the percentage of the time each verb occurs as the *root* of the parse.

```
>>> verbs = corpus.interrogate(P, r'^vb', show=L)
>>> roots = corpus.interrogate(F, 'root', show=L)
>>> relv = verbs.edit('%', roots.results)
```

4.5 Keywording

corpkit treats keywording as an editing task, rather than an interrogation task. This makes it easy to get key nouns, or key Agents, or key grammatical features. To do keywording, use the K operation:

```
>>> from corpkit import *
### * imports predefined global variables like K and SELF
>>> keywords = result.edit(K, SELF)
```

This finds out which words are key in each subcorpus, compared to the corpus as a whole. You can compare subcorpora directly as well. Below, we compare the plays subcorpus to the novels subcorpus.

. code-block:: python

```
>>> from corpkit import *
>>> keywords = result.edit(K, result.ix['novels'], just_subcorpora='plays')
```

You could also pass in word frequency counts from some other source. A wordlist of the *British National Corpus* is included:

```
>>> keywords = result.edit(K, 'bnc')
```

The default keywording metric is *log-likelihood*. If you'd like to use *percentage difference*, you can do:

```
>>> keywords = result.edit(K, 'bnc', keyword_measure='pd')
```

4.6 Sorting

You can sort results using the `sort_by` keyword. Possible values are:

- ‘name’ (alphabetical)
- ‘total’ (most common first)
- ‘infreq’ (inverse total)
- ‘increase’ (most increasing)
- ‘decrease’ (most decreasing)
- ‘turbulent’ (by most change)
- ‘static’ (by least change)
- ‘p’ (by p value)
- ‘slope’ (by slope)
- ‘intercept’ (by intercept)
- ‘r’ (by correlation coefficient)
- ‘stderr’ (by standard error of the estimate)
- ‘<subcorpus>’ by total in <subcorpus>

```
>>> inc = result.edit(sort_by='increase', keep_stats=False)
```

Many of these rely on Scipy’s linregress function. If you want to keep the generated statistics, use `keep_stats=True`.

4.7 Calculating trends, P values

`keep_stats=True` will cause slopes, p values and stderr to be calculated for each result.

4.8 Saving results

You can save edited results to disk.

```
>>> edited.save('savename')
```

4.9 Exporting results

You can generate CSV data very easily using Pandas:

```
>>> result.results.to_csv()
```

4.10 Next step

Once you've edited data, it's ready to visualise. Hit next to learn how to use the `visualise()` method.

VISUALISING RESULTS

One thing missing in a lot of corpus linguistic tools is the ability to produce high-quality visualisations of corpus data. `corpkit` uses the `corpkit.interrogation.Interrogation.visualise` method to do this.

- *Basics*
- *Plot type*
- *Plot style*
- *Figure and font size*
- *Title and labels*
- *Subplots*
- *TeX*
- *Legend*
- *Colours*
- *Saving figures*
- *Other options*
- *Multiplotting*

Note: Most of the keyword arguments from Pandas' `plot` method are available. See their documentation for more information.

5.1 Basics

`visualise()` is a method of all `corpkit.interrogation.Interrogation` objects. If you use `from corpkit import *`, it is also monkey-patched to Pandas objects.

Note: If you're using a *Jupyter Notebook*, make sure you use `%matplotlib inline` or `%matplotlib notebook` to set the appropriate backend.

A common workflow is to interrogate a corpus, relative results, and visualise:

```
>>> from corpkit import *
>>> corpus = Corpus('data/P-parsed', load_saved=True)
>>> counts = corpus.interrogate({T: r'MD < __'})
>>> reldat = counts.edit('%', SELF)
>>> reldat.visualise('Modals', kind='line', num_to_plot=ALL).show()
```

```
### the visualise method can also attach to the df:
>>> reldat.results.visualise(...).show()
```

The current behaviour of `visualise()` is to return the `pyplot` module. This allows you to edit figures further before showing them. Therefore, there are two ways to show the figure:

```
>>> data.visualise().show()
```

```
>>> plt = data.visualise()
>>> plt.show()
```

5.2 Plot type

The `visualise` method allows `line`, `bar`, horizontal bar (`barh`), `area`, and `pie` charts. Those with `seaborn` can also use '`heatmap`' ([docs](#)). Just pass in the type as a string with the `kind` keyword argument. Arguments such as `robust=True` can then be used.

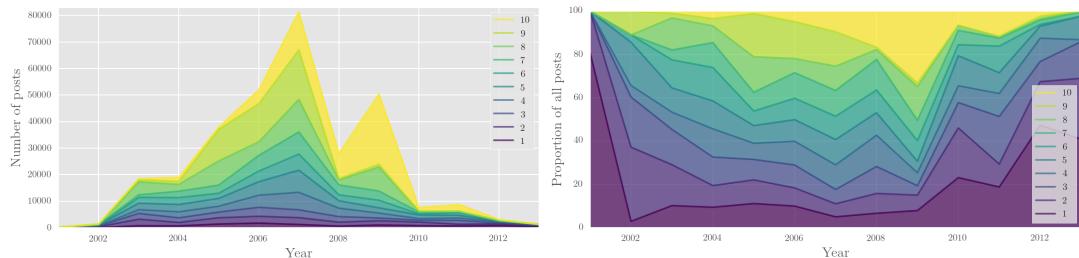
```
>>> data.visualise(kind='heatmap', robust=True, figsize=(4, 12),
...                  x_label='Subcorpus', y_label='Event').show()
```



Fig. 5.1: Heatmap example

Stacked area/line plots can be made with `stacked=True`. You can also use `filled=True` to at-

tempt to make all values sum to 100. Cumulative plotting can be done with `cumulative=True`. Below is an area plot beside an area plot where `filled=True`. Both use the `vidiris` colour scheme.



5.3 Plot style

You can select from a number of styles, such as `ggplot`, `fivethirtyeight`, `bmh`, and `classic`. If you have `seaborn` installed (and you should), then you can also select from `seaborn` styles (`seaborn-paper`, `seaborn-dark`, etc.).

5.4 Figure and font size

You can pass in a tuple of `(width, height)` to control the size of the figure. You can also pass an integer as `fontsize`.

5.5 Title and labels

You can label your plot with `title`, `x_label` and `y_label`:

```
>>> data.visualise('Modals', x_label='Subcorpus', y_label='Relative frequency')
```

5.6 Subplots

`subplots=True` makes a separate plot for every entry in the data. If using it, you'll probably also want to use `layout=(rows, columns)` to specify how you'd like the plots arranged.

```
>>> data.visualise(subplots=True, layout=(2, 3)).show()
```

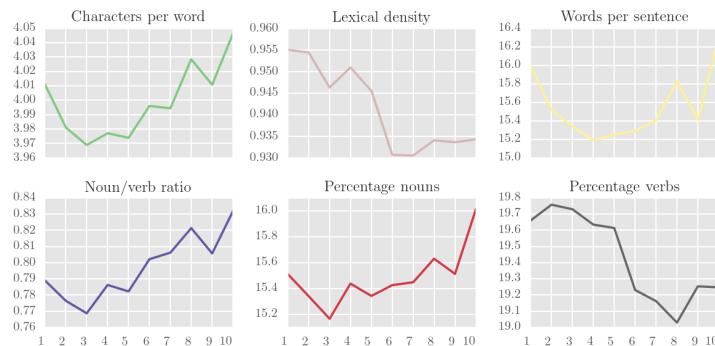


Fig. 5.2: Line charts using subplots and layout specification

5.7 TeX

If you have LaTeX installed, you can use `tex=True` to render text with LaTeX. By default, `visualise()` tries to use LaTeX if it can.

5.8 Legend

You can turn the legend off with `legend=False`. Legend placement can be controlled with `legend_pos`, which can be:

Margin	Figure		Margin
outside upper left	upper left	upper right	outside upper right
outside center left	center left	center right	outside center right
outside lower left	lower left	lower right	outside lower right

The default value, '`best`', tries to find the best place automatically (without leaving the figure boundaries).

If you pass in `draggable=True`, you should be able to drag the legend around the figure.

5.9 Colours

You can use the `colours` keyword argument to pass in:

1. A colour name recognised by `matplotlib`
2. A hex colour string
3. A colourmap object

There is an extra argument, `black_and_white`, which can be set to `True` to make greyscale plots. Unlike `colours`, it also updates line styles.

5.10 Saving figures

To save a figure to a project's `images` directory, you can use the `save` argument. `output_format='png' / 'pdf'` can be used to change the file format.

```
>>> data.visualise(save='name', output_format='png')
```

5.11 Other options

There are a number of further keyword arguments for customising figures:

Argument	Type	Action
<i>grid</i>	<i>bool</i>	Show grid in background
<i>rot</i>	<i>int</i>	Rotate x axis labels n degrees
<i>shadow</i>	<i>bool</i>	Shadows for some parts of plot
<i>ncol</i>	<i>int</i>	n columns for legend entries
<i>explode</i>	<i>list</i>	Explode these entries in pie
<i>partial_pie</i>	<i>bool</i>	Allow plotting of pie slices
<i>legend_frame</i>	<i>bool</i>	Show frame around legend
<i>legend_alpha</i>	<i>float</i>	Opacity of legend
<i>reverse_legend</i>	<i>bool</i>	Reverse legend entry order
<i>transpose</i>	<i>bool</i>	Flip axes of DataFrame
<i>logx/logy</i>	<i>bool</i>	Log scales
<i>show_p_val</i>	<i>bool</i>	Try to show p value in legend
<i>interactive</i>	<i>bool</i>	Experimental <code>mpld3</code> use

A number of these and other options for customising figures are also described in the `corpkit.interrogation.Interrogation.visualise` method documentation.

5.12 Multiplotting

The `corpkit.interrogation.Interrogation` also comes with a `corpkit.interrogation.Interrogation.multiplot` method, which can be used to show two different kinds of chart within the same figure.

The first two arguments for the function are two *dict* objects, which configure the larger and smaller plots.

For the second dictionary, you may pass in a *data* argument, which is an `corpkit.interrogation.Interrogation` or similar, and will be used as separate data for the subplots. This is useful, for example, if you want your main plot to show absolute frequencies, and your subplots to show relative frequencies.

There is also *layout*, which you can use to choose an overall grid design. You can also pass in a list of tuples if you like, to use your own layout. Below is a complete example, focussing on objects in risk processes:

```
>>> from corpkit import *
>>> from corpkit.dictionaries import *
### parse a collection of text files
>>> corpora = Corus('data/news')
### make dependency parse query: get get 'object' of risk process
>>> query = {F: roles.participant2, GL: r'\brisk', GF: roles.process}
### interrogate corpus, return lemma form, no coreference
>>> result = corpus.interrogate(query, show=[L], coref=False)
### generate relative frequencies, skip closed class, and sort
>>> inc = result.edit('%', SELF,
                     sort_by='increase',
                     skip_entries=wordlists.closedclass)
### visualise as area and line charts combined
>>> inc.multiplot({'title': 'Objects of risk processes, increasing',
                  'kind': 'area',
                  'x_label': 'Year',
                  'y_label': 'Percentage of all results'},
                  {'kind': 'line'}, layout=5)
```

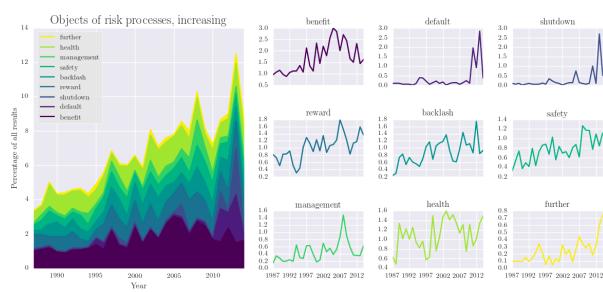


Fig. 5.3: *multiplot* example

USING LANGUAGE MODELS

Warning: Language modelling is currently deprecated, while the tool is updated to use *CONLL* formatted data, rather than *CoreNLP XML*. Sorry!

Language models are probability distributions over sequences of words. They are common in a number of natural language processing tasks. In corpus linguistics, they can be used to judge the similarity between texts.

corpkit's `make_language_model()` method makes it very easy to generate a language model:

```
>>> corpus = Corpus('threads')
# save as models/savename.p
>>> lm = corpus.make_language_model('savename')
```

One simple thing you can do with a language model is pass in a string of text:

```
>>> text = ("We can compare an arbitrary string against the models \"\
...         \"created for each subcorpus, in order to find out how \"\
...         \"similar the text is to the texts in each subcorpus... \"")
# get scores for each subcorpus, and the corpus as a whole
>>> lm.score(text)

01      -4.894732
04      -4.986471
02      -5.060964
03      -5.096785
05      -5.106083
07      -5.226934
06      -5.338614
08      -5.829444
09      -5.874777
10      -6.351399
Corpus   -5.285553
```

You can also pass in `corpkit.corpus.Subcorpus` objects, subcorpus names or `corpkit.corpus.File` instances.

6.1 Customising models

Under the hood, *corpkit* interrogates the corpus using some special parameters, then builds a model from the results. This means that you can pass in arbitrary arguments for the `interrogate()` method:

```
>>> lm = corpus.make_language_model('lemma_model',
...                                 show=L,
...                                 just_speakers='MAHSA',
...                                 multiprocess=2)
```

6.2 Compare subcorpora

You can find out which subcorpora are similar using the `score()` method:

```
>>> lm.score('1996')
```

Or get a complete *DataFrame* of values using `score_subcorpora()`:

```
>>> df = lm.score_subcorpora()
```

6.3 Advanced stuff

Note: Coming soon

MANAGING PROJECTS

corpkit has a few other bits and pieces designed to make life easier when doing corpus linguistic work. This includes methods for loading saved data, for working with multiple corpora at the same time, and for switching between command line and graphical interfaces. Those things are covered here.

- *Loading saved data*
- *Managing multiple corpora*
- *Using the GUI*

7.1 Loading saved data

When you're starting a new session, you probably don't want to start totally from scratch. It's handy to be able to load your previous work. You can load data in a few ways.

First, you can use `corpkit.load()`, using the name of the filename you'd like to load. By default, `corpkit` looks in the `saved_interrogations` directory, but you can pass in an absolute path instead if you like.

```
>>> import corpkit
>>> nouns = corpkit.load('nouns')
```

Second, you can use `corpkit.loader()`, which provides a list of items to load, and asks the user for input:

```
>>> nouns = corpkit.loader()
```

Third, when instantiating a `Corpus` object, you can add `load_saved=True` keyword argument to load any saved data belonging to this corpus as an attribute.

```
>>> corpus = Corpus('data/psyc-parsed', load_saved=True)
```

A final alternative approach stores all interrogations within an `corpkit.interrogation.Interrodict` object object:

```
>>> r = corpkit.load_all_results()
```

7.2 Managing multiple corpora

corpkit can handle one further level of abstraction for both Corpus and Interrogations. `corpkit.corpus.Corpora` models a collection of `corpkit.corpus.Corpus` objects. To create one, pass in a directory containing corpora, or a list of paths/Corpus objects:

```
>>> from corpkit import Corpora
>>> corpora = Corpora('data')
```

Individual corpora can be accessed as attributes, by index, or as keys:

```
>>> corpora.first  
>>> corpora[0]  
>>> corpora['first']
```

You can use the `interrogate()` method to search them, using the same arguments as you would for `interrogate()`.

Interrogating these objects often returns an `corpkit.interrogation.Interrodict` object, which models a collection of DataFrames.

Editing can be performed with `edit()`. The editor will iterate over each DataFrame in turn, generally returning another Interrodict.

Note: There is no `visualise()` method for Interrodict objects.

`multiindex()` can turn an Interrodict into a *Pandas MultiIndex*:

```
>>> multiple_res.multiindex()
```

`collapse()` will collapse one dimension of the Interrodict. You can collapse the x axis ('x'), the y axis ('y'), or the Interrodict keys ('k'). In the example below, an Interrodict is collapsed along each axis in turn.

```
>>> d = corpora.interrogate({F: 'compound', GL: r'^risk'}, show=L)  
>>> d.keys()  
  
['CHT', 'WAP', 'WSJ']  
  
>>> d['CHT'].results  
  
....  health  cancer  security  credit  flight  safety  heart  
1987      87       25       28      13       7       6       4  
1988      72       24       20      15       7       4       9  
1989     137       61       23      10       5       5       6  
  
>>> d.collapse(axis=Y).results  
  
...  health  cancer  credit  security  
CHT    3174    1156    566    697  
WAP    2799     933    582   1127  
WSJ    1812     680   2009    537  
  
>>> d.collapse(axis=X).results  
  
...  1987  1988  1989  
CHT    384    328    464  
WAP    389    355    435  
WSJ    428    410    473  
  
>>> d.collapse(axis=K).results  
  
...  health  cancer  credit  security  
1987     282     127      65      93  
1988     277     100      70     107  
1989     379     253      83      91
```

`topwords()` quickly shows the top results from every interrogation in the Interrodict.

```
>>> data.topwords(n=5)
```

Output:

TBT	%	UST	%	WAP	%	WSJ	%
health	25.70	health	15.25	health	19.64	credit	9.22
security	6.48	cancer	10.85	security	7.91	health	8.31

cancer	6.19	heart	6.31	cancer	6.55	downside	5.46
flight	4.45	breast	4.29	credit	4.08	inflation	3.37
safety	3.49	security	3.94	safety	3.26	cancer	3.12

7.3 Using the GUI

corpkit is also designed to work as a GUI. It can be started in bash with:

```
$ python -m corpkit.gui
```

The GUI can understand any projects you have defined. If you open it, you can simply select your project via Open Project and resume work in a graphical environment.

CHAPTER EIGHT

OVERVIEW

corpuskit comes with a dedicated interpreter, which receives commands in a natural language syntax like these:

```
> set mydata as corpus
> search corpus for pos matching 'JJ.*'
> call result 'adjectives'
> edit adjectives by skipping subcorpora matching 'books'
> plot edited as line chart with title as 'Adjectives'
```

It's a little less powerful than the full Python API, but it is easier to use, especially if you don't know Python. You can also switch instantly from the interpreter to the full API, so you only need the API for the really tricky stuff.

The syntax of the interpreter is based around *objects*, which you do things to, and *commands*, which are actions performed upon the objects. The example below uses the *search* command on a *corpus* object, which produces new objects, called *result*, *concordance*, *totals* and *query*. As you can see, very complex searches can be performed using an English-like syntax:

```
> search corpus for lemma matching '^t' and pos matching 'VB' \
...     excluding words matching 'try' \
...     showing word and dependent-word \
...     with preserve_case
> result
```

This shows us results for each subcorpus:

.	I/think	I/thought	and/turned	me/told	and/took	I/told	...
chapter1	5	3	2	2	1	3	...
chapter2	7	2	5	3	0	2	...
chapter3	5	5	4	4	1	0	...
chapter4	3	7	1	0	3	1	...
chapter5	7	7	2	1	4	2	...
chapter6	2	0	0	2	1	0	...
chapter7	6	2	6	1	1	3	...
chapter8	3	1	2	2	1	1	...
chapter9	5	7	1	4	6	3	...

8.1 Objects

The most common objects you'll be using are:

Object	Contains
<i>corpus</i>	Dataset selected for parsing or searching
<i>result</i>	Search output
<i>edited</i>	Results after sorting, editing or calculating
<i>concordance</i>	Concordance lines from search
<i>features</i>	General linguistic features of corpus
<i>wordclasses</i>	Distribution of word classes in corpus
<i>postags</i>	Distribution of POS tags in corpus
<i>lexicon</i>	Distribution of lexis in the corpus
<i>figure</i>	Plotted data
<i>query</i>	Values used to perform search or edit
<i>previous</i>	Object created before last
<i>sampled</i>	A sampled corpus
<i>wordlists</i>	A list of wordlists for searching, editing

When you start the interpreter, these are all empty. You'll need to run commands in order to fill them with data. You can also create your own object names using the `call` command.

8.2 Commands

You do things to the objects via commands. Each command has its own syntax, designed to be as similar to natural language as possible. Below is a table of common commands, an explanation of their purpose, and an example of their syntax

Com-mand	Purpose	Syntax
<code>new</code>	Make a new project	<i>new project <name></i>
<code>set</code>	Set current corpus	<i>set <corpusname></i>
<code>parse</code>	Parse corpus	<i>parse corpus with [options]*</i>
<code>search</code>	Search a corpus for linguistic feature, generate concordance	<i>search corpus for [feature matching pattern]* showing [feature]* with [options]*</i>
<code>edit</code>	Edit results or edited results	<i>edit result by [skipping subcorpora/entries matching pattern]* with [options]*</i>
<code>calcule-tate</code>	Calculate relative frequencies, keyness, etc.	<i>calculate result/edited as operation of denominator</i>
<code>sort</code>	Sort results or concordance	<i>sort result/concordance by value</i>
<code>plot</code>	Visualise result or edited result	<i>plot result/edited as line chart with [options]*</i>
<code>show</code>	Show any object	<i>show object</i>
<code>anno-tate</code>	Add annotations to corpus based on search results	<i>annotate all with field as <fieldname> and value as m</i>
<code>unan-notate</code>	Delete annotation fields from corpus	<i>unannotate <fieldname> field</i>
<code>sample</code>	Get a random sample of subcorpora or files from a corpus	<i>sample 5 subcorpora of corpus</i>
<code>call</code>	Name an object (i.e. make a variable)	<i>call object '<name>'</i>
<code>export</code>	Export result, edited result or concordance to string/file	<i>export result to string/csv/latex/file <filename></i>
<code>save</code>	Save data to disk	<i>save object to <filename></i>
<code>load</code>	Load data from disk	<i>load object as result</i>
<code>store</code>	Store something in memory	<i>store object as <name></i>
<code>fetch</code>	Fetch something from memory	<i>fetch <name> as object</i>
<code>help</code>	Get help on an object or command	<i>help command/object</i>
<code>history</code>	See previously entered commands	<i>history</i>
<code>ipython</code>	Enter IPython with objects available	<i>ipython</i>
<code>py</code>	Execute Python code	<i>py 'print("hello world")'</i>
!	Run a line of bash shell	<i>!ls -al data</i>

In square brackets with asterisks are recursive parts of the syntax, which often also accept *not* operators. <*text*> denotes places where you can choose an identifier, filename, etc.

In the pages that follow, the syntax is provided for the most common commands. You can also type the name of the command with no arguments into the interpreter, in order to show usage examples.

8.3 Prompt features

- You can use *history*, *clear*, *ls* and *cd* commands as you would in the shell
- You can execute arbitrary bash commands by beginning the line with an exclamation point (e.g. `!rm data/*`)
- You can use semicolons to put multiple commands on a line (currently needs a space **before and after** the semicolon)
- There is no piping or output redirection (yet), but you can use the *export* and *save* commands to export results
- You can use backslashes to continue writing on the next line
- You can write scripts and pass them to the *corpkit* interpreter

The below is therefore a possible (but terrible) way to write code in *corpkit*:

```
> !du -h data ; set mycorp ; search corpus for words \
... matching any \
... excluding wordlists.closedclass \
... showing lemma and pos ; concordance
```


SETUP

- *Dependencies*
- *Accessing*
- *The prompt*

9.1 Dependencies

To use the interpreter, you'll need *corpkit* installed. To use all features of the interpreter, you will also need *readline* and *IPython*.

9.2 Accessing

With *corpkit* installed, you can start the interpreter in a couple of ways:

```
$ corpkit
# or
$ python -m corpkit.env
```

You can start it from a Python session, too:

```
>>> from corpkit import env
>>> env()
```

9.3 The prompt

When using the interpreter, the prompt (the text to the left of where you type your command) displays the directory you are in (with an asterisk if it does not appear to be a *corpkit* project) and the currently active corpus, if any:

```
corpkit@junglebook: no-corpus>
```

When you see it, *corpkit* is ready to accept commands!

MAKING PROJECTS AND CORPORA

The first two things you need to do when using *corpkit* are to create a project, and to create (and optionally parse) a corpus. These steps can all be accomplished quickly using shell commands. They can also be done using the interpreter, however.

Once you're in *corpkit*, the command below will create a new project called *iran-news*, and move you into it.

```
> new project named iran-news
```

10.1 Adding a corpus

Adding a corpus simply copies it to the project's data directory. The syntax is simple:

```
> add '.../.../my_corpus'
```

10.2 Parsing a corpus

To parse a text file, folder of text files, or folder of folder of text files, you first `set` the corpus, and then use the `parse` command:

```
> set my_corpus as corpus
> parse corpus
```

10.3 Tokenising, POS tagging and lemmatising

If you don't want/need full parses, or if you aren't working with English, you might want to use the `tokenise` method.

```
> set abstracts as corpus
> tokenise corpus
```

POS tagging and lemmatisation are switched on by default, but you could also disable them:

```
> tokenise corpus with postag as false and lemmatise as false
```

10.4 Working with metadata

Parsing/tokenising can be made way cooler when your data has some metadata in it. The metadata will be transferred over to the parsed version of the corpus, and then you can search or filter by metadata features, use metadata values as symbolic subcorpora, or display metadata alongside concordances.

Metadata should take the form of an XML tag at the end of a line, which could be a sentence or a paragraph:

```
I hope everyone is hanging in with this blasted heat. As we all know being hot, sticky,  
stressed and irritated can bring on a mood swing super fast. So please make sure your  
all takeing your meds and try to stay out of the heat. <metadata username="Emz45"  
totalposts="5063" currentposts="4051" date="2011-07-13" postnum="0" threadlength="1">
```

Then, parse with metadata:

```
> parse corpus with metadata
```

The parser output will look something like:

```
# sent_id 1  
# parse=(ROOT (S (NP (PRP I)) (VP (VBP hope) (SBAR (S (NP (NN everyone)) (VP (VBZ is) (VP  
↪(VBG hanging) (PP (IN in) (IN with) (NP (DT this) (VBN blasted) (NN heat))))))) (. .)))  
# speaker=Emz45  
# totalposts=5063  
# threadlength=1  
# currentposts=4051  
# stage=10  
# date=2011-07-13  
# year=2011  
# postnum=0  
1 1 I I PRP O 2 nsubj 0 1  
1 2 hope hope VBP O 0 ROOT 1,5,11 —  
1 3 everyone everyone NN O 5 nsubj 0 —  
1 4 is be VBZ O 5 aux 0 —  
1 5 hanging hang VBG O 2 ccomp 3,4,10 —  
1 6 in in IN O 10 case 0 —  
1 7 with with IN O 10 case 0 —  
1 8 this this DT O 10 det 0 2  
1 9 blasted blast VBN O 10 amod 0 2  
1 10 heat heat NN O 5 nmod:with 6,7,8,9 2*  
1 11 . . . O 2 punct 0 —
```

The next page will show you how to search the corpus you've built, and to work with metadata if you've added it.

INTERROGATING CORPORA

The most powerful thing about *corpkit* is its ability to search parsed corpora for very complex constituency, dependency or token level features.

Before we begin, make sure you've set the corpus as the thing to search:

```
> set nyt-parsed as corpus  
# you could also try just typing `set` ...
```

Note: By default, when using the interpreter, searching also produces concordance lines. If you don't need them, you can use toggle `conc` to switch them off, or on again. This can dramatically speed up processing time.

11.1 Search examples

```
> search corpus ### interactive search helper  
> search corpus for words matching ".*"  
> search corpus for words matching "[A-M]" showing lemma and word with case_sensitive  
> search corpus for cql matching '[pos="DT"] [pos="NN"]' showing pos and word with coref  
> search corpus for function matching roles.process showing dependent-lemma  
> search corpus for governor-lemma matching processes.verbal showing governor-lemma, lemma  
> search corpus for words matching any and not words matching wordlists.closedclass  
> search corpus for trees matching '/NN.?/ >># NP'  
> search corpus for pos matching NNP showing ngram-word and pos with gramsize as 3  
> etc.
```

Under the surface, what you are doing is selecting a *Corpus* object to search, and then generating arguments for the `interrogate()` method. These arguments, in order, are:

1. *search* criteria
2. *exclude* criteria
3. *show* values
4. Keyword arguments

Here is a syntax example that might help you see how the command gets parsed. Note that there are two ways of setting *exclude* criteria.

```
> search corpus \  
... for words matching 'ing$' and \  
... not lemma matching 'being' and \  
... pos matching 'NN' \  
... excluding words matching wordlists.closedclass \  
... showing lemma and pos and function \  
... with preserve_case and \  
... not no_punct and \  
... excludemode as 'all'
```

select object
search criterion
exclude criterion
search criterion
exclude criterion
show values
boolean keyword arg
bool keyword arg
keyword arg

11.2 Working with metadata

By default, *corpkit* treats folders within your corpus as subcorpora. If you want to treat files, rather than folders, as subcorpora, you can do:

```
> search corpus for words matching 'ing$' with subcorpora as files
```

If you have metadata in your corpus, you can use the metadata value as the subcorpora:

```
> search corpus for words matching 'ing$' with subcorpora as speaker
```

If you don't want to keep specifying the subcorpus structure every time you search a corpus, you have a couple of choices. First, you can set the default subcorpus value with `for` for the session with `set subcorpora`. This applies the filter globally, to whatever corpus you search:

```
# use speaker metadata as subcorpora
> set subcorpora as speaker
# ignore folders, use files as subcorpora
> set subcorpora as files
```

You can also define metadata filters, which skip sentences matching a metadata feature, or which keep only sentences matching a metadata feature:

```
# if you have a `year` metadata field, skip this decade
> set skip year as '^201'
# if you want only this decade:
> set keep year as '^201'
```

If you want to set subcorpora and filters for a corpus, rather than globally, you can do this by passing in the values when you select the corpus:

```
> set mydata-parsed as corpus with year as subcorpora and \
... just year as '^201' and skip speaker as 'chomsky'
# forget filters for this corpus:
> set mydata-parsed
```

11.3 Sampling a corpus

Sometimes, your corpus is too big to search quickly. If this is the case, you can use the `sample` command to create a randomise sample of the corpus data:

```
> sample 3 subcorpora of corpus
> sample 100 files of corpus
```

If you pass in a float, it will try to get a proportional amount of data: `sample 0.33 subcorpora of corpus` will return a third of the subcorpora in the corpus.

A sampled corpus becomes an object called `sampled`. You can then refer to it when searching:

```
> search sampled for words matching '^[abcde]'
```

Global metadata filters and subcorpus declarations will be observed when searching this corpus as well.

CONCORDANCING

By default, every search also produces concordance lines. You can view them by typing `concordance`. This opens an interactive display, which can be scrolled and searched—hit `h` to get help on possible commands.

12.1 Customising appearance

The first thing you might want to do is adjust how concordance lines are displayed:

```
# hide subcorpus name, speaker name
> show concordance with columns as lmr
# enlarge window
> show concordance with columns as lmr and window as 60
# limit number of results to 100
> show concordance with columns as lmr and window as 60 and n as 100
```

The values you enter here are persistant—the window size, number of lines, etc. will remain the same until you shut down the interpreter or provide new values.

12.2 Sorting

Sorting can be by column, or by word.

```
# middle column, first word
> sort concordance by M1
# left column, last word
> sort concordance by L1
# right column, third word
> sort concordance by R3
# by index (original order)
> sort concordance by index
```

12.3 Colouring

One nice feature is that concordances can be coloured. This can be done through either indexing or regular expression matches. Note that `background` can be added to colour the background instead of the foreground, and `dim/bright` can be used to adjust text brightness. This means that you can code lines for multiple phenomena. Background highlighting could mark the argument structure, foreground highlighting could mark the mood type, and bright and dim could be used to mark exemplars or false positives.

Note: If you're using Python 2, you may find that colouring concordance lines causes some interference with `readline`, making it difficult to select or search previous commands. This is a limitation of `readline` in Python 2. Use Python 3 instead!

```
# colour by index
> mark 10 blue
> mark -10 background red
> mark 10-15 cyan
> mark 15- magenta
# reset all
> mark - reset
```

```
# regular expression methods: specify column(s) to search
> mark m '^PRP.*' yellow
> mark r 'be(ing)' background green
> mark lm 'JJR$' dim
# reset via regex
> mark m '.*' reset
```

You can then sort by colour with *sort concordance by scheme*. If you export the concordances to a file (*export concordance as csv to conc.csv*), colour information will be added in additional columns.

12.4 Editing

To edit concordance lines, you can use the same syntax as you would use to edit results:

```
> edit concordance by skipping subcorpora matching '[123]$'
> edit concordance by keeping entries matching 'PRP'
```

Perhaps faster is the use of *del* and *keep*. For these, specify the column and the criteria using the same methods as you would for colouring:

```
> del m matching 'this'
> keep l matching '^I\s'
> del 10-20
```

12.5 Recalculating results from concordance lines

If you've deleted some concordance lines, you can update the `result` object to reflect these changes with *calculate result from concordance*.

12.6 Working with metadata

You can use `show_conc_metadata` when interrogating or concordancing to collect and display metadata alongside concordance results:

```
> search corpus for words matching any with show_conc_metadata
> concordance
```

ANNOTATING YOUR CORPUS

Another thing you might like to do is add metadata or annotations to your corpus. This can be done by simply editing corpus files, which are stored in a human-readable format. You can also automate annotation, however.

To do annotation, you first run a search command and generate a concordance. After deleting any false positives from the concordance, you can use the annotate command to annotate each sentence for which a concordance line exists.

`annotate`` works a lot like the `mark`, `keep`, and `del` commands to begin with, but has some special syntax at the end, which controls whether you annotate using *tags*, or *fields and values*.

13.1 Tagging sentences

The first way of annotating is to add a **tag** to one or more sentences:

```
> search corpus for pos matching NNP and word matching 'daisy'  
> annotate m matching '^daisy$' with tag 'has_daisy'
```

You can use *all* to annotate every single concordance line:

```
> search corpus for governor-function matching nsubjpass \  
... showing governor-lemma and lemma  
> annotate all with tag 'passive'
```

If you try to run this code, you actually get a *dry run*, showing you what would be modified in your corpus. Once you're happy with it, you can do `toggle annotation` to turn file writing on, and then run the previous line again (use the up arrow to get it!).

13.2 Creating fields and values

More complex than adding tags is adding **fields and values**. This creates a new metadata category with multiple possible realisations. Below, we tag an sentence sentences based on their containing certain kinds of processes

```
> search corpus for function matching roles.process showing lemma  
> mark m matching processes.verbal red  
# annotate by colour  
> annotate red with field as process \  
... and value as 'verbal'  
# annotate without colouring first  
> annotate m matching processes.mental with field as process \  
... and value as 'mental'
```

You can also use `m` as the value, which passes in the text from the middle column of the concordance.

```
> search corpus for pos matching NNP showing word  
> annotate m matching [gatsby, daisy, tom] \  
... with field as character and value as m
```

The moment these values have been added to your text, you can do really powerful things with them. You can, for example, use them as subcorpora, or use them as filters for the sentences being processed.

```
> set subcorpora as process  
> set skip character as 'gatsby'  
> set skip passive tag
```

Now, the subcorpora will be the different processes (*verbal*, *mental* and *none*), and any sentence annotated as containing the gatsby character, or the passive tag, will be ignored.

13.3 Removing annotations

To remove a tag or a field across the dataset, the commands are very simple. Note that again, you need to toggle annotation to actually alter any files.

```
> unannotate character field  
> unannotate typo tag  
> unannotate all tags
```

CHAPTER
FOURTEEN

EDITING RESULTS

Once you have generated a *result* object via the *search* command, you can edit the result in a number of ways. You can delete, merge or otherwise alter entries or subcorpora; you can do statistics, and you can sort results.

Editing, calculating and sorting each create a new object, called *edited*. This means that if you make a mistake, you still have access to the original *result* object, without needing to run the search again.

14.1 The edit command

When using the *edit* command, the main things you'll want to do is skip, keep, span or merge results or subcorpora.

```
> edit result by keeping subcorpora matching '[01234]'  
> edit result by skipping entries matching wordlists.closedclass  
# merge has a slightly different syntax, because you need  
# to specify the name to merge under  
> edit result by merging entries matching 'be|have' as 'aux'
```

Note: The syntax above works for concordance lines too, if you change *result* to *concordance*. Merging is not possible.

14.2 Doing basic statistics

The *calculate* command allows you to turn the absolute frequencies into relative frequencies, keyness scores, etc.

```
> calculate result as percentage of self  
> calculate edited as percentage of features.clauses  
> calculate result as keyness of self
```

If you want to run more complicated operations on the results, you might like to use the *ipython* command to enter an IPython session, and then manipulate the Pandas objects directly.

14.3 Sorting results

The *sort* command allows you to change the search result order.

Possible values are *total*, *name*, *infreq*, *increase*, *decrease*, *static*, *turbulent*.

```
> sort result by total  
# requires scipy  
> sort edited by increase
```

CHAPTER
FIFTEEN

PLOTTING

You can plot results and edited results using the `plot` method, which interfaces with `matplotlib`.

```
> plot edited as bar chart with title as 'Example plot' and x_label as 'Subcorpus'  
> plot edited as area chart with stacked and colours as Paired  
> plot edited with style as seaborn-talk # defaults to line chart
```

There are many possible arguments for customising the figure. The table below shows some of them.

```
> plot edited as bar chart with rot as 45 and logy and \  
...     legend_alpha as 0.8 and show_p_val and not grid
```

Argument	Type	Action
<code>grid</code>	<code>bool</code>	Show grid in background
<code>rot</code>	<code>int</code>	Rotate x axis labels n degrees
<code>shadow</code>	<code>bool</code>	Shadows for some parts of plot
<code>ncol</code>	<code>int</code>	n columns for legend entries
<code>explode</code>	<code>list</code>	Explode these entries in pie
<code>partial_pie</code>	<code>bool</code>	Allow plotting of pie slices
<code>legend_frame</code>	<code>bool</code>	Show frame around legend
<code>legend_alpha</code>	<code>float</code>	Opacity of legend
<code>reverse_legend</code>	<code>bool</code>	Reverse legend entry order
<code>transpose</code>	<code>bool</code>	Flip axes of DataFrame
<code>logx/logy</code>	<code>bool</code>	Log scales
<code>show_p_val</code>	<code>bool</code>	Try to show p value in legend

Note: If you want to set a boolean value, you can just say `and` value or `not` value. If you like, however, you could write it more fully as `with` value as `true/false` as well.

SETTINGS AND MANAGEMENT

The interpreter can do a number of other useful things. They are outlined here.

16.1 Managing data

You should be able to store most of the objects you create in memory using the `store` command:

```
> store result as 'good_result'  
> show store  
> fetch 'good_result' as result
```

A more permanent solution is to use `save` and `load`:

```
> save result as 'good_result'  
> ls saved_interrogations  
> load 'good_result' as result
```

An alternative approach is to create variables using the `call` command:

```
> search corpus for words matching any  
> call result anyword  
> calculate anyword as percentage of self
```

A variable can also be a simple string, which you can then add into searches:

```
> call '/NN.?/' >># NP' headnoun  
> search corpus for trees matching headnoun
```

To forget a variable, just do `remove <name>`.

16.2 Toggles and settings

- Using `toggle interactive`, You can switch between interactive mode, where results and concordances are shown in a way that you can manipulate directly, and non-interactive mode, where results and concordances are simply printed to the console.
- Using `toggle conc`, you can tell `corpuskit` not to produce concordances. This can be much faster, especially when there are a lot of results.
- You can set the number of decimals displayed when viewing results with `set decimal to <n>`

16.3 Switching to IPython

When the interpreter constrains you, you can switch to IPython with `ipython`. Your objects are available there under the same name. When you're done there, do `quit` to return to the `corpuskit` interpreter.

16.4 Running scripts

You can also write and run scripts. If you make a file, `participants.cki`, containing:

```
#!/usr/bin/env corpkit

set mydata-parsed as corpus
search corpus for function matching roles.participant showing lemma
export result as csv to part.csv
```

You can run it from the terminal with:

```
corpkit participants.cki
# or, directly, if there's a shebang and chmod +x:
./participants.cki
```

which will leave you with a CSV file at `exported/part.csv`. This approach can be handy if you need to pipe `stdout` or `stderr`, or if you want to call `corpkit` within a shell script.

Note: When running a script, interactivity will automatically be switched off, and concordancing disabled if the script does not appear to need it.

CHAPTER
SEVENTEEN

CORPUS CLASSES

Much of *corplib*'s functionality comes from the ability to work with `Corpus` and `Corpus`-like objects, which have methods for parsing, tokenising, interrogating and concordancing.

17.1 *Corpus*

class `corplib.corpus.Corpus` (*path*, `**kwargs`)
Bases: `object`

A class representing a linguistic text corpus, which contains files, optionally within subcorpus folders.

Methods for concordancing, interrogating, getting general stats, getting behaviour of particular word, etc.

Unparsed, tokenised and parsed corpora use the same class, though some methods are available only to one or the other. Only unparsed corpora can be parsed, and only parsed/tokenised corpora can be interrogated.

subcorpora

A list-like object containing a corpus' subcorpora.

Example

```
>>> corpus.subcorpora
<corplib.corpus.Datalist instance: 12 items>
```

speakerlist

Lazy-loaded data.

files

A list-like object containing the files in a folder.

Example

```
>>> corpus.subcorpora[0].files
<corplib.corpus.Datalist instance: 240 items>
```

all_filepaths

Lazy-loaded data.

all_files

Lazy-loaded data.

tfidf (*search={‘w’: ‘any’}*, *show=[‘w’]*, `**kwargs`)

Generate TF-IDF vector representation of corpus using interrogate method. All args and kwargs go to `interrogate()`.

Returns Tuple: the vectoriser and matrix

features

Generate and show basic stats from the corpus, including number of sentences, clauses, process types, etc.

Example

	..	Characters	Tokens	Words	Closed class words	Open class words	Clauses
01		26873	8513	7308	4809	3704	2212
02		25844	7933	6920	4313	3620	2270
03		18376	5683	4877	3067	2616	1640
04		20066	6354	5366	3587	2767	1775

wordclasses

Lazy-loaded data.

postags

Lazy-loaded data.

lexicon

Lazy-loaded data.

configurations (*search*, ***kwargs*)

Get the overall behaviour of tokens or lemmas matching a regular expression. The search below makes DataFrames containing the most common subjects, objects, modifiers (etc.) of ‘see’:

Parameters **search** (*dict*) – Similar to *search* in the [interrogate\(\)](#) method.

Valid keys are:

- *W/L* match word or lemma
- *F*: match a semantic role (‘*participant*’, ‘*process*’ or ‘*modifier*’). If *F* not specified, each role will be searched for.

Example

>>> see = corpus.configurations({L: 'see', F: 'process'}, show=L)
>>> see.has_subject.results.sum()
i 452
it 227
you 162
we 111
he 94

Returns [corpkit.interrogation.Interrodict](#)

interrogate (*search*=‘w’, **args*, ***kwargs*)

Interrogate a corpus of texts for a lexicogrammatical phenomenon.

This method iterates over the files/folders in a corpus, searching the texts, and returning a [corpkit.interrogation.Interrogation](#) object containing the results. The main options are *search*, where you specify search criteria, and *show*, where you specify what you want to appear in the output.

Example

>>> corpus = Corpus('data/conversations-parsed')
show lemma form of nouns ending in 'ing'
>>> q = {W: r'ing\$', P: r'^N'}
>>> data = corpus.interrogate(q, show=L)
>>> data.results
.. something anything thing feeling everything nothing morning
01 14 11 12 1 6 0 1
02 10 20 4 4 8 3 0
03 14 5 5 3 1 0 0
...

Parameters **search** (*dict*) – What part of the lexicogrammar to search, and what criteria to match. The *keys* are the thing to be searched, and values are the criteria. To search parse trees, use the *T* key, and a Tregex query as the value. When searching dependencies, you can use any of:

	Match	Governor	Dependent	Head
Word	<i>W</i>	<i>G</i>	<i>D</i>	<i>H</i>
Lemma	<i>L</i>	<i>GL</i>	<i>DL</i>	<i>HL</i>
Function	<i>F</i>	<i>GF</i>	<i>DF</i>	<i>HF</i>
POS tag	<i>P</i>	<i>GP</i>	<i>DP</i>	<i>HP</i>
Word class	<i>X</i>	<i>GX</i>	<i>DX</i>	<i>HX</i>
Distance from root	<i>A</i>	<i>GA</i>	<i>DA</i>	<i>HA</i>
Index	<i>I</i>	<i>GI</i>	<i>DI</i>	<i>HI</i>
Sentence index	<i>S</i>	<i>SI</i>	<i>SI</i>	<i>SI</i>

Values should be regular expressions or wordlists to match.

Example

```
>>> corpus.interrogate({T: r'/NN.?/ < /^t/'}) # T- nouns, via trees
>>> corpus.interrogate({W: '^t': P: r'^v'}) # T- verbs, via dependencies
```

Parameters

- **searchmode** (*str – ‘any’/‘all’*) – Return results matching any/all criteria
- **exclude** (*dict – {L: ‘be’}*) – The inverse of *search*, removing results from search
- **excludemode** (*str – ‘any’/‘all’*) – Exclude results matching any/all criteria
- **query** (*str, dict or list*) – A search query for the interrogation. This is only used when *search* is a *str*, or when multiprocessing. When *search* is a *str*, the search criteria can be passed in as ‘query’, in order to allow the simpler syntax:

```
>>> corpus.interrogate(GL, '(think|want|feel)')
```

When multiprocessing, the following is possible:

```
>>> q = {'Nouns': r'/NN.?/', 'Verbs': r'/VB.?/'}
### return an :class:`corpkit.interrogation.Interrogation` object with_
˓→multiindex:
>>> corpus.interrogate(T, q)
### return an :class:`corpkit.interrogation.Interrogation` object_
˓→without multiindex:
>>> corpus.interrogate(T, q, show=C)
```

- **show** (*str/list of strings*) – What to output. If multiple strings are passed in as a *list*, results will be colon-separated, in the supplied order. Possible values are the same as those for *search*, plus options n-gramming and getting collocates:

Show	Gloss	Example
N	N-gram word	<i>The women were</i>
NL	N-gram lemma	<i>The woman be</i>
NF	N-gram function	<i>det nsubj root</i>
NP	N-gram POS tag	<i>DT NNS VBN</i>
NX	N-gram word class	<i>determiner noun verb</i>
B	Collocate word	<i>The_were</i>
BL	Collocate lemma	<i>The_be</i>
BF	Collocate function	<i>det_root</i>
BP	Collocate POS tag	<i>DT_VBN</i>
BX	Collocate word class	<i>determiner_verb</i>

- **lemmatise** (*bool*) – Force lemmatisation on results. **Deprecated: instead, output a lemma form with the ‘show’ argument**
- **lemmatag** (*‘n’/‘v’/‘a’/‘r’/False*) – When using a Tregex/Tgrep query, the tool will attempt to determine the word class of results from the query. Passing in a *str* here will tell the lemmatiser the expected POS of results to lemmatise. It only has an affect if trees are being searched and lemmata are being shown.

- **save** (*str*) – Save result as pickle to *saved_interrogations/<save>* on completion
- **gramsize** (*int*) – Size of n-grams (default 1, i.e. unigrams)
- **multiprocess** (*int/bool (bool determines automatically)*) – How many parallel processes to run
- **files_as_subcorpora** (*bool*) – (**Deprecated, use subcorpora=files**). Treat each file as a subcorpus, ignoring actual subcorpora if present
- **conc** (*bool/only*) – Generate a concordance while interrogating, store as *.concordance* attribute
- **coref** (*bool*) – Also get coreferents for search matches
- **tgrep** (*bool*) – Use *TGrep* for tree querying. *TGrep* is less expressive than *Tregex*, and is slower, but can work without Java. This option may be turned on internally if Java is not found.
- **subcorpora** (*str/list*) – Use a metadata value as subcorpora. Passing a list will create a multiindex. ‘file’ and ‘folder’/‘default’ are also possible values.
- **just_metadata** (*dict*) – One or more metadata fields and criteria to filter sentences by. Only those matching will be kept. Criteria can be a list of words or a regular expression. Passing {‘speaker’: ‘ENVER’} will search only sentences annotated with speaker=ENVER.
- **skip_metadata** (*dict*) – A field and regex/list to filter sentences by. The inverse of just_metadata.
- **discard** (*int/float*) – When returning many (i.e. millions) of results, memory can be a problem. Setting a discard value will ignore results occurring infrequently in a subcorpus. An int will remove any result occurring *n* times or fewer. A float will remove this proportion of results (i.e. 0.1 will remove 10 per cent)

Returns A `corpkit.interrogation.Interrogation` object, with *.query*, *.results*, *.totals* attributes. If multiprocessing is invoked, result may be multiindexed.

sample (*n, level=f*)

Get a sample of the corpus

Parameters

- **n** (*int/float*) – amount of data in the the sample. If an *int*, get *n* files. if a *float*, get *float* * 100 as a percentage of the corpus
- **level** (*str*) – sample subcorpora (s) or files (f)

Returns a Corpus object

delete_metadata ()

Delete metadata for corpus. May be needed if corpus is changed

metadata

Lazy-loaded data.

parse (*corenlpPath=False, operations=False, copula_head=True, speaker_segmentation=False, memory_mb=False, multiprocess=False, split_texts=400, outname=False, metadata=False, coref=True, *args, **kwargs*)

Parse an unparsed corpus, saving to disk

Parameters

- **corenlpPath** (*str*) – Folder containing corenlp jar files (use if *corpkit* can’t find it automatically)
- **operations** (*str*) – Which kinds of annotations to do

- **speaker_segmentation** (*bool*) – Add speaker name to parser output if your corpus is script-like
- **memory_mb** (*int*) – Amount of memory in MB for parser
- **copula_head** (*bool*) – Make copula head in dependency parse
- **split_texts** – Split texts longer than *n* lines for parser memory
- **multiprocess** (*int*) – Split parsing across n cores (for high-performance computers)
- **folderise** (*bool*) – If corpus is just files, move each into own folder
- **output_format** (*str*) – Save parser output as *xml*, *json*, *conll*
- **outname** (*str*) – Specify a name for the parsed corpus
- **metadata** (*bool*) – Use if you have XML tags at the end of lines containing metadata

Example

```
>>> parsed = corpus.parse(speaker_segmentation=True)
>>> parsed
<corpkit.corpus.Corpora instance: speeches-parsed; 9 subcorpora>
```

Returns The newly created *corpkit.corpus.Corpora*

tokenise (*postag=True*, *lemmatise=True*, **args*, ***kwargs*)
Tokenise a plaintext corpus, saving to disk

Parameters **nltk_data_path** (*str*) – Path to tokeniser if not found automatically**Example**

```
>>> tok = corpus.tokenise()
>>> tok
<corpkit.corpus.Corpora instance: speeches-tokenised; 9 subcorpora>
```

Returns The newly created *corpkit.corpus.Corpora*

concordance (**args*, ***kwargs*)
A concordance method for Tregex queries, CoreNLP dependencies, tokenised data or plaintext.

Example

```
>>> wv = ['want', 'need', 'feel', 'desire']
>>> corpus.concordance({L: wv, F: 'root'})
0   01  1-01.txt.conll          But , so I   feel      like i do that for w
1   01  1-01.txt.conll          I   felt      a little like oh , i
2   01  1-01.txt.conll          he 's a difficult man I   feel      like his work ethic
3   01  1-01.txt.conll          So I   felt      like i recognized li
...
...
```

Arguments are the same as *interrogate()*, plus a few extra parameters:**Parameters**

- **only_format_match** (*bool*) – If *True*, left and right window will just be words, regardless of what is in *show*
- **only_unique** (*bool*) – Return only unique lines
- **maxconc** (*int*) – Maximum number of concordance lines

Returns A *corpkit.interrogation.Concordance* instance, with columns showing filename, subcorpus name, speaker name, left context, match and right context.

interroplot(*search*, ***kwargs*)

Interrogate, relativise, then plot, with very little customisability. A demo function.

Example

```
>>> corpus.interroplot(r'/NN.?/ >># NP')
```

Parameters

- **search** (*dict*) – Search as per *interrogate()*
- **kwargs** (*keyword arguments*) – Extra arguments to pass to *visualise()*

Returns *None* (but show a plot)**save**(*savename=False*, ***kwargs*)

Save corpus instance to file. There's not much reason to do this, really.

```
>>> corpus.save(filename)
```

Parameters **savename** (*str*) – Name for the file**Returns** *None***make_language_model**(*name*, *search={‘w’: ‘any’}*, *exclude=False*, *show=[‘w’, ‘+1mw’]*, ***kwargs*)

Make a language model for the corpus

Parameters

- **name** (*str*) – a name for the model
- **kwargs** (*keyword arguments*) – keyword arguments for the *interrogate()* method

Returns a *corpkit.model.MultiModel***annotate**(*conclines*, *annotation*, *dry_run=True*)

Annotate a corpus

Parameters

- **conclines** – a Concordance or DataFrame containing matches to annotate
- **annotation** (*str/dict*) – a tag or field and value
- **dry_run** (*bool*) – Show the annotations to be made, but don't do them

Returns *None***unannotate**(*annotation*, *dry_run=True*)

Delete annotation from a corpus

Parameters **annotation** (*str/dict*) – a tag or field and value**Returns** *None*

17.2 Corpora

class *corpkit.corpus.Corpora*(*data=False*, ***kwargs*)

Bases: *corpkit.corpus.Datalist*

Models a collection of Corpus objects. Methods are available for interrogating and plotting the entire collection. This is the highest level of abstraction available.

Parameters **data** (*str/list*) – Corpora to model. A *str* is interpreted as a path containing corpora. A *list* can be a list of corpus paths or *corpkit.corpus.Corpus* objects.)

parse (**kwargs)
Parse multiple corpora

Parameters **kwargs** – Arguments to pass to the `parse()` method.

Returns `corpkit.corpus.Corpora`

features

Generate and show basic stats from the corpus, including number of sentences, clauses, process types, etc.

Example

```
>>> corpus.features
..  Characters  Tokens  Words  Closed class words  Open class words  Clauses
01      26873     8513    7308                  4809            3704      2212
02      25844     7933    6920                  4313            3620      2270
03      18376     5683    4877                  3067            2616      1640
04      20066     6354    5366                  3587            2767      1775
```

postags

Lazy-loaded data.

wordclasses

Lazy-loaded data.

lexicon

Lazy-loaded data.

17.3 Subcorpus

class `corpkit.corpus.Subcorpus` (*path*, *datatype*, ***kwa*)
Bases: `corpkit.corpus.Corpora`

Model a subcorpus, containing files but no subdirectories.

Methods for interrogating, concordancing and configurations are the same as `corpkit.corpus.Corpora`.

17.4 File

class `corpkit.corpus.File` (*path*, *dirname=False*, *datatype=False*, ***kwa*)
Bases: `corpkit.corpus.Corpora`

Models a corpus file for reading, interrogating, concordancing.

Methods for interrogating, concordancing and configurations are the same as `corpkit.corpus.Corpora`, plus methods for accessing the file contents directly as a *str*, or as a Pandas DataFrame.

read(**kwargs)

Read file data. If data is pickled, unpickle first

Returns *str/unpickled data*

document

Return a DataFrame representation of a parsed file

trees

Lazy-loaded data.

plain

Lazy-loaded data.

17.5 *Datalist*

```
class corpkit.corpus.Datalist(data, **kwargs)
    Bases: list

    interrogate(*args, **kwargs)
        Interrogate the corpus using interrogate\(\)

    concordance(*args, **kwargs)
        Concordance the corpus using concordance\(\)

    configurations(search, **kwargs)
        Get a configuration using configurations\(\)
```

INTERROGATION CLASSES

Once you have searched a `Corpus` object, you'll want to be able to edit, visualise and store results. Remember that upon importing `corplib`, any `pandas.DataFrame` or `pandas.Series` object is monkey-patched with `save`, `edit` and `visualise` methods.

18.1 *Interrogation*

```
class corpkit.interrogation.Interrogation(results=None, totals=None, query=None, concordance=None)
```

Bases: `object`

Stores results of a corpus interrogation, before or after editing. The main attribute, `results`, is a Pandas object, which can be edited or plotted.

results = None

pandas `DataFrame` containing counts for each subcorpus

totals = None

pandas `Series` containing summed results

query = None

`dict` containing values that generated the result

concordance = None

pandas `DataFrame` containing concordance lines, if concordance lines were requested.

edit (*args, **kwargs)

Manipulate results of interrogations.

There are a few overall kinds of edit, most of which can be combined into a single function call. It's useful to keep in mind that many are basic wrappers around `pandas` operations—if you're comfortable with `pandas` syntax, it may be faster at times to use its syntax instead.

Basic mathematical operations

First, you can do basic maths on results, optionally passing in some data to serve as the denominator. Very commonly, you'll want to get relative frequencies:

Example

```
>>> data = corpus.interrogate({W: r'^t'})  
>>> rel = data.edit('%', SELF)  
>>> rel.results  
.. to that the then ... toilet tolerant tolerate ton  
01 18.50 14.65 14.44 6.20 ... 0.00 0.00 0.11 0.00  
02 24.10 14.34 13.73 8.80 ... 0.00 0.00 0.00 0.00  
03 17.31 18.01 9.97 7.62 ... 0.00 0.00 0.00 0.00
```

For the operation, there are a number of possible values, each of which is to be passed in as a `str`:

+, -, /, *, %: self explanatory

k: calculate keywords

a: get distance metric

SELF is a very useful shorthand denominator. When used, all editing is performed on the data. The totals are then extracted from the edited data, and used as denominator. If this is not the desired behaviour, however, a more specific *interrogation.results* or *interrogation.totals* attribute can be used.

In the example above, *SELF* (or ‘self’) is equivalent to:

Example

```
>>> rel = data.edit('%', data.totals)
```

Keeping and skipping data

There are four keyword arguments that can be used to keep or skip rows or columns in the data:

- just_entries*
- just_subcorpora*
- skip_entries*
- skip_subcorpora*

Each can accept different input types:

- str*: treated as regular expression to match
- list*:
 - of integers: indices to match
 - of strings: entries/subcorpora to match

Example

```
>>> data.edit(just_entries=r'^fr',  
...             skip_entries=['free', 'freedom'],  
...             skip_subcorpora=r'[0-9]')
```

Merging data

There are also keyword arguments for merging entries and subcorpora:

- merge_entries*
- merge_subcorpora*

These take a *dict*, with the new name as key and the criteria as value. The criteria can be a str (regex) or wordlist.

Example

```
>>> from dictionaries.wordlists import wordlists  
>>> mer = {'Articles': ['the', 'an', 'a'], 'Modals': wordlists.modals}  
>>> data.edit(merge_entries=mer)
```

Sorting

The *sort_by* keyword argument takes a *str*, which represents the way the result columns should be ordered.

- increase*: highest to lowest slope value

- *decrease*: lowest to highest slope value
- *turbulent*: most change in y axis values
- *static*: least change in y axis values
- *total/most*: largest number first
- *infreq/least*: smallest number first
- *name*: alphabetically

Example

```
>>> data.edit(sort_by='increase')
```

Editing text

Column labels, corresponding to individual interrogation results, can also be edited with *replace_names*.

Parameters `replace_names` (*str/list of tuples/dict*) – Edit result names, then merge duplicate entries

If *replace_names* is a string, it is treated as a regex to delete from each name. If *replace_names* is a dict, the value is the regex, and the key is the replacement text. Using a list of tuples in the form (*find, replacement*) allows duplicate substitution values.

Example

```
>>> data.edit(replace_names={r'object': r'[di]obj'})
```

Parameters `replace_subcorpus_names` (*str/list of tuples/dict*) – Edit subcorpus names, then merge duplicates. The same as *replace_names*, but on the other axis.

Other options

There are many other miscellaneous options.

Parameters

- `keep_stats` (*bool*) – Keep/drop stats values from dataframe after sorting
- `keep_top` (*int*) – After sorting, remove all but the top *keep_top* results
- `just_totals` (*bool*) – Sum each column and work with sums
- `threshold` (*int(bool)*) –

When using results list as dataframe 2, drop values occurring fewer than n times.
If not keywording, you can use:

'high': denominator total / 2500

'medium': denominator total / 5000

'low': denominator total / 10000

If keywording, there are smaller default thresholds

- `span_subcorpora` (*tuple – (int, int2)*) – If subcorpora are numerically named, span all from *int* to *int2*, inclusive
- `projection` (*tuple – (subcorpus_name, n)*) – multiply results in subcorpus by n
- `remove_above_p` (*bool*) – Delete any result over *p*
- `p` (*float*) – set the p value

- **revert_year** (*bool*) – When doing linear regression on years, turn annual subcorpora into 1, 2 ...
- **print_info** (*bool*) – Print stuff to console showing what's being edited
- **spelling** (*str* – ‘US’/‘UK’) – Convert/normalise spelling:

Keywording options

If the operation is *k*, you're calculating keywords. In this case, some other keyword arguments have an effect:

Parameters `keyword_measure` – what measure to use to calculate keywords:

ll: log-likelihood ‘pd’: percentage difference

type `keyword_measure`: *str*

Parameters

- **selfdrop** (*bool*) – When keywording, try to remove target corpus from reference corpus
- **calc_all** (*bool*) – When keywording, calculate words that appear in either corpus

Returns `corpkit.interrogation.Interrogation`

sort (*way*, ***kwargs*)

visualise (*title*=‘’, *x_label*=*None*, *y_label*=*None*, *style*=‘ggplot’, *figsize*=(8, 4), *save*=*False*, *legend_pos*=‘best’, *reverse_legend*=‘guess’, *num_to_plot*=7, *tex*=‘try’, *colours*=‘Accent’, *cumulative*=*False*, *pie_legend*=*True*, *rot*=*False*, *partial_pie*=*False*, *show_totals*=*False*, *transparent*=*False*, *output_format*=‘png’, *interactive*=*False*, *black_and_white*=*False*, *show_p_val*=*False*, *indices*=*False*, *transpose*=*False*, ***kwargs*)

Visualise corpus interrogations using *matplotlib*.

Example

```
>>> data.visualise('An example plot', kind='bar', save=True)
<matplotlib figure>
```

Parameters

- **title** (*str*) – A title for the plot
- **x_label** (*str*) – A label for the x axis
- **y_label** (*str*) – A label for the y axis
- **kind** (*str* (‘line’/‘bar’/‘barh’/‘pie’/‘area’/‘heatmap’)) – The kind of chart to make
- **style** (*str* (‘ggplot’/‘bmh’/‘fivethirtyeight’/‘seaborn-talk’/etc)) – Visual theme of plot
- **figsize** (*tuple* – (*int*, *int*)) – Size of plot
- **save** (*bool*/*str*) – If *bool*, save with *title* as name; if *str*, use *str* as name
- **legend_pos** (*str* (‘upper right’/‘outside right’/etc)) – Where to place legend
- **reverse_legend** (*bool*) – Reverse the order of the legend
- **num_to_plot** (*int*/‘all’) – How many columns to plot
- **tex** (*bool*) – Use TeX to draw plot text
- **colours** (*str*) – Colourmap for lines/bars/slices
- **cumulative** (*bool*) – Plot values cumulatively
- **pie_legend** (*bool*) – Show a legend for pie chart

- **partial_pie** (*bool*) – Allow plotting of pie slices only
- **show_totals** (*str* – ‘legend’/‘plot’/‘both’) – Print sums in plot where possible
- **transparent** (*bool*) – Transparent .png background
- **output_format** (*str* – ‘png’/‘pdf’) – File format for saved image
- **black_and_white** (*bool*) – Create black and white line styles
- **show_p_val** (*bool*) – Attempt to print p values in legend if contained in df
- **indices** (*bool*) – To use when plotting “distance from root”
- **stacked** (*str*) – When making bar chart, stack bars on top of one another
- **filled** (*str*) – For area and bar charts, make every column sum to 100
- **legend** (*bool*) – Show a legend
- **rot** (*int*) – Rotate x axis ticks by *rot* degrees
- **subplots** (*bool*) – Plot each column separately
- **layout** (*tuple* – (*int*, *int*)) – Grid shape to use when *subplots* is True
- **interactive** (*list* – [1, 2, 3]) – Experimental interactive options

Returns matplotlib figure

multiplot (*leftdict*={}, *rightdict*={}, ***kwargs*)

language_model (*name*, **args*, ***kwargs*)

Make a language model from an Interrogation. This is usually done directly on a `corpkit.corpus.Corpus` object with the `make_language_model()` method.

save (*savename*, *savedir*=‘*saved_interrogations*’, ***kwargs*)

Save an interrogation as pickle to *savedir*.

Example

```
>>> o = corpus.interrogate(W, 'any')
### create ./saved_interrogations/savename.p
>>> o.save('savename')
```

Parameters

- **savename** (*str*) – A name for the saved file
- **savedir** (*str*) – Relative path to directory in which to save file
- **print_info** (*bool*) – Show/hide stdout

Returns None

quickview (*n*=25)

view top *n* results as painlessly as possible.

Example

```
>>> data.quickview(n=5)
0: to      (n=2227)
1: that    (n=2026)
2: the     (n=1302)
3: then    (n=857)
4: think   (n=676)
```

Parameters *n* (*int*) – Show top *n* results

Returns None

```
tabview(**kwargs)
asciiplot(row_or_col_name, axis=0, colours=True, num_to_plot=100, line_length=120,
            min_graph_length=50, separator_length=4, multivalue=False, human_readable='si',
            graphsymbol='*', float_format='{:,.2f}', **kwargs)
A very quick ascii chart for result

rel(denominator='self', **kwargs)

keyness(measure='ll', denominator='self', **kwargs)

multiindex(indexnames=None)
Create a pandas.MultiIndex object from slash-separated results.
```

Example

```
>>> data = corpus.interrogate({W: 'st$'}, show=[L, F])
>>> data.results
.. just/admod almost/admod last/amod
01      79        12       6
02     105        6       7
03      86        10       1
>>> data.multiindex().results
Lemma      just almost last first most
Function   advmod advmod amod amod advmod
0          79    12    6    2    3
1         105    6    7    1    3
2          86    10    1    3    0
```

Parameters **indexnames** (list of strings) – provide custom names for the new index, or leave blank to guess.

Returns *corpkit.interrogation.Interrogation*, with *pandas.MultiIndex* as

results attribute

topwords(datatype='n', n=10, df=False, sort=True, precision=2)

Show top n results in each corpus alongside absolute or relative frequencies.

Parameters

- **datatype** (str (n/k/%)) – show abs/rel frequencies, or keyness
- **n** (int) – number of result to show
- **df** (bool) – return a DataFrame
- **sort** (bool) – Sort results, or show as is
- **precision** (int) – float precision to show

Example

```
>>> data.topwords(n=5)
1987      % 1988      % 1989      % 1990      %
health    25.70  health   15.25  health   19.64  credit   9.22
security  6.48   cancer   10.85  security  7.91   health   8.31
cancer    6.19   heart    6.31   cancer   6.55   downside 5.46
flight    4.45   breast   4.29   credit   4.08   inflation 3.37
safety    3.49   security  3.94   safety   3.26   cancer   3.12
```

Returns None

perplexity()

Pythonification of the formal definition of perplexity.

input: a sequence of chances (any iterable will do) output: perplexity value.

from https://github.com/zeffi/NLP_class_notes

```
entropy()
entropy(pos.edit(merge_entries=mergetags, sort_by='total').results.T

shannon()
```

18.2 Interrodict

class corpkit.interrogation.**Interrodict** (*data*)

Bases: collections.OrderedDict

A class for interrogations that do not fit in a single-indexed DataFrame.

Individual interrogations can be looked up via dict keys, indexes or attributes:

Example

```
>>> out_data['WSJ'].results
>>> out_data.WSJ.results
>>> out_data[3].results
```

Methods for saving, editing, etc. are similar to corpkit.corpus.Interrogation. Additional methods are available for collapsing into single (multi-indexed) DataFrames.

This class is now deprecated, in favour of a multiindexed DataFrame.

edit (**args*, ***kwargs*)

Edit each value with `edit()`.

See `edit()` for possible arguments.

Returns A `corpkit.interrogation.Interrodict`

multiindex (*indexnames=False*)

Create a `pandas.MultiIndex` version of results.

Example

```
>>> d = corpora.interrogate({F: 'compound', GL: '^risk'}, show=L)
>>> d.keys()
['CHT', 'WAP', 'WSJ']
>>> d['CHT'].results
.... health cancer security credit flight safety heart
1987    87     25      28     13      7     6     4
1988    72     24      20     15      7     4     9
1989   137     61      23     10      5     5     6
>>> d.multiindex().results
...
Corpus Subcorpus
CHT    1987        87     25     13      28     20
          1988        72     24     15      20     12
          1989       137     61     10      23     10
WAP    1987        83     44      8      44     10
          1988        83     27     13      40      6
          1989       95     77     18      25     12
WSJ    1987        52     27     33      4     21
          1988        39     11     37      9     22
          1989       55     47     43      9     24
```

Returns A `corpkit.interrogation.Interrogation`

save (*savename*, *savedir='saved_interrogations'*, ***kwargs*)

Save an interrogation as pickle to *savedir*.

Parameters

- **savename** (*str*) – A name for the saved file

- **savedir** (str) – Relative path to directory in which to save file
- **print_info** (bool) – Show/hide stdout

Example

```
>>> o = corpus.interrogate(W, 'any')
## create ``saved_interrogations/savename.p``
>>> o.save('savename')
```

Returns None**collapse** (axis='y')

Collapse Interrodict on an axis or along interrogation name.

Parameters **axis** (str: x/y/n) – collapse along x, y or name axis**Example**

```
>>> d = corpora.interrogate({F: 'compound', GL: r'^risk'}, show=L)

>>> d.keys()
['CHT', 'WAP', 'WSJ']

>>> d['CHT'].results
.... health cancer security credit flight safety heart
1987     87      25      28     13      7      6      4
1988     72      24      20     15      7      4      9
1989    137      61      23     10      5      5      6

>>> d.collapse().results
... health cancer credit security
CHT    3174    1156    566    697
WAP    2799     933    582   1127
WSJ    1812     680   2009    537

>>> d.collapse(axis='x').results
... 1987 1988 1989
CHT   384   328   464
WAP   389   355   435
WSJ   428   410   473

>>> d.collapse(axis='key').results
... health cancer credit security
1987     282     127      65     93
1988     277     100      70    107
1989     379     253      83     91
```

Returns A *corpkit.interrogation.Interrogation***topwords** (datatype='n', n=10, df=False, sort=True, precision=2)

Show top n results in each corpus alongside absolute or relative frequencies.

Parameters

- **datatype** (str (n/k/*)) – show abs/rel frequencies, or keyness
- **n** (int) – number of result to show
- **df** (bool) – return a DataFrame
- **sort** (bool) – Sort results, or show as is
- **precision** (int) – float precision to show

Example

```
>>> data.topwords(n=5)
          %      UST          %      WAP          %      WSJ          %
TBT      25.70  health      15.25  health     19.64  credit      9.22
```

security	6.48	cancer	10.85	security	7.91	health	8.31
cancer	6.19	heart	6.31	cancer	6.55	downside	5.46
flight	4.45	breast	4.29	credit	4.08	inflation	3.37
safety	3.49	security	3.94	safety	3.26	cancer	3.12

Returns None

visualise (*shape='auto'*, *truncate=8*, ***kwargs*)

Attempt to visualise Interrodict by using subplots

Parameters

- **shape** (*tuple*) – Layout for the subplots (e.g. (2, 2))
- **truncate** (*int*) – Only process the first *n* items in the class: `corpkit.interrogation.Interrodict`
- **kwargs** (*keyword arguments*) – specifications to pass to `plotter()`

copy ()

flip (*truncate=30*, *transpose=True*, *repeat=False*, **args*, ***kwargs*)

Change the dimensions of `corpkit.interrogation.Interrodict`, making column names into keys.

Parameters

- **truncate** (*int/all*) – Get first *n* columns
- **transpose** (*bool*) – Flip rows and columns:
- **repeat** (*bool*) – Flip twice, to move columns into key position
- **kwargs** – Arguments to pass to the `edit()` method

Returns `corpkit.interrogation.Interrodict`

get_totals ()

Helper function to concatenate all totals

18.3 Concordance

class `corpkit.interrogation.Concordance` (*data*)

Bases: `pandas.core.frame.DataFrame`

A class for concordance lines, with methods for saving, formatting and editing.

format (*kind='string'*, *n=100*, *window=35*, *print_it=True*, *columns='all'*, *metadata=True*, ***kwargs*)

Print concordance lines nicely, to string, LaTeX or CSV

Parameters

- **kind** (*str*) – output format: *string/latex/csv*
- **n** (*int/all*) – Print first *n* lines only
- **window** (*int*) – how many characters to show to left and right
- **columns** (*list*) – which columns to show

Example

```
>>> lines = corpus.concordance({T: r'/*NN.?/* >># NP'}, show=L)
## show 25 characters either side, 4 lines, just text columns
>>> lines.format(window=25, n=4, columns=[L,M,R])
0           we 're in tucson      , then up north to flagst
1   e 're in tucson , then up north      to flagstaff , then we we
```

```
2 tucson , then up north to flagstaff , then we went through th
3 through the grand canyon area and then phoenix and i sp
```

Returns None

calculate()

Make new Interrogation object from (modified) concordance lines

shuffle(inplace=False)

Shuffle concordance lines

Parameters `inplace (bool)` – Modify current object, or create a new one

Example

```
>>> lines[:4].shuffle()
3 01 1-01.txt.conll through the grand canyon area and then phoenix_
←and i sp
1 01 1-01.txt.conll e 're in tucson , then up north to flagstaff ,_
←then we we
0 01 1-01.txt.conll we 're in tucson , then up north to_
←flagst
2 01 1-01.txt.conll tucson , then up north to flagstaff , then we went_
←through th
```

edit(*args, **kwargs)

Delete or keep rows by subcorpus or by middle column text.

```
>>> skipped = conc.edit(skip_entries=r'to_?match')
```

less(kwargs)**

FUNCTIONS

corpkit contains a small set of standalone functions.

19.1 `as_regex`

`corpkit.other.as_regex(lst, boundaries='w', case_sensitive=False, inverse=False, compile=False)`

Turns a wordlist into an uncompiled regular expression

Parameters

- **lst** (*list*) – A wordlist to convert
- **boundaries** (*str* -- 'word'/'line'/'space'; *tuple* -- (*leftboundary*, *rightboundary*)) –
- **case_sensitive** (*bool*) – Make regular expression case sensitive
- **inverse** (*bool*) – Make regular expression inverse matching

Returns regular expression as string

19.2 `load`

`corpkit.other.load(savename, loadaddir='saved_interrogations')`

Load saved data into memory:

```
>>> loaded = load('interro')
```

will load ./*savename*/*loadaddir* as loaded

Parameters

- **savename** (*str*) – Filename with or without extension
- **loadaddir** (*str*) – Relative path to the directory containing *savename*
- **only_concs** (*bool*) – Set to True if loading concordance lines

Returns loaded data

19.3 `load_all_results`

`corpkit.other.load_all_results(data_dir='saved_interrogations', **kwargs)`

Load every saved interrogation in *data_dir* into a dict:

```
>>> r = load_all_results()
```

Parameters `data_dir` (*str*) – path to saved data

Returns dict with filenames as keys

19.4 *new_project*

`corpkit.other.new_project` (*name*, *loc*=`'.'`, `**kwargs`)

Make a new project in `loc`.

Parameters

- `name` (*str*) – A name for the project
- `loc` (*str*) – Relative path to directory in which project will be made

Returns None

WORDLISTS

20.1 Closed class word types

Various wordlists, mostly for subtypes of closed class words

```
corpkit.dictionaries.wordlists.wordlists = wordlists(pronouns=['all', 'another', 'any', 'anybody', 'anyone',  
    wordlists(pronouns, conjunctions, articles, determiners, prepositions, connectors, modals, closedclass, stop-  
    words, titles, whpro)
```

20.2 Systemic functional process types

Inflected verbforms for systemic process types.

```
corpkit.dictionaries.process_types.processes
```

20.3 Stopwords

A list of arbitrary stopwords.

```
corpkit.dictionaries.stopwords.stopwords
```

20.4 Systemic/dependency label conversion

Systemic-functional to dependency role translation.

```
corpkit.dictionaries.roles.roles = roles(actor=['agent', 'agent', 'csubj', 'nsubj'], adjunct=[('prep|nmod')(_|:).*  
    roles(actor, adjunct, any, auxiliary, circumstance, classifier, complement, deictic, epithet, event, existential,  
    finite, goal, modal, modifier, numerative, participant, participant1, participant2, polarity, postmodifier,  
    predicator, premodifier, process, qualifier, subject, textual, thing)
```

20.5 BNC reference corpus

BNC word frequency list.

```
corpkit.dictionaries.bnc.bnc
```

20.6 Spelling conversion

A dict with U.S. English spellings as keys, U.K. spellings as values.

```
corpkit.dictionaries.word_transforms.usa_convert
```

Cite

If you'd like to cite *corpkit*, you can use:

```
McDonald, D. (2015). corpkit: a toolkit for corpus linguistics. Retrieved from  
https://www.github.com/interrogator/corpkit. DOI: http://doi.org/10.5281/zenodo.28361
```

INDEX

all_filepaths (corpkit.corpus.Corpus attribute), 57
all_files (corpkit.corpus.Corpus attribute), 57
annotate() (corpkit.corpus.Corpus method), 62
as_regex() (in module corpkit.other), 75
asciiplot() (corpkit.interrogation.Interrogation method), 70

calculate() (corpkit.interrogation.Concordance method), 74
collapse() (corpkit.interrogation.Interrodict method), 72
Concordance (class in corpkit.interrogation), 73
concordance (corpkit.interrogation.Interrogation attribute), 65
concordance() (corpkit.corpus.Corpus method), 61
concordance() (corpkit.corpus.Datalist method), 64
configurations() (corpkit.corpus.Corpus method), 58
configurations() (corpkit.corpus.Datalist method), 64
copy() (corpkit.interrogation.Interrodict method), 73
corpkit.dictionaries.bnc (built-in variable), 77
corpkit.dictionaries.process_types.processes (built-in variable), 77
corpkit.dictionaries.stopwords.stopwords (built-in variable), 77
corpkit.dictionaries.word_transforms.usa_convert (built-in variable), 78
Corpora (class in corpkit.corpus), 62
Corpus (class in corpkit.corpus), 57

Datalist (class in corpkit.corpus), 64
delete_metadata() (corpkit.corpus.Corpus method), 60
document (corpkit.corpus.File attribute), 63

edit() (corpkit.interrogation.Concordance method), 74
edit() (corpkit.interrogation.Interrodict method), 71
edit() (corpkit.interrogation.Interrogation method), 65
entropy() (corpkit.interrogation.Interrogation method), 70

features (corpkit.corpus.Corpora attribute), 63
features (corpkit.corpus.Corpus attribute), 57
File (class in corpkit.corpus), 63
files (corpkit.corpus.Corpus attribute), 57
flip() (corpkit.interrogation.Interrodict method), 73
format() (corpkit.interrogation.Concordance method), 73

get_totals() (corpkit.interrogation.Interrodict method), 73
Interrodict (class in corpkit.interrogation), 71
interrogate() (corpkit.corpus.Corpus method), 58
interrogate() (corpkit.corpus.Datalist method), 64
Interrogation (class in corpkit.interrogation), 65
interroplot() (corpkit.corpus.Corpus method), 61

keyness() (corpkit.interrogation.Interrogation method), 70

language_model() (corpkit.interrogation.Interrogation method), 69
less() (corpkit.interrogation.Concordance method), 74
lexicon (corpkit.corpus.Corpora attribute), 63
lexicon (corpkit.corpus.Corpus attribute), 58
load() (in module corpkit.other), 75
load_all_results() (in module corpkit.other), 75

make_language_model() (corpkit.corpus.Corpus method), 62
metadata (corpkit.corpus.Corpus attribute), 60
multiindex() (corpkit.interrogation.Interrodict method), 71
multiindex() (corpkit.interrogation.Interrogation method), 70
multiplot() (corpkit.interrogation.Interrogation method), 69

new_project() (in module corpkit.other), 76

parse() (corpkit.corpus.Corpora method), 63
parse() (corpkit.corpus.Corpus method), 60
perplexity() (corpkit.interrogation.Interrogation method), 70
plain (corpkit.corpus.File attribute), 63
postags (corpkit.corpus.Corpora attribute), 63
postags (corpkit.corpus.Corpus attribute), 58

query (corpkit.interrogation.Interrogation attribute), 65
quickview() (corpkit.interrogation.Interrogation method), 69

read() (corpkit.corpus.File method), 63
rel() (corpkit.interrogation.Interrogation method), 70
results (corpkit.interrogation.Interrogation attribute), 65
roles (in module corpkit.dictionaries.roles), 77

sample() (corpkit.corpus.Corpus method), [60](#)
save() (corpkit.corpus.Corpus method), [62](#)
save() (corpkit.interrogation.Interrodict method), [71](#)
save() (corpkit.interrogation.Interrogation method), [69](#)
shannon() (corpkit.interrogation.Interrogation method),
 [71](#)
shuffle() (corpkit.interrogation.Concordance method),
 [74](#)
sort() (corpkit.interrogation.Interrogation method), [68](#)
speakerlist (corpkit.corpus.Corpus attribute), [57](#)
subcorpora (corpkit.corpus.Corpus attribute), [57](#)
Subcorpus (class in corpkit.corpus), [63](#)

tabview() (corpkit.interrogation.Interrogation method),
 [69](#)
tfidf() (corpkit.corpus.Corpus method), [57](#)
tokenise() (corpkit.corpus.Corpus method), [61](#)
topwords() (corpkit.interrogation.Interrodict method),
 [72](#)
topwords() (corpkit.interrogation.Interrogation
 method), [70](#)
totals (corpkit.interrogation.Interrogation attribute), [65](#)
trees (corpkit.corpus.File attribute), [63](#)

unannotate() (corpkit.corpus.Corpus method), [62](#)

visualise() (corpkit.interrogation.Interrodict method),
 [73](#)
visualise() (corpkit.interrogation.Interrogation
 method), [68](#)

wordclasses (corpkit.corpus.Corpora attribute), [63](#)
wordclasses (corpkit.corpus.Corpus attribute), [58](#)
wordlists (in module corpkit.dictionaries.wordlists), [77](#)

